Imperial College London
Department of Bioengineering

Computational modelling of neural mechanisms underlying natural speech perception

Mikolaj Aleksander Kegler
Supervisor: Prof. Tobias Reichenbach
Co-Supervisor: Prof. Mauricio Barahona

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Bioengineering at Imperial College London
March 2022, London, United Kingdom

Copyright Declaration

Declaration of Originality

I declare that the work presented in this thesis is my own, except otherwise acknowledged. Some parts have been conducted in collaboration with other researchers. This is clearly indicated where relevant.

Acknowledgements

To my late grandparents.

This thesis wouldn't be complete without acknowledging many people without whom finishing this PhD would not be possible or, at least, would be much less enjoyable.

I want to acknowledge my closest co-workers and dear friends from the Sensory Neuroengineering lab. I would like to especially thank Dr Octave Etard, whose mentorship during my MSc project fuelled my passion for research and contributed to my decision to apply for the PhD programme. This journey wouldn't be the same without my great friends: Marina, Hugo, Antonio, Laura, Katerina, Shabnam, Mahmoud, Enrico, Pierre, Mike, Anirudh, and Mathilde. Thank you for all the fantastic moments together and for everything I learned from you.

I want to express my extreme gratitude to my supervisors, Prof. Tobias Reichenbach and Prof. Mauricio Barahona, who offered me this unique opportunity. Thank you for giving me the freedom to design my research agenda and for providing excellent mentoring over the years.

I want to thank all my close friends, without whom the last five years in London wouldn't be worth so much to me. Thank you, Martyna, Mario B., Hristo, Jean-Charles, Julia, Clara, Irene, Tomek, Emil, Guillem, Mario V., and Jean. Thanks to all our experiences together, I can now wholeheartedly call London my second home, and you, my second family. Although we are separated by hundreds of miles, I want to thank my friends in Poland for their continuous support and for always welcoming me back home with open arms. Thank you, Kasia R., Gosia, Bartosz, Jakub, Grzegorz, Marcin, Ola, Zuza, Monia, and Kasia O..

Throughout my PhD, I also spent time exploring other fields of science as an intern at Logitech (2019) and Amazon (2021). I want to express my gratitude to Jean-Michel and Milos at Logitech, and Trausti and Tarun at Amazon, for giving me the opportunities to join their teams for an exciting summer of research. I want to thank all my mentors and co-workers at Logitech and Amazon for their incredibly warm welcoming and for everything I learned from them. These unique experiences substantially broadened my scientific horizons and undoubtedly shifted the trajectory of my professional career.

Last but not least, I would like to wholeheartedly thank my entire family, without whose continuous support, trust, and patience, I would have never been able to reach this point.

Abstract

Humans are highly skilled at the analysis of complex auditory scenes. In particular, the human auditory system is characterized by incredible robustness to noise and can nearly effortlessly isolate the voice of a specific talker from even the busiest of mixtures. However, neural mechanisms underlying these remarkable properties remain poorly understood. This is mainly due to the inherent complexity of speech signals and multi-stage, intricate processing performed in the human auditory system. Understanding these neural mechanisms underlying speech perception is of interest for clinical practice, brain-computer interfacing and automatic speech processing systems.

In this thesis, we developed computational models characterizing neural speech processing across different stages of the human auditory pathways. In particular, we studied the active role of slow cortical oscillations in speech-in-noise comprehension through a spiking neural network model for encoding spoken sentences. The neural dynamics of the model during noisy speech encoding reflected speech comprehension of young, normal-hearing adults. The proposed theoretical model was validated by predicting the effects of non-invasive brain stimulation on speech comprehension in an experimental study involving a cohort of volunteers. Moreover, we developed a modelling framework for detecting the early, high-frequency neural response to the uninterrupted speech in non-invasive neural recordings. We applied the method to investigate top-down modulation of this response by the listener's selective attention and linguistic properties of different words from a spoken narrative. We found that in both cases, the detected responses of predominantly subcortical origin were significantly modulated, which supports the functional role of feedback, between higher- and lower levels stages of the auditory pathways, in speech perception.

The proposed computational models shed light on some of the poorly understood neural mechanisms underlying speech perception. The developed methods can be readily employed in future studies involving a range of experimental paradigms beyond these considered in this thesis.

Contents

List of Figures

List of Tables

Motivations

"All models are wrong, but some are useful."
— George E. P. Box, (Box 1979)

Human speech is one of the most complex auditory stimuli in the natural world and requires intricate coordination of over one hundred muscles to produce it (Fant 1970). This complexity makes speech special and allows us to convey substantial amounts of information in a relatively short time span (Pellegrino et al. 2011). On the other end of the communication chain, the message, encoded by the speaker pronouncing the words, need to be decoded by the listener to extract the meaning of individual words and the whole utterance.

To allow effective information transfer, both parts of the human vocal communication system need to be robust to a range of internal and external interferences. For instance, in noisy environments such as bars or restaurants, we instinctively speak louder to increase the relative signal-to-noise ratio of our voice with respect to the noisy background. This phenomenon is commonly known as *Lombard reflex* (Junqua et al. 1999). In the same scenario, the human auditory system needs to exhibit incredible robustness to decode the message in the presence of strong acoustic interference. To achieve such an ability to decode complex messages, and do it effectively in adverse conditions, the human auditory system implements unique neural mechanisms to process speech. Most notably, understanding spoken language requires substantial cognitive resources to extract meaning from perceived sounds. This can be illustrated by comparing the terms *hearing* and *listening*. While seemingly similar and commonly associated with the perception of an auditory stimulus, they vastly differ from the brain's perspective. *Hearing* can be equated to perceiving certain sound as being able to transform it from the acoustic waveform to the neural code. However, *listening* indicates the active involvement in the task, which is crucial for decoding and/or extracting the meaning from the perceived stimulus.

As such, *listening* with the involvement of cognitive resources is critical for effective communication, especially in challenging conditions. One of the classic examples illustrating the power of the human auditory system and the importance of the listener's attention and involvement is a *cocktail party scenario* (Cherry 1953). In this scenario, several sources of speech are mixed to simulate conversations taking place at an imaginary *cocktail party*. Even though the auditory scene contains several talkers, normal-hearing listeners can easily follow one of the voices while ignoring other talkers. The selection of the voice to follow depends on the listener's attentional focus, including not following any of the talkers and ignoring them all. The above example illus-

trates the role attention and involvement play in the perception of speech and clearly highlights the difference between *hearing* (all the speakers talking at once) and *listening* (to only one of them).

Despite decades of research, neural mechanisms underlying natural speech perception still remain poorly understood. This applies to both processes involved in *hearing*, such as the implementation of machinery responsible for translating the sound pressure waveform to neural code, as well as those crucial for *listening*, such as the ability to effectively solve the *cocktail party* problem. Motivations for building computational models that accurately describe these neural mechanisms go beyond pushing the boundary of fundamental knowledge. In particular, understanding neural principles of speech perception is important for many applications, including clinical practice, brain-computer interfaces (BCIs) and automatic speech processing systems.

Firstly, understanding mechanisms underlying speech perception in a normal-hearing population would allow the creation of normative models. Such models could be used to diagnose hearing impairments, such as sensorineural hearing loss, or cognitive dysfunctions, such as attention disorders. Importantly, investigation of a specific clinical population would allow adjusting the model to understand specific changes in neural processing giving rise to different disorders (Frisina et al. 1997). Such models could be furthermore used for individual profiling, for example, in the automated fitting of hearing aids or cochlear implants (Vanheusden et al. 2020).

Secondly, access to an accurate computational model estimating the evolution of a certain biomarker as a function of a stimulus or the subject's state (for example, attention) is critical for auditory BCIs. These systems can range from "*smart*" brain-steered auditory prostheses, which would allow the selective amplification of a speech source based on the listener's attentional focus (Geirnaert et al. 2021b), through portable non-invasive systems for use in education (Davidesco et al. 2021), to algorithms for the assessment of consciousness in patients in locked-in state (Kübler et al. 2009). In either of these BCI applications, the efficacy of the decoding systems usually relies on the fundamental understanding of the neural mechanisms underlying speech perception.

Thirdly, understanding neural mechanisms underlying speech perception can be revolutionary for automatic speech processing systems. Currently, such systems are predominantly implemented as deep neural networks (DNNs), which tend to be computationally heavy and require a considerable amount of energy to achieve a satisfactory or even "super-human" performance (Yu et al. 2016). In contrast, the human auditory system and the brain require very little energy to operate and often outperform DNNs in a range of auditory tasks (Spille et al. 2018). While part of the differences can be attributed to a vastly different *hardware* implementation (processors in modern computers vs biological neurons), the translation of certain algorithmic principles from the brain to DNNs could improve the efficacy of the latter (Marković et al. 2020).

In this thesis, we sought to develop computational models characterizing neural mechanisms

underlying natural speech perception. In particular, we employed electroencephalography (EEG) and transcranial alternating current stimulation (tACS) to respectively record or perturb the neural activity of young normal-hearing volunteers listening to speech. Based on the experimental data and recent theories, we designed computational models for a) the effects of tACS intervention on the cortical processing of speech in noise and the associated speech-in-noise comprehension; b) the early, high-frequency neural response to speech tracking the speaker's pitch. Furthermore, we used the latter model to study how the listener's selective attention and the linguistic properties of different words in spoken narratives influence this early, high-frequency neural response to speech.

Chapter 1

Introduction

This chapter provides a broad introduction of key concepts and previous work related to the original studies introduced in Chapters 2 - 5, which contain their own detailed introductions. As such, section 1.1 provides the reader with a general overview of the field and relevant state-of-the-art research and identifies current knowledge gaps that this thesis aims to address. Section 1.2 provides the outline of the subsequent chapters and their respective aims.

1.1 Background

1.1.1 Overview of the human auditory system

Hearing is one of five mammalian senses. As one of the main instruments for vocal communication, the human auditory system evolved to become incredibly versatile, adaptive, and robust to cope with various, often challenging, conditions. Temporarily putting speech-specific neural processing aside, this section will introduce general auditory processing pathways converting the acoustic pressure wave into the firing of neurons in cortical circuits associated with auditory perception.

In general, the human auditory system (Fig. 1.1) can be split into two major parts: the auditory periphery and the central auditory system (Pickles 1998). The auditory periphery consists of *outer ear*, *middle ear* and *inner ear*. In the *outer ear*, the pinna guides the acoustic waveform towards the *ear canal* and the *tympanic membrane*, which is a gateway to the middle ear. While the intricate shape of the outer ear provides important spectral cues for sound localization, in all experimental studies considered in this thesis, the stimuli were presented to the participants diotically (i.e., the same signal to both ears, commonly referred to as *mono* audio) via insert earphones, not a loudspeaker. As a result, the sound localization cues were absent in the experiments, and thus, the topics of sound localization will not be covered in this introduction. However, for an in-depth review of the topic, please see Ashida et al. 2011. After propagating through the *outer ear*, the sound pressure waveform causes the *tympanic membrane* to vibrate. These vibrations are then transmitted through a set of three bones (*malleus*, *incus* and *stapes*) to the *oval window* of the *cochlea*. The role of the middle ear is to match the high impedance of the fluid-filled *cochlea* and the low impedance of the air through which the sound

is propagating (Maier et al. 2016).

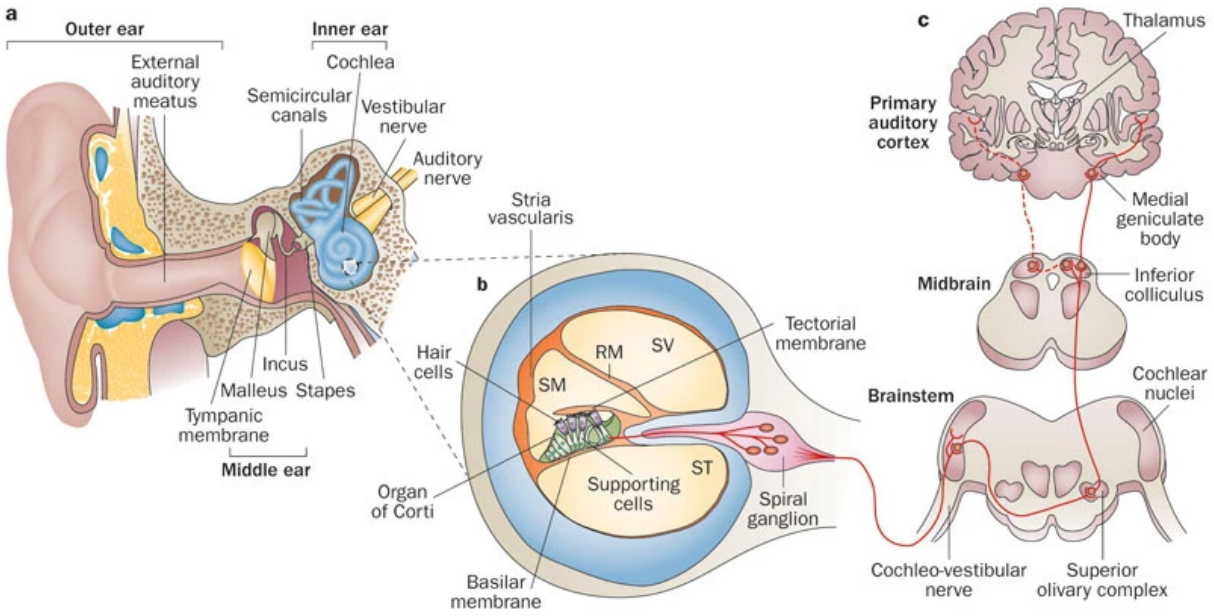

Figure 1.1: The human auditory system. Figure reproduced with permission from Ng et al. 2013.

The *cochlea* in the inner ear is a processing stage at which vibrations of the oval window are transduced into electrical impulses in the *auditory nerve* (Pickles 1998). The *cochlea* is a snail-shaped fluid-filled structure with two compartments (*scalae*) separated by the *basilar membrane*. As the *oval window* vibrates, the *basilar membrane* vibrates accordingly due to the propagation of vibrations in the fluid. Since the *basilar membrane* varies in its stiffness and width across the cochlea (wider and stiffer towards the base, near the *oval window*), the response of its different parts is frequency-dependent. In particular, each part of the *basilar membrane* is tuned to respond the most to a different frequency, and the entire *basilar membrane* effectively performs spectral decomposition of the input sound. Such frequency-mapping across the *basilar membrane* is known as *tonotopy* and is maintained across consecutive stages of the human auditory pathways.

The conversion of *basilar membrane* vibrations to electrical impulses is performed in the *organ of Corti* placed on top of the membrane (Maoiléidigh et al. 2019). Among several types of cells inside the *organ of Corti*, the *inner hair cells* placed along the *basilar membrane* carry out the mechano-electrical transduction in the inner ear. In particular, the *hair cells* are sensitive to the displacement of the *basilar membrane*. Following their displacement, the cell membrane depolarizes, and when the depolarization exceeds the firing threshold, an action potential is sent down the cell's axon and subsequently depolarizes neurons in the auditory nerve. Note that, since inner hair cells are located along the tonotopically-organized *basilar membrane*, the activity of the auditory nerve fibres will also maintain the tonotopy.

The next stage in the auditory processing pathways is the brainstem (Irvine 2012). The signal from the auditory nerve fibres is projected to the *cochlear nucleus* and subsequently the midbrain. In the midbrain, the ascending information is propagated either to the *inferior colliculus* or the *superiory olivary complex*. The former is considered an information transfer hub for sensory information, including sound, and the latter integrates information from both ears and plays an important role in sound localization. As indicated before, this review will not focus on sound localization. While the exact role of the *inferior colliculus* is not fully understood, it is known to contribute to the integration of information, both within the auditory modality, by integrating primary and previously segregated information, as well as across sensory modalities. The last subcortical nucleus is the *medial geniculate body* of the *thalamus*.

From the *thalamus*, the information is finally projected to the auditory cortices located bilaterally in the temporal lobes of the brain. It takes the neural signal about 50 ms to reach this stage since the sound enters the ear canal (Pickles 1998). Brainstem responses are typically associated with latencies between single ms to approximately 20 ms, depending on the exact stage of the brainstem. The cortex is considered the main centre where the processing related to the encoding of auditory stimuli and/or extraction of information happens. The cortical circuits involved in auditory perception implement a hierarchical encoding of the auditory input (Sharpee et al. 2011). In particular, areas in the primary auditory cortex encode spectral content of the auditory stimulus in a tonotopic fashion, which is preserved all the way from the inner ear (Saenz et al. 2014). In addition to the tonotopic frequency mapping, cortical circuits implement a variety of spectro-temporal filters, which gradually extract more complex representations of the stimulus.

Like many other cortical areas, the auditory cortex is characterized by dense, complex connectivity patterns, both in terms of local connections and long-range projections between distant regions of the brain. This is particularly relevant for speech and language processing (Hickok et al. 2007), which is known to interact with the prefrontal cortex, motor cortex, as well as Broca's and Wernicke's areas, which are considered critical for speech production and perception, respectively. As such, the cortical processing of sound should not be considered a strictly feed-forward process. Especially higher-level cognitive processing heavily relies on the exchange of information between different cortical areas.

The cortical activity also modulates the earlier stages of the auditory pathways. In particular, extensive networks of efferent connections can relay information from higher-level cortical areas all the way to the inner ear (Pickles 1998; Saiz-Alía et al. 2021; Terreros et al. 2015; Winer 2005). While these connections are not direct, the cortical circuits can modulate the activity of the auditory brainstem, in particular, the aforementioned *inferior colliculus*. The descending connections between cortical and subcortical areas form the *corticofugal pathways*. Subsequently, the subcortical nuclei can project the descending signal back to the *cochlea* in the *inner ear*. The main role of the efferent connection is to provide adaptation for the lower-level stages of the auditory pathways in a *feedback* fashion, in contrast to the above described *feed-*

forward processing of the sound. While the functional role of these *feedback* projections is not fully understood, previous studies showed that it contributes to the protection from acoustic trauma (Maison et al. 2000), as well as acts as a top-down frequency filter needed in different scenarios (Terreros et al. 2015). The latter mechanism is known to be controlled by the selective attention of the listener and thus is critical for coping with noisy environments.

In summary, the above section provided a general overview of the human auditory pathways, including its peripheral and central parts, as well as *feed-forward* (afferent) and *feedback* (efferent) processing. While the auditory system across mammalian genera shares many similarities, both in terms of structure and function, humans stand out from other species because of their ability to speak. In particular, over the centuries, the human auditory system developed speech-specific adaptations (Fitch et al. 1997). The remainder of this introduction will focus on speech-specific neural processing unique for humans. The next sections will review methods used for recording neural responses to speech or for perturbing neural processing to manipulate speech perception. The further parts of this introduction will extend the above general description with a review of neural mechanisms characteristic to speech, both on cortical and subcortical levels, and specific to feed-forward and feedback auditory processing pathways.

1.1.2 Non-invasive recording of neural responses to speech

In order to study neural mechanisms of speech processing, a suitable neuroimaging toolkit is necessary. While the vast majority of fundamental research on auditory perception has been conducted in an animal model, the studies of speech perception can be performed only in humans (Fitch et al. 1997). Spoken language can be used as a stimulus presented to non-human animals, and some of them can learn to interpret speech by reacting differently to different words or phrases (Kluender et al. 2012). This way, it is possible to study how complex spectro-temporal patterns present in speech are processed across the mammalian auditory pathways. However, here, we will focus on studying neural mechanisms of speech perception in humans. This section will introduce the methodological framework for studying neural responses across the human auditory system. In particular, the leading modalities will be compared and narrowed down to methods most relevant for the work conducted in this thesis.

While the vast majority of studies in animal models can record the neural responses directly from the brain circuits or single neurons using invasive probes, that is rarely available in healthy humans (Engel et al. 2005). Because of the ethical reasons and inherent risk associated with the surgical procedure required to implant the recording device in the brain, the direct recordings of human brain activity are incredibly rare to come across. In fact, the implantation of recording probes for long-term continuous monitoring is limited only to clinical populations at risk, such as those suffering from drug-resistant depression or seizures, or disabled volunteers, for whom access to a brain-computer interface can vastly improve their quality of life. Implantable probes (such as *Utah arrays*), electrocorticography (ECoG), intracranial (iEEG) and stereotactic electroencephalography (sEEG) belong to the currently most common invasive neuroimaging methods

used in human studies. While studies having access to invasively recorded data are crucial for understanding neural mechanisms of speech perception, they are usually investigated in healthy populations using non-invasive neuroimaging methods. In particular, the four leading non-invasive brain imaging modalities available to researchers nowadays are electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imagining (fMRI) and functional near-infrared spectroscopy (fNIRS). The four modalities are briefly introduced below.

EEG is the oldest of the four methods, with the neural activity in humans recorded for the first time in the 1920s by Hans Berger (İnce et al. 2020). Electroencephalogram (i.e., the signal captured using EEG) represents the electrical activity of the brain picked up by electrodes located on the participant's scalp. Because of the skull, tissue, and scalp, the source electrical signal is greatly attenuated, and it is not possible to distinguish activation of single neurons, but rather detect synchronous activations of large populations of cells (Hari et al. 2017; Klonowski 2009). As a result, the amplitude of neural signal picked up by EEG is very low, usually of the order of single microvolts. In addition, a typical encephalogram contains many sources of noise ranging from the cross-talk between other electrical signals generated in the body, such as an electrocardiogram (ECG) or electrooculogram (EOG), reflecting the heart activity and the eye movement, respectively, to the intrinsic noise of the EEG recording equipment. The latter can be attributed to the electrical properties of the amplifier or impedances between recording electrodes and the scalp. As such, EEG typically has a highly negative signal-to-noise ratio (SNR), meaning that the background noise and/or unwanted signals exceed the source signal of interest.

In contrast to EEG, magnetoencephalography (MEG) records small changes in the magnetic field resulting from the current flow in the population of neurons (Hari et al. 2017). Omitting fine details, the signal obtained using MEG has typically higher SNR than EEG, and the two modalities are similar in terms of signal utility and associated analysis techniques. In order to record small MEG signals, orders of magnitude lower than the background magnetic activity, the recordings require highly specialized equipment and infrastructure. In particular, the devices need to be equipped with sensitive magnetometers, which require liquid helium cooling to achieve optimal working conditions. In addition, the MEG device needs to be located in the magnetically shielded rooms in order to pick up signals orders of magnitude lower than the earth's magnetic field. This is associated with a high cost of devices and their usage, resulting in limited availability of the technology worldwide. Importantly, recent efforts focus on miniaturizing the MEG to allow the participant to move freely within the shielded room (Boto et al. 2018), which is one of the limitations of MEG. In contrast, EEG is already available in a portable form factor, which can be easily *taken for a walk* (Debener et al. 2012; Hölle et al. 2021).

Unlike EEG and MEG (often jointly referred to as M/EEG), fMRI and fNIRS operate on a vastly different basis (Buxton 2013; Ferrari et al. 2012). In particular, they do not record the electrical activity of the brain, but changes in the oxygenation of blood supplied to different brain regions. Such blood-oxygenation-level-dependent (BOLD) signal correlates well with the activation of neural populations since oxygen is required to fuel them. Without getting into fine

details, fMRI measures the changes in magnetic properties of tissue, while fNIRS is an optical imaging technique that measures the difference in the optical wavelength as the light propagates through the oxygenated or de-oxygenated blood. fMRI deriving from the MRI structural imaging offers superb spatial resolution, often below 1 mm, compared to any other non-invasive neuroimaging modality. However, to achieve such exceptional spatial resolution, the fMRI requires long signal acquisition times, and as a result, standard fMRI protocols can rarely achieve sub-second temporal resolution. fNIRS is associated with a better temporal resolution of even below 100 ms. However, the method offers poorer spatial resolution than fMRI, mainly due to the attenuation and scattering of the light in the tissue. Unlike fMRI, it does not require an expensive and large MRI scanner, and most available systems are portable and allow participants to move freely. Since fNIRS does not record electrical signals, it is suitable for monitoring the brain activity of cochlear implant users, who are not eligible for fMRI scans, and whose M/EEG responses to auditory stimuli are contaminated with large stimulation artefacts (Sevy et al. 2010).

The two families of methods, M/EEG and fMRI & fNIRS, are commonly used for studying neural mechanisms of speech processing in humans. However, as outlined above, they offer vastly different trade-offs, mainly in terms of their respective spatial and temporal resolution. M/EEG reflects direct electrical activity of the brain and offers sub-millisecond temporal resolution (usually limited by the sampling rate of the recording device). However, the volume conduction of tissue between the neural source and sensors outside the scalp limits the spatial resolution of the modality. While it is possible to estimate the source of activity from sensor-space signals, the efficacy of the method relies on the availability of precise structural data reflecting the anatomy of the participant. In contrast, fMRI offers sub-millimetre spatial resolution, which is priceless for the precise mapping of brain networks involved in speech processing. However, the method is associated with a poor temporal resolution of about 1 second, which corresponds to the duration of a phrase or a short sentence in spoken language. EEG and fMRI can be recorded simultaneously to overcome the above limitations. This, however, requires an MRI-compliant EEG system and intensive post-processing of the EEG signal contaminated by currents induced by even the tiniest movements of the electrodes and cables in the magnetic field of the MRI scanner (Ritter et al. 2006).

From the perspective of studying the neural mechanisms of speech processing, there is, unfortunately, not a single modality that offers both good spatial and temporal resolution. On the one hand, since speech processing involves the communication of different, distant brain regions, fine spatial resolution is desired to pinpoint the activation of different brain structures. On the other, speech is a complex stimulus with a great range of temporal dynamics even in a single utterance. Ranging from the slow temporal fluctuation of the energy envelope of sound up until the high-frequency temporal fine structure. As such, the choice of a suitable neuroimaging method depends on research questions and the experimental paradigm of interest.

Historically, both in M/EEG and fNIRS & fMRI, the most common paradigm was to study event-related responses obtained by averaging signals recorded for many repetitions of the same

or similar stimuli to overcome the low SNR of the recorded data (Makeig et al. 2004). Arguably, listening to the same short auditory stimuli, such as syllables, words, or short sentences, repeated many times does not reflect how most humans experience the world. Thus, in recent years many studies attempted to step out of the highly-structured experimental paradigms and study the brain using naturalistic, ecologically-relevant continuous speech stimuli (Brodbeck et al. 2020b). Studies included in the main chapters of this thesis follow this notion and focus on the processing of natural, unaltered speech in the form of long narratives, such as audiobooks or sets of spoken sentences.

Since the naturalistic, continous speech stimuli contain little to no repetitions, the recorded neural signals need entirely different treatment in comparison to the previous research focused on event-related activity. This lack of repetitions can be problematic for fMRI and fNIRS characterized by poor temporal resolution, and thus they are less common for studying continuous speech processing of longer uninterrupted narratives. The experiments with their use are typically designed to contrast brain responses to short sentences in different conditions, such as varying levels of background noise, or compare responses to speech and non-speech stimuli. In contrast, M/EEG offering superb temporal resolution is particularly suited for studying rapid neural dynamics associated with natural speech processing (Brodbeck et al. 2020b).

However, because of the low SNR of M/EEG, the identification of neural activity underlying continuous speech processing is not trivial, and the signal requires special treatment to study the neural mechanism of continuous speech processing. Since the raw signal cannot be directly interpreted due to the low SNR and non-repetitive stimuli, the relationship between auditory stimuli and the associated M/EEG signal is most often studied using computational models. Temporal response function (TRF) is currently one of the most common modelling frameworks for mapping naturalistic continuous stimuli to neural recordings, especially in the field of speech processing (Fig. 1.2) (Brodbeck et al. 2020b; Crosse et al. 2016; Crosse et al. 2021).

TRF commonly refers to the linear modelling framework for mapping a feature of naturalistic, continuous stimulus, for example energy fluctuations, or onsets of words, to the neural responses usually measured using M/EEG (Brodbeck et al. 2020b; Crosse et al. 2016; Crosse et al. 2021) (Fig. 1.2b). In particular, the linear model $\alpha_{c,\tau}$ is optimized to predict the multichannel neural response $r(t, c)$ at channel c and time t, from the stimulus feature $s(t)$:

$$r(t, c) = \sum_{\tau=1}^{T} \alpha_{c,\tau} s(t - \tau), \tag{1.1}$$

where τ corresponds to the latency between the stimulus and the response. Typically, models include a range of T time lags. As a result, the coefficients of the linear model $\alpha_{c,\tau}$ correspond to the strength of the neural response at channel c and τ latency between the stimulus feature and the response (Haufe et al. 2014). The TRF is often referred to as the "*forward model*", because the mapping direction follows the flow of information across the auditory system (i.e., stimulus $\longrightarrow$ response).

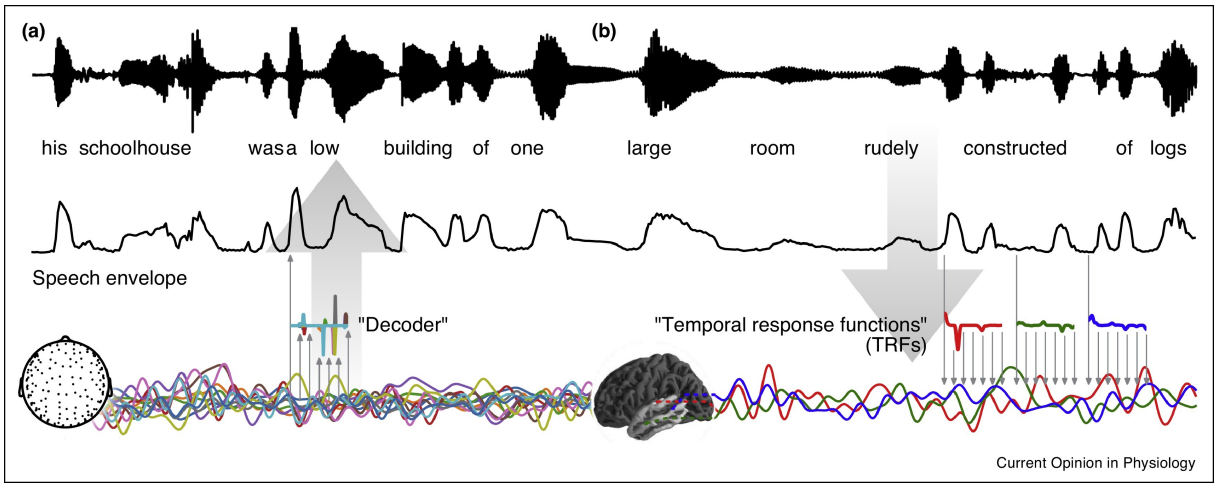

Figure 1.2: Modelling of neural responses to continuous speech. Linear models are optimized to find the optimal mapping between the stimulus feature (for example, speech envelope fluctuations) and the associated neural response. The model can be a "*decoder*" optimized for reconstructing the stimulus from time-lagged neural response **(a)**, or a "*temporal response function*" (TRF) optimized for predicting neural responses from the stimulus feature **(b)**. Figure reproduced with permission from Brodbeck et al. 2020b.

In contrast, it is also possible to optimize the "*decoder*", which reconstructs the stimulus feature from the multichannel neural response (Fig. 1.2a):

$$s(t) = \sum_{c=1}^{C} \sum_{\tau=1}^{T} \beta_{c,\tau} r(t+\tau, c), \tag{1.2}$$

where C corresponds to the number of channels, and the remaining notation is the same as for the *forward model*. Since, in this case, the mapping direction is the opposite to the information flow across the auditory pathways (i.e., stimulus $\longleftarrow$ response), this model is often referred to as the "*backward model*". While the coefficients of the model β should not be directly interpreted as the neural response (Haufe et al. 2014), the model's stimulus reconstruction performance reflects how well is the stimulus feature represented in the neural response. This type of model is particularly popular in BCI applications, to, for example, decode the user attentional focus from the strength of the stimulus feature reconstruction (Biesmans et al. 2016; Geirnaert et al. 2021b; O'Sullivan et al. 2015).

The above-described models are usually optimized using regularized ridge regression (Crosse et al. 2021; Friedman et al. 2001):

$$\alpha = (X^T X + \lambda I)^{-1} X^T y, \tag{1.3}$$

where, α corresponds to the model coefficients; X is the design matrix containing time-lagged stimulus (*forward model*) or time-lagged multichannel response (*backward model*); y represents the model reconstruction target, either multichannel neural response (*forward model*) or stim-

ulus feature (*backward model*); I is an identity matrix and λ is the regularization parameter. Since the problem underlying the models is usually ill-posed, the use of regularization (usually l_2) is essential for numerical stability and for preventing the model coefficients from exploding or vanishing (Crosse et al. 2021; Tikhonov 1963).

It is worth highlighting that the amount of regularization is the only user-controlled hyperparameter of the model[1] and thus needs to be selected with caution to prevent potential overfitting (Crosse et al. 2021). The amount of regularization can be set to the a-priori-selected fixed value (common strategy for the forward models, e.g., Etard et al. 2019a; Fiedler et al. 2019) or optimized to yield the optimal predictive performance of the model. In the latter case, it is crucial to set aside a portion of the data not used to optimize the model and the regularization parameter, which would serve to provide an unbiased evaluation of the fully-optimized model (Crosse et al. 2021).

Importantly, regularized ridge regression is not the only way to optimize linear models for studying the neural response to natural speech, with Boosting being often proposed as an alternative approach (David et al. 2007; Ding et al. 2012). In brief, boosting algorithm is an iterative, sparse estimation technique using a greedy coordinate descent. Starting from all-zero model coefficients, the algorithm incrementally adds small, fixed values to decrease the mean square error (MSE) at each iteration. The algorithm is stopped as subsequent iterations do not reduce the MSE anymore. Compared to the ridge regression, with a smooth l_2 penalty, the boosting algorithm tends to yield sparser forward models with more pronounced, sharper peaks (Kulasingham et al. 2022). One practical advantage of the boosting algorithm is the lack of the regularization parameter, which needs to be either user-defined or carefully optimized in the case of ridge regression. In a recent systematic comparison of the two methods, Kulasingham et al. 2022 showed that (sparse) boosting and (smooth) ridge regression yield comparable performance in estimating neural responses to speech.

The above introduction to the methodology used for studying neural responses to continuous speech is meant to be general and acquaint the reader with key aspects of the framework. Importantly, neural responses to continuous speech can also be modelled using non-linear models, such as deep neural networks (Keshishian et al. 2020; Yang et al. 2015). While the latter family of models allows studying non-linear mechanisms involved in speech processing, the models are inherently more difficult to interpret. Applications of the above outlined linear- and non-linear modelling frameworks for studying neural mechanisms of speech processing on cortical and subcortical levels, alongside the review of key findings, are presented in the next sections of this introduction.

[1] Excepting the number of channels and/or time lags.

1.1.3 Non-invasive brain stimulation for modulation of speech perception

While neuroimaging methods outlined in the previous section have been extensively applied to study the brain for decades, they are not flawless scientific discovery tools. It is worth noting that imaging the brain allows to "only" observe how the activity of particular regions changes as a function of stimulus presentation, experimental condition or participant's behaviour. As such, through only observation, the conclusion about the exact neural mechanisms giving rise to certain behavioural changes in the participants cannot be confidently formulated (Bergmann et al. 2021). In particular, the observed changes of neural activity can be directly induced by the stimulus or occur indirectly, as an epiphenomenon of other processes not directly related to the stimulus processing.

In other words, from just neuroimaging, one cannot clearly determine the causality of certain neural mechanisms on either stimulus encoding or behavioural changes, such as speech comprehension. In fact, without a direct intervention targeting the neural activity of interest, it is not possible to definitively determine whether the changes in the neural activity *enable* certain behavioural changes or whether behavioural changes lead to the alteration of neural activity. The above arguments can be summarized as *correlation* not equating *causality*. In studies of neural mechanisms of speech perception, the awareness of the above caveat is critical for preventing incorrect interpretations of neuroimaging results.

Fortunately, there exist tools for the direct modulation of neural activity in humans to causally study the role of neural mechanisms on perception or behaviour. Similarly, as for neuroimaging, the methods can be split into invasive and non-invasive. Invasive methods involve stimulation of brain tissue directly through implantable electrodes, and thus their use is limited to patient populations at risk, which requires continuous long-term neurological observation. In turn, non-invasive brain stimulation (NIBS) methods recently had their revival and became particularly popular in the fields of psychology and cognitive neuroscience (Dayan et al. 2013; Herrmann et al. 2013; Miniussi et al. 2013; Parkin et al. 2015). Since the latter category of methods applies to a much larger population, we will focus on how they are used for studying neural mechanisms of speech processing.

The methods for NIBS can be split into: transcranial magnetic stimulation (TMS) (Hallett 2007), transcranial current stimulation (TCS) (Paulus 2011) and focused ultrasound stimulation (FUS) (Kubanek 2018). Among the three, TCS is most commonly used to study neural mechanisms of continuous speech processing (Brodbeck et al. 2020b), TMS has also been applied, but mainly in studies investigating neural mechanisms of language-specific processing (Devlin et al. 2007), and finally, FUS, as a relatively recently developed method, has not been yet applied to study speech perception. Thus, this review will predominantly focus on the mechanisms of TCS and its application for studying neural mechanisms of natural speech processing.

TCS uses weak electrical current applied to the scalp to modulate the cortical activity in the brain (Paulus 2011). Since the electrical signal can take an arbitrary shape, limited only

by parameters of the current generator, it allows greater flexibility in terms of the stimulation protocol design than TMS, which uses strong magnetic pulses to induce current flow in the cortical neurons and cause them to fire action potentials. Compared to TMS, TCS applied below the skin sensation threshold, usually below 2 mA, produces only weak modulation because most of the applied current is attenuated in the tissue before reaching the brain. As a result, TCS typically weakly depolarizes membranes of neurons in the superficial layers of the auditory cortex by single millivolts (Baltus et al. 2018). However, even such small depolarization applied to a population of cortical neurons can impact auditory perception, as evidenced by many experimental studies (Heimrath et al. 2016; Helfrich et al. 2014).

TCS can be split into two main categories depending on the shape of the stimulation signal: transcranial direct current stimulation (tDCS) and transcranial alternating current stimulation (tACS). The goal of the former is to constantly modulate the excitability of the neuronal population in a certain brain region and thus causally investigate its role in speech or auditory perception. In contrast, the goal of tACS, usually taking the shape of a sine wave, is to selectively excite and inhibit the intrinsic or stimulus-related activity of a neuronal network. By perturbing the neural activity using either of the TCS types, it is possible to observe how the perception or comprehension of speech changes, as a function of stimulation parameters, for instance, placement of the stimulation electrodes or stimulation intensity. Based on the outcomes and initial hypothesis, the significant modulation or lack thereof sheds light on the causal role of targeted neural mechanisms in speech perception (Bergmann et al. 2021) (Fig. 1.3A).

While TCS is a valuable tool complementing neuroimaging for studying the neural mechanisms of speech perception, the design of the experiments and interpretation of the results need to be done carefully to avoid misinterpretation. In particular, since the stimulation is non-invasive, it is virtually impossible to target a specific brain region without stimulating other neighbouring areas due to the volume conduction of the tissue and stimulation hardware limitations. As such, some of the TCS-induced effects might emerge not through the targeted neural mechanisms but indirectly, through the stimulation of other networks (Fig. 1.3B).

A common way for controlling the emergence of the effects through non-target mechanisms is to employ unrelated stimulation protocol as a control condition. For example, stimulating other brain areas or using the stimulation signal mismatched with the targeted brain activity would indicate what behavioural changes are associated with the lack of effects. To further understand the impact of TCS on brain activity, neuroimaging can be employed to monitor changes in neural responses as a function of the stimulation protocol. While commonly paired with M/EEG (Helfrich et al. 2014) and fMRI (Zoefel et al. 2018), TCS contaminates the M/EEG signal with artefacts that are not trivial to suppress, especially for alternating non-rhythmic currents (Noury et al. 2016). fMRI is not impacted by stimulation artefacts but requires special MRI-compliant TCS devices.

Although TCS has been extensively used to study neural dynamics of cognition, the exact

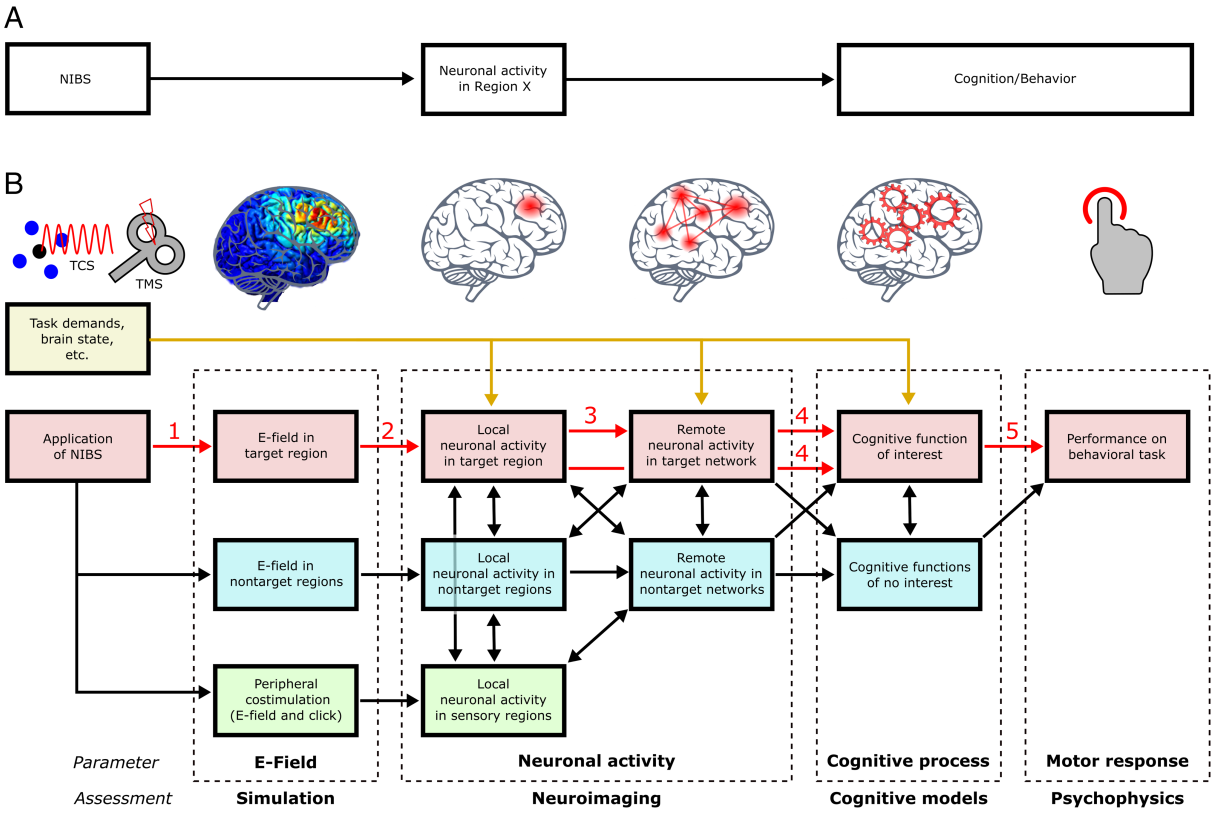

Figure 1.3: Causal diagram for non-invasive brain stimulation (NIBS) studies in cognitive neuroscience. **(A)** Simplified diagram for causality in NIBS studies. The stimulation is expected to affect one particular brain region associated with certain neural activity underlying the participant's perception or behaviour. **(B)** Extended causal diagram illustrating possible mechanisms through which the effects of NIBS emerge. Red arrows indicate the causal information flow of the neural mechanism targeted by NIBS. Yellow arrows reflect the influence of the task on the activity of the targeted network. Black arrows illustrate other possible interactions. (1) Firstly, the focality of the stimulation and the electric field (E-Field) induced in the brain is limited by the stimulation equipment and volume conduction in the skin, scalp and skull tissue, especially in TCS. (2,3) Secondly, the indirect stimulation might impact other brain networks (both local and remote, depending on the stimulation paradigm) not directly involved in the task. (4,5) Finally, the changes in neural activity of both targeted and non-target brain networks may impact cognitive mechanisms and cause behavioural effects. Reproduced with permission from Bergmann et al. 2021.

mechanisms through which the weak electrical current influences cortical networks and produces behavioural effects remain unknown. Computational modelling has been recently proposed as a promising tool for explaining the impact of TCS on behaviour (Fröhlich et al. 2015). In particular, finite-element models estimating the strength of electric fields induced in the brain are commonly used to optimize the stimulation setup (Datta et al. 2009; Huang et al. 2019a) and can explain some of the inter-subject variability observed in the experiments (Kasten et al. 2019). The other family of functional models explaining the effects of TCS on dynamics in the cortical networks is much less advanced. While, the spiking neural networks were developed to study the impact of external stimulation current on the population dynamics (Ali et al. 2013; Cakan et al. 2020; Herrmann et al. 2016) and plasticity (Farahani et al. 2021; Kronberg et al.

2020), computational models for the effects of TCS on speech processing were not yet developed[2].

1.1.4 Cortical speech processing

Understanding neural mechanisms of speech and language processing have been one of the key areas of auditory neuroscience for decades. To date, most of the studies focused on investigating cortical responses to speech stimuli using M/EEG or fMRI. As discussed in the previous sections, M/EEG, due to its high temporal resolution, is particularly suited for studying the dynamics of neural responses to speech. Prior to the development of recent frameworks for modelling neural responses to continuous, uninterrupted speech, most studies focused on event-related potentials (ERPs). In particular, the analysis of evoked responses to syllables, words or short sentences became a natural extension of research investigating auditory (not speech-specific) responses to pure tones or other short, synthetic stimuli (Vaughan Jr et al. 1970; Wood et al. 1971).

The cortical auditory evoked potential (CAEP) is a stereotypical M/EEG response elicited by the sound stimulus. CAEP is typically obtained by averaging responses to many stimulus repetitions, as the activity elicited by a single stimulus presentation has a very low SNR. Having averaged many responses to combat low SNR of the encephalogram, CAEP emerges as a stereotypical response characterized by a complex of negative and positive peaks, N1 and P2, occurring at about 100 and 200 ms after the onset of the stimulus, respectively. While the response tends to be similar across participants, it can be influenced by the listener's attention (Picton et al. 1974). Importantly, the later components of the CAEP, after 200 ms, are typically associated with higher-level cognitive processing. In particular, positive peak P300, occuring at about 300 ms, is known to be strongly modulated by the listener's attention (Picton 1992). This attentional effect is consistent in the population of healthy adults and robust enough to be utilized in P300-based BCIs (Klobassa et al. 2009).

ERP studies investigating neural mechanisms of speech or language processing typically employ short stimuli, such as isolated syllables, words or short sentences. Although limited in terms of their ecological validity, ERP studies unravelled many cortical mechanisms of speech processing, such as attentional modulation of the response (Hansen et al. 1983) as well as correlates of linguistic expectancy (Kutas et al. 2011, 1984). While ERPs still remain the default tool for studying brain dynamics, more and more studies employ naturalistic continuous speech stimuli, such as audiobooks. In particular, Lalor et al. 2010 was the first to show that coefficients of linear TRF models reflect the CAEP obtained in the classic ERP paradigm.

Due to the methodological advances and gradually stepping away from the ERP framework, endogenous cortical oscillations, and rhythmic activity in the brain in general, gathered the interest of the community studying neural mechanisms of speech processing. In particular, these oscillations, occurring across a broad range of frequencies, coincide with the rate of different linguistic units in spoken language (Brodbeck et al. 2020b; Giraud et al. 2012; Meyer 2018). For

[2]Excepting Kegler et al. 2021 presented in this thesis (chapter 3).

instance, lower frequencies in the theta frequency range (4 – 8 Hz) reflect the rate of syllables, while the activity in the gamma band (above 25 Hz) is similar to the rate at which phonemes are produced. These observations, backed by experimental evidence, became the foundation of the theory of speech encoding through coupled cortical oscillations (Giraud et al. 2012) (Fig. 1.4).

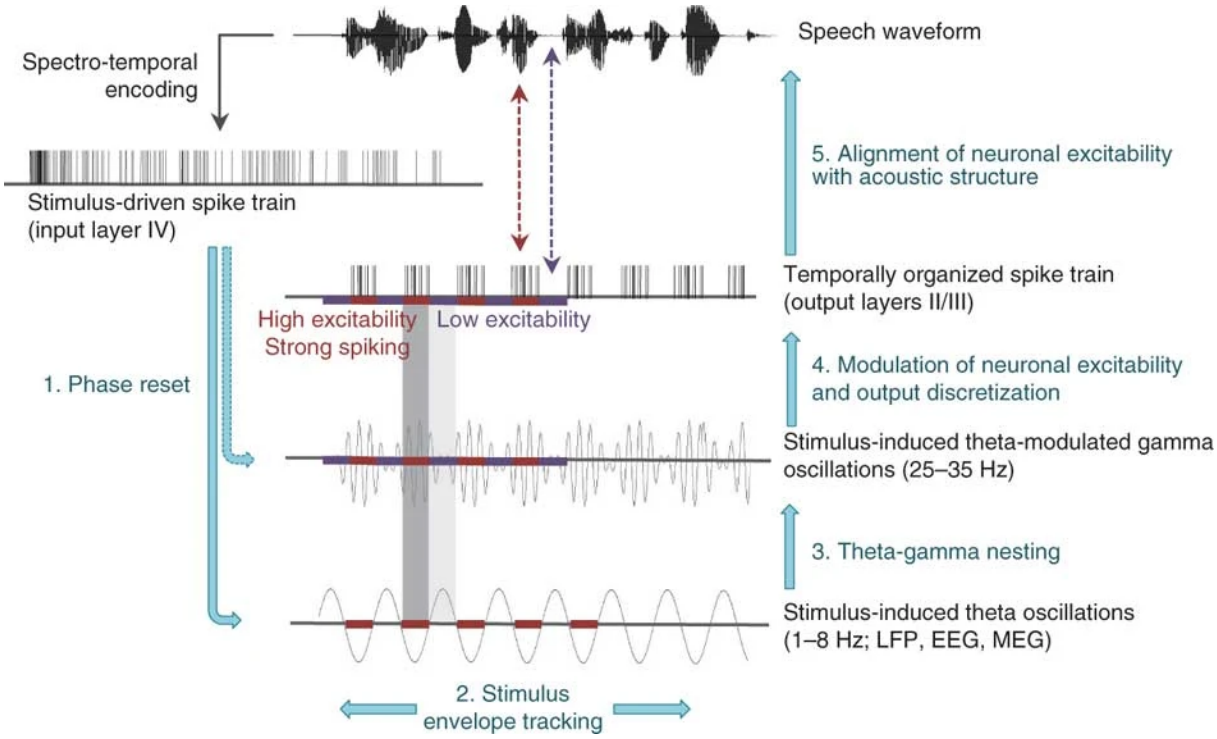

Figure 1.4: A theory of early cortical speech encoding through coupled neural oscillations. The speech signal processed through the auditory periphery and subcortical structures arrives at the input layer IV of the primary auditory cortex as a stimulus-driven spike train. (1) Such stimulus-driven spike train elicits a reset of theta oscillations in superficial cortical layers. (2) After the phase reset, the low-frequency oscillations track the envelope of the input speech signal. (3) The oscillatory activity in the higher gamma-band is modulated by slower theta-band oscillations. (4) Such theta-modulated gamma power controls the excitability of neurons generating the feedforward signal to higher-order cortical areas. (5) The cross-frequency coupled oscillations segment the input stimulus-driven spike train into smaller chunks corresponding to syllables in the input utterance. Reproduced with permission from Giraud et al. 2012.

According to the theory, cortical oscillations play an active role in the rapid segregation of acoustic information and thus facilitate neural encoding of speech. In particular, the cross-frequency coupling of oscillations in the theta and gamma bands is key for the efficient encoding of speech. As the acoustic waveform is encoded into neural signals through the neural machinery in the auditory periphery and the brainstem, it arrives as a stimulus-driven spike train at layer IV of the auditory cortex (Sakata et al. 2009). The presence of the speech in the acoustic input causes the intrinsic slower oscillations in the delta and theta bands (1 – 8 Hz) to follow the energy envelope of the input utterance (Ghitza 2011; Giraud et al. 2007; Gross et al. 2013). Since the slow oscillatory rhythm is matched to the rate of phrases or syllables in spoken language, this mechanism is postulated to be involved in the parsing of upcoming information and

segregating it into smaller segments (Brodbeck et al. 2020b; Giraud et al. 2012; Kayser et al. 2012; Meyer 2018). This phenomenon is commonly referred to as *cortical tracking* of speech or *neural entrainment to speech* (Obleser et al. 2019).

It is known that gamma-band neural activity in the auditory cortex is responsible for the encoding of the spectral content of the input utterance (Brosch et al. 2002). In particular, the frequency of endogenous gamma bursts coincides with the rate at which the phonemes occur in spoken language. Thus, the gamma-band oscillations are typically associated with the encoding of phonemic information (Giraud et al. 2012; Gross et al. 2013). Through the cross-frequency coupling between the oscillatory activity in the theta and gamma frequency ranges, the characteristic nesting of neural oscillations occurs (Schroeder et al. 2009). In particular, even at rest, without sensory input, slow oscillations modulate those in the higher frequency range (Jensen et al. 2007). However, in the presence of speech stimulus, this coupling becomes stronger. During the encoding of speech, the slower oscillations start following the energy envelope fluctuations in the input and, at the same time, modulate the faster intrinsic activity in the gamma band. Such temporal gating causes the rhythmic facilitation of the gamma-band activity by increasing the excitability of neurons. This, in turn, allows segmentation of the stimulus-driven spike train, encoded through the gamma oscillations, to be segmented into smaller chunks typically associated with syllables. Each encoded segment represents the acoustic content of the syllable, as encoded by the gamma-band activity. Notably, the proposed cortical tracking mechanism is capable of adapting to different rates of speech production without loss of comprehension (Giraud et al. 2012; Hyafil et al. 2015). However, compromised cortical tracking is associated with degraded comprehension of time-compressed speech (Ahissar et al. 2001; Nourski et al. 2009).

Although still often debated (Doelling et al. 2021) and missing some experimental evidence, the theory of speech encoding through coupled neural oscillations is currently a leading model for speech-specific neural coding in the auditory cortex. Over the years, neuroimaging studies have provided convincing evidence for the strong phase-locking of oscillations in the theta-band to the envelope of speech (Brodbeck et al. 2020b; Luo et al. 2007; Obleser et al. 2019). In recent years, the validity of the theory and the impacts of acoustic and top-down cognitive effects on the cortical tracking of speech has been extensively studied using continuous speech stimuli. In particular, M/EEG studies has shown that the cortical tracking correlates with speech intelligibility manipulated by distorting or masking the speech stimulus using background noise (Etard et al. 2019b; Iotzov et al. 2019; Lesenfants et al. 2019a; van Canneyt et al. 2021a; Vanthornhout et al. 2018). Similarly, the level at which the stimulus was presented also impacts the neural speech tracking (Verschueren et al. 2021). In turn, tACS studies have shown that stimulating the brain at low frequencies, matched to those involved in cortical tracking, can modulate the participants' speech-in-noise comprehension (Kadir et al. 2019; Keshavarzi et al. 2020a, 2020b; Riecke et al. 2018; Zoefel et al. 2018). The latter suggests that neural oscillation indeed play a causal role in speech processing.

Recent studies have also shown that low-frequency brain activity tracks not only the energy

envelope of speech. In particular, replacing the envelope with a spectrogram-like feature indicating how the energy changes in different frequency bands allows to better predict cortical response than the envelope (Daube et al. 2019; Di Liberto et al. 2015). Fiedler et al. 2017 showed that the acoustic onsets (usually represented as a first derivative of the envelope) also reliably predict neural responses. Moreover, the phonemic features, representing onsets of different phonemes, were identified as good predictors of low-frequency cortical responses to speech (Di Liberto et al. 2015; Lesenfants et al. 2019b). Indeed, invasive brain recordings showed that distinct regions of the auditory cortex were involved in the grouping of local acoustic onsets and other acoustic features (Hamilton et al. 2018). In addition, the low-frequency activity is involved in tracking slowly-fluctuating prosodic features, such as pitch contour (Tang et al. 2017; Teoh et al. 2019). Finally, it is worth noting that although the above-outlined acoustic features share a similar frequency range and some of them are correlated, they tend to explain the unique variability of neural response to speech (Brodbeck et al. 2018a; Daube et al. 2019).

In addition to extrinsic aspects like the objective intelligibility of speech stimulus, the cortical tracking is also impacted by intrinsic factors, such as age and related sensorineural hearing loss (Peelle et al. 2016). In particular, both hearing loss and age, often highly correlated, are somewhat paradoxically associated with an enhanced cortical tracking of speech envelope (Brodbeck et al. 2018b; Decruy et al. 2019, 2020; Fuglsang et al. 2020; Presacco et al. 2016), and longer latency of the underlying neural responses, as compared to normal-hearing listeners (Gillis et al. 2021a). The exact neural mechanisms behind this counter-intuitive finding are not clear. This effect might be due to the age-related decline in neural inhibition, which can lead to more pronounced neural responses (Schmidt et al. 2010). This modulation of cortical tracking is, however, not directly associated with hearing aid usage, as Vanheusden et al. 2020 showed the lack of difference in cortical tracking of clean, unaltered speech in hearing aid users when they were using their devices or not. On the contrary, the neural responses to speech-in-noise are impacted by the use of noise cancellation algorithms often implemented in hearing aids (Alickovic et al. 2020; Fiedler et al. 2021). As such, the modulation of cortical tracking in hearing-impaired participants may not directly correlate with the audibility of speech stimulus but rather with increased cognitive demand while listening to speech masked by background noise.

This chapter covered an introduction to the leading theory of continuous speech encoding in the auditory cortex and how the acoustic properties of input can influence the cortical mechanisms of speech processing. However, as highlighted above and in the previous parts of this introduction, the neural processing of speech is top-down modulated by the higher-level cognitive functions. A review of recent studies and proposed mechanisms for speech-specific cognitive top-down modulation are introduced in the following chapters of this introduction.

1.1.5 Subcortical speech processing

While the cortical responses tend to be studied most often, the role of the lower stages of the auditory pathways in speech processing is investigated as well, especially in recent years. Con-

sidering the contribution of the auditory periphery and brainstem in the bottom-up auditory processing of sound, they have been extensively studied to understand the neural mechanisms underlying hearing impairments (Verhulst et al. 2018). Taking advantage of similarities of the auditory periphery between mammalian species, subcortical mechanisms of bottom-up sensory neural coding of sound are currently relatively well understood (Zilany et al. 2014).

However, as mentioned before, it is not possible to study speech processing in animal models beyond investigating how complex speech stimuli are encoded. Studying continuous speech processing mechanisms of subcortical structures in healthy adults is particularly challenging due to the deep location of the generators, unlike cortical sources, which are in closer proximity to the electrodes placed on the scalp. The acquisition of auditory brainstem responses (ABR) follows a similar protocol as recording cortical ERPs. Evoked ABR is typically obtained by averaging response to many repetitions of short stimuli, such as clicks or pure tones. Since the response source is in the deep brain structures, the SNR of the signal picked up by the scalp electrodes is much lower, and many more repetitions of the stimulus are needed to extract a meaningful evoked response. However, since the latency of the ABR is typically below 10 ms, the rate of stimulus presentation can be much higher, as compared to the recording of cortical ERPs spanning over hundreds of milliseconds.

Due to the above-outlined limitations, studying ABR to complex speech stimuli has been technically challenging. Because obtaining a stable evoked response requires recording hundreds or thousands of repetitions of stimuli, most studies employ short speech stimuli, such as syllables, vowels or short words. Because of the complexity of the stimulus used, as compared to clicks or pure tones, the methodology is referred to as complex ABR (cABR) (Skoe et al. 2010). In comparison to the cortical evoked response to analogous stimuli, the cABR is characterized by much lower latency (below 10 ms) and much higher frequency (Fig. 1.5). In particular, the typical cABR to a syllable can be split into three main components, the onset response, the transient phase associated with the response to the unvoiced consonant (e.g., */b/*, */f/*, */g/*) and the steady-state phase associated with the response to the voiced vowel (e.g., */a/*, */e/*, */o/*). Notably, the steady-state response is phase-locked to the pitch of the speaker's voice (usually between 70 and 300 Hz). Analogous steady-state phase-locking is characteristic of ABR to sustained pure tone, known as frequency following response (FFR). Because of the very low SNR of cABR, its low latency and high frequency, the signal acquisition setup typically requires a high sampling rate and low-noise pre-amplifiers. Unlike cortical responses, the cABR recordings usually employ only a couple of recording electrodes located at the centre of the head, which are referenced to the earlobe or linked mastoids (Skoe et al. 2010).

Although cABR methodology greatly facilitated the investigation of subcortical mechanisms associated with speech processing, the common experimental paradigms involving the presentation of thousands of repeated short stimuli still lack ecological validity. This is particularly limiting for studying the role of higher-level cognitive mechanisms involved in speech processing. While cABR differs depending on the listener's language experience and musical training (Bidel-

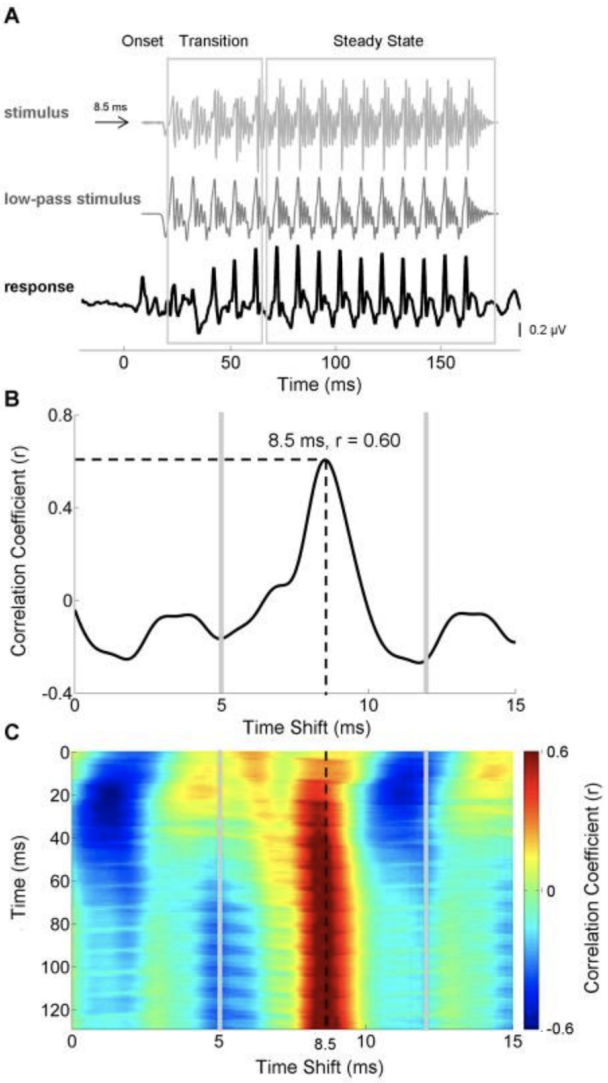

Figure 1.5: Subcortical response to speech. A: The cABR response (black) to a syllable */da/* is compared to the evoking stimulus (top) and its low-pass filtered version. Notice that the stimulus consists of an unvoiced (silent) consonant */d/* and a voiced vowel */a/*. The unvoiced consonant produces the transient response (transition), while the voiced vowel results in a stable steady-state response phase-locked to the speaker's pitch. **B**: Cross-correlation of the low-pass filtered stimulus and the evoked cABR. The correlation peaks at 8.5 ms, reflecting the transition delay associated with the steady-state portion of the response. **C**: Correlation between the evoked cABR response and low-pass-filtered stimulus using overlapped 40-ms windows. The strongest correlation, and thus phase-locking, is obtained for the latter voiced portion of the syllable. Reproduced with permission from Skoe et al. 2010.

man et al. 2011; Krizman et al. 2019), the methodology is not suited for unravelling dynamic online modulation of the response through changes in the non-repetitive acoustic input or top-down cognitive modulation.

In recent years, many studies focused on overcoming the limitation of the cABR by developing methods for detecting brainstem responses to continuous speech in the form of, for example, audiobooks. Forte et al. 2017 was the first to propose a cross-correlation based method for de-

tecting the ABR to natural, uninterrupted speech. In particular, the method uses a fundamental waveform feature extracted from speech through empirical mode decomposition (Huang et al. 2006). The fundamental waveform feature vibrates according to the talker's instantaneous pitch frequency. A similar effect can be obtained by filtering the signal, as depicted in Fig. 1.5A. Since the fundamental waveform represents only the voiced portion of the speech signal, the detected response will reflect the steady-state portion of cABR. Indeed, having cross-correlated the fundamental waveform and the recorded scalp EEG, the authors' showed the highest correlation peak at 9 ms. This result matched the cABR responses to repeated syllables (Skoe et al. 2010) (Fig. 1.5).

An alternative method for detecting the brainstem response to the continuous speech was proposed in Maddox et al. 2018 and further refined and extensively validated in Polonenko et al. 2021. In contrast to Forte et al. 2017, the method uses deconvolution (Lalor et al. 2010) to predict neural responses from half-wave rectified speech signal. The half-wave rectified speech shares similarities with a set of glottal pulses represented as unit responses. The authors' validated their method by detecting click-evoked ABR, which matched the standard approaches obtained through averaging responses to individual stimuli. When applied to neural data elicited by continuous speech stimulus, the method yields a pronounced peak at about 6 ms. Notably, the model could also detect the activity of subsequent parts of the auditory pathways, middle latency responses (between 20 – 50 ms) and late cortical responses (50+ ms).

When applied to detecting brainstem responses to continuous speech, the two above-outlined methods yielded similar results in a systematic comparison (Bachmann et al. 2021; Bachmann et al. 2020). However, unlike the purely pitch-based method from Forte et al. 2017, the deconvolution of half-wave rectified speech (Maddox et al. 2018) involves unvoiced portions of speech. This difference might explain why the latter approach tends to closely resemble click-evoked ABR, while the former is more similar to steady-state FFR to voiced portions of syllables.

Since their invention, the above-outlined approaches, and the novel method introduced in chapter 4 of this thesis (Etard et al. 2019a), have been employed to study subcortical responses to speech. In particular, Saiz-Alía et al. 2019; van Canneyt et al. 2021d showed that ABR to continuous speech is strongly modulated by the voice characteristics. This acoustic modulation of the detected response depended mainly on the talker's pitch range, with the ABR to high-pitched talkers being weaker. In the computational modelling study, Saiz-Alía et al. 2020 showed that the inverse relationship between the speaker's pitch and the strength of the ABR to their voice can originate from decreasing phase-locking capability of the brainstem for higher frequencies (Joris et al. 2013).

Furthermore, the methodology was applied to study the effects of age and hearing loss on the early neural phase-locking to the speaker's pitch. In van Canneyt et al. 2021b the authors showed that the early subcortical tracking of the speaker's pitch (at about 10 ms latency) decayed with age but was not influenced by the degree of hearing loss. This finding is in agreement

with recent evidence for age-related decrease in FFR phase-locking to non-speech stimuli (Anderson et al. 2012). The authors also found a significant contribution from late tracking of the speaker's fundamental frequency at about 40 ms in the hearing-impaired participants. While the mechanism of this somewhat unexpected enhancement of cortical phase-locking to pitch is unknown, the authors propose that it might be modulated through increased listening effort and attention of the hearing-impaired listeners.

While the recently developed methodology for detecting ABR to continuous, uninterrupted speech is a precious tool for understanding mechanisms of natural speech processing, the exact origin of the FFR measured using non-invasive methods is still debated. In particular, Coffey et al. 2016 recently showed the significant cortical contributions to the FFR measured using MEG. FFR has been believed to originate from multiple sources in the brainstem (predominantly inferior colliculus), as evidenced by the source localization studies using M/EEG in humans and single-unit recordings in an animal model (Bidelman 2015; Chandrasekaran et al. 2010; Sohmer et al. 1977). The low latency of the response (<10 ms) and the fact that upon lesioning inferior colliculus, the FFR is eradicated (Sohmer et al. 1977) further support the predominantly subcortical origin of the response. According to the recent opinion paper Coffey et al. 2019, the sources of FFR can span across the auditory system, but the emphasis of different sources might depend on the experimental setup, including neuroimaging modality, sensor layout, referencing and stimuli. In particular, EEG and MEG offer different tradeoffs in terms of detecting responses in cortical and deep regions (Piastra et al. 2021). Bidelman 2018 furthermore showed that the contribution of cortical and subcortical sources to the FFR depends on the stimulus frequency and the presence of cortical sources in the response decays above 100 Hz.

Although the exact source of the FFR is currently debated, a biophysically-plausible computational model of subcortical neural responses to speech introduced in Saiz-Alía et al. 2020 replicated results presented in recent studies investigating ABR to natural speech (Forte et al. 2017; Saiz-Alía et al. 2019). Notably, the model implemented only auditory periphery and subcortical circuits, which suggests the predominantly subcortical source of the ABR to natural speech obtained in the experimental studies.

Despite the possible contribution of cortical sources to the FFR, the recently developed methodology allows monitoring early neural responses tracking the fundamental frequency of the speaker's voice. The possibility of using continuous, uninterrupted speech in the experiments allows designing novel paradigms to study how cognitive processes influence this, predominantly subcortical, response. This line of research is critical for understanding the top-down relationship between cortical and subcortical activity involved in speech processing. Due to the lack of suitable methods for detecting subcortical responses to naturalistic speech stimuli, the functional role of the efferent feedback interactions in natural speech perception has not been extensively studied.

1.1.6 Top-down modulation of neural speech encoding

The previous chapters outlined mechanisms of speech encoding in cortical and subcortical structures. While this bottom-up sensory processing enables speech perception, the top-down modulation of these responses is instrumental for the effective neural encoding of speech and extracting meaning from the utterances. In particular, higher-level cognitive processes related to selective attention and linguistic processing underlie effective neural speech processing.

The importance of selective attention in speech perception has been known for decades, starting with the seminal experiments implementing the cocktail party problem (Cherry 1953). Over the past decades, behavioural studies investigated mechanisms underlying the attention-based filtering of the target talker from the mixture. According to the recent theories, auditory attention starts with the formation of auditory objects and source segregation (Bregman 1994). While the exact definition of an auditory object is still debated (Griffiths et al. 2004), the sounds that are co-modulated or harmonically related tend to be grouped together. Notably, speech exhibits both of these properties, and thus individual voices at a *cocktail party* tend to be grouped as separate objects (Remez et al. 1994).

The exact mechanisms of efficient stream segregation in the human auditory system are not fully understood. However, it is known that neural spectro-temporal filters implemented across the human auditory pathways contribute to untangling the neural representations of mixed sources (Elhilali et al. 2008). The success of source separation, in terms of understanding target talker, depends on a range of factors, such as spatial separation of the sources (in binaural hearing), their relative loudness or differences in pitch, rhythm and timbre (De Cheveigne 2005; Fawcett et al. 2015). Notably, in some severely noisy auditory scenes, or in the case of hearing impairment, the reliable separation of sources might not be possible. Following the segregation of sources, the attentional focus is defined by the saliency of the sources (bottom-up, exogenous attention) and the listener's intention (top-down endogenous attention) (Fawcett et al. 2015), however, it can also be influenced by working memory. Importantly, according to the object-based mechanism, attention itself impacts the object formation (Fritz et al. 2007; Shamma et al. 2011; Shinn-Cunningham 2008). As such, the acoustic cues facilitating the source segregation will also impact the endogenous attentional selection (i.e., it is difficult to focus on one of two identical voices without spatial cues).

Assuming the successful formation and segregation of auditory objects, attention provides a top-down sensitivity control, which enhances the neural representation of the attended object, with respect to the others in the auditory scene (Scharf et al. 1987). However, the exact mechanism through which the enhancement of the target object occurs remains unknown. Currently, the two main models for attentional filtration are debated. In the *early* model, the attended object is filtered at the early stages of the auditory processing, and thus only its neural representation persists. According to the *late* model, auditory objects are processed in parallel to produce their neural representations and attention enhances only that of the attended auditory object. While both of the theories are supported by experimental evidence, the *perceptual load*

theory proposes both of them being involved in the selective attention (Lavie 1995). In particular, it postulates that the attentional mechanism depends on the perceptual demands of the task. When there are only a few sources in the auditory scenes, the *late* attentional selection is implemented. However, with the increasing number of sources and/or additional background noise, the *early* attentional selection can be involved to attempt to filter out irrelevant objects early and thus facilitate the segregation of relevant auditory objects.

In recent years, many studies investigating the neural correlates of attention in natural speech processing focused on cortical tracking of speech (Brodbeck et al. 2018a; Ding et al. 2013; Golumbic et al. 2013; Mesgarani et al. 2012, 2014; O'Sullivan et al. 2015) (Fig. 1.6). Recent experimental evidence showed that slow neural oscillations in the theta and delta frequency bands phase-lock to the envelope of the attended speaker more than to that of the ignored talker(s) (Ding et al. 2012, 2014; Golumbic et al. 2013, 2012). In turn, the neural response to continuous speech estimated using TRF methodology shows that selective attention significantly alters the amplitude of the response, its latency and leads to the emergence of additional components (Brodbeck et al. 2020a; Fiedler et al. 2019). In particular, it is well established that in a typical two-talker cocktail party scenario, early components of the neural response, often associated with acoustic processing, are represented for both attend and ignored talkers and most of the attentional modulation is associated with later components (Brodbeck et al. 2018a, 2020a; Fiedler et al. 2019). Notably, the attentional modulation of cortical tracking of speech is robust enough to allow decoding of attentional focus using only short segments of brain activity recordings (Das et al. 2020; Fuglsang et al. 2017; Geirnaert et al. 2021b; Mirkovic et al. 2015).

Although many studies investigated changes in the neural responses to attended and ignored talkers, it is currently not clear whether the neural representation of the attended talker is enhanced or that of the ignored talker suppressed. Results from an EEG study reported in Hausfeld et al. 2018 indicate that neural tracking of the attended talker is enhanced since the representation of the ignored talker(s) is more similar to the mixture of ignored talkers, rather than their voices in isolation. However, Fiedler et al. 2019 showed that in the case of acoustically dominant ignored talker, additional late components emerge in the neural response (Fig. 1.6C). This suggests the existence of active mechanisms for suppressing the interferer. Recently Keshavarzi et al. 2021 conducted a tACS study, in which the neural activity of participants listening to a two-talker cocktail party was perturbed. The stimulation signal was either derived from the speech envelope of the attended or the ignored talker. Under either stimulation protocol, the participants' comprehension was significantly modulated. That suggests the existence of cortical representation of both voices and supports the notion of *late* attention mechanism.

Like selective attention, linguistic processing, fundamental for extracting meaning from the utterance, is commonly associated with a top-down influence from higher-level language-specific areas and modulation of late components of neural responses to speech (Hickok et al. 2007; Kutas et al. 2011). In recent years, many studies showed that slow neural oscillations track linguistic features, such as onsets of words, their semantic similarity or context-independent

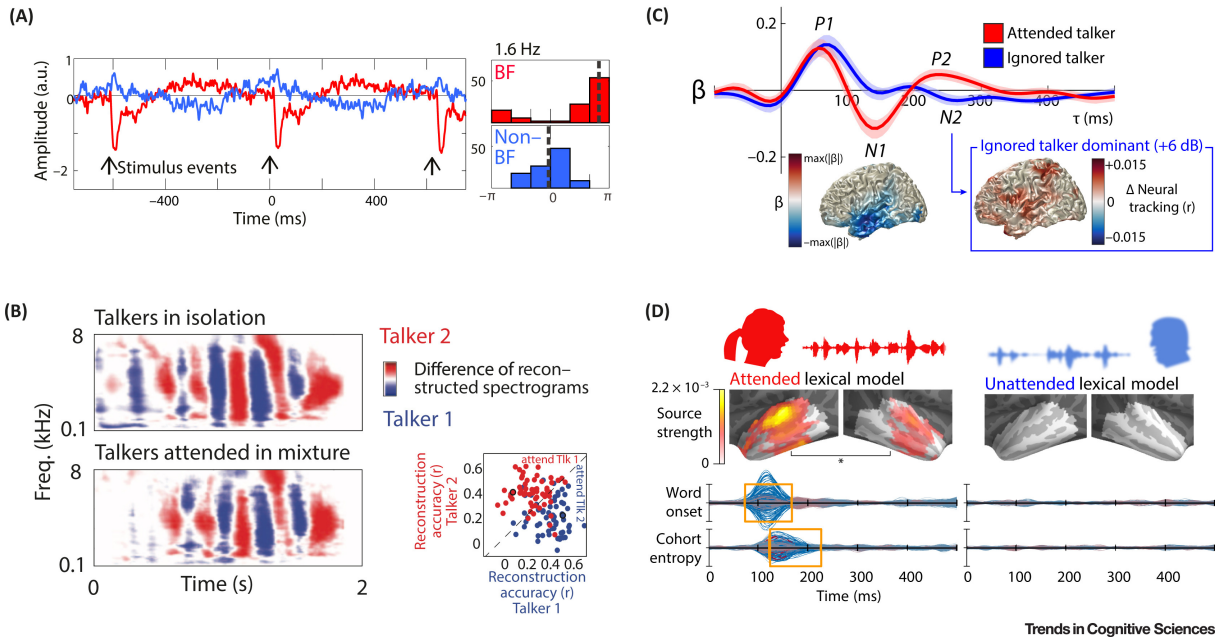

Figure 1.6: Attention modulation of cortical speech encoding. (A) In the macaque primary auditory cortex (A1), neural entrainment acts as a spectrotemporal filter. Activity in the sites responding to the attended stimulus (red) produces a strong response and are in anti-phase to those tuned to respond to a different, unattended stimulus (blue) (Lakatos et al. 2013). **(B)** The spectrograms of the talkers in a *cocktail party* can be reconstructed from the gamma-band activity recorded using ECoG over the temporal lobe. The representation of the talkers' voices is similar when presented in isolation (top) and together (bottom). Inset: The reconstruction of the speakers' voices from ECoG depends on the attentional focus of the listener (red: attend talker 2; blue: attend talker 1) (Mesgarani et al. 2012). **(C)** Selective attention significantly impacts the EEG temporal response function. The early neural response between 100-200 ms, originating from the temporal cortex (lower left), occurs for both attended and ignored talkers. However, when the ignored talker is acoustically dominant, in terms of SNR, an additional response component (N2) emerges, which has contributions from other brain regions, including frontoparietal areas (blue inset). This emergence of the additional component is also associated with improved neural tracking of the ignored talker (Fiedler et al. 2019). **(D)** Selective attention influences tracking of the acoustic properties of the voices and simple linguistic features, such as word onsets. It also significantly modulates the representation of the complex linguistic features derived from the local context, such as cohort entropy (i.e., how uncertain a phoneme is, based on the previously heard phonemes). The effect of attention is largely binary, with the representation of the linguistic features being only present for the attended talker (Brodbeck et al. 2018a). Reproduced with permission from Obleser et al. 2019.

word frequency, as well as context-dependent features quantifying the uncertainty of predicting a word from its local context (Brodbeck et al. 2022, 2018c; Broderick et al. 2018; Gillis et al. 2021b; Koskinen et al. 2020; Weissbart et al. 2020). While neural tracking of speech envelope and phonetic features can predict comprehension and listening effort (Decruy et al. 2020; Etard et al. 2019b; Iotzov et al. 2019; Lesenfants et al. 2019a), the neural representation of linguistic features reflects the second language proficiency (Di Liberto et al. 2021). Notably, the cortical tracking of linguistic features is strongly modulated by attention. In particular, Brodbeck et al. 2018a showed that only linguistic features associated with the attended talker are tracked by

the low-frequency neural activity measured using MEG (Fig. 1.6D). This suggests that while acoustics of both the attended and ignored voices are segregated and represented at the cortical level, only the target speaker, on which the listener is focused, undergoes linguistic encoding.

While the abundance of evidence for the top-down attentional modulation of cortical responses to speech exists, the functional role of efferent corticofugal pathways, projecting from cortical areas to the brainstem, in speech processing remain poorly understood. In particular, the attentional modulation of evoked auditory brainstem responses has been historically debated due to many inconclusive or even contradictory results (Brix 1984; Galbraith et al. 2003; Lehmann et al. 2014; Varghese et al. 2015). These discrepancies in results reported to date may partially be attributed to the suboptimal experimental paradigm, which involved rapid repetitions of short stimuli. This might make it challenging for the participants to sustain their attention during the experiment. Furthermore, a high rate of stimulus presentation might lead to adaptation and gradual reduction of the response amplitude (Neupane et al. 2014).

While previous studies sought modulation of the recorded evoked response as a function of attention, Price et al. 2021 investigated the functional connectivity between the sources located in the primary auditory cortex and the brainstem (Fig. 1.7). In particular, the experimental protocol involved a presentation of repeated vowels to participants whose EEG was recorded throughout the experiment. The participants were asked to either focus on the listening task or passively listen to the stimuli. Furthermore, the authors' added background noise to the vowels to assess the functional role of corticofugal pathways in noise suppression. The functional connectivity, representing top-down corticofugal modulation, significantly decreased when participants were passively listening to the stimuli masked with background noise. This finding supports the existence of top-down attentional modulation of the auditory brainstem responses by cortical sources to support speech processing in adverse conditions. However, the use of an isolated vowel as a repeatedly presented stimulus has the same limitations as previous cABR studies, and thus the effects might not translate to realistic auditory scenes.

To overcome the limitations of previous studies using short repeated stimuli, the methods for modelling subcortical response to continuous speech stimuli have been developed (Etard et al. 2019a; Forte et al. 2017; Maddox et al. 2018; Polonenko et al. 2021; van Canneyt et al. 2021c). In particular, naturalistic, non-repetitive speech stimuli, such as audiobooks, alleviate the problem of maintaining participants' focus and prevent neural adaptation. These methods have been applied to study attentional modulation of subcortical responses to continuous speech in a classic two-talker cocktail party (Forte et al. 2017; Maddox et al. 2018). In particular, Forte et al. 2017 found a significant attentional modulation of the auditory brainstem response phase-locked to the speaker's pitch. However, Maddox et al. 2018 did not find a significant difference in the response estimated using the linear decorrelation method. The source of this discrepancy might originate from the inclusion of the unvoiced parts of speech in the latter method. In particular, the response detected in Forte et al. 2017 represented the steady-state portion of the ABR phase-locked to the speaker's pitch. Maddox et al. 2018, however, included both

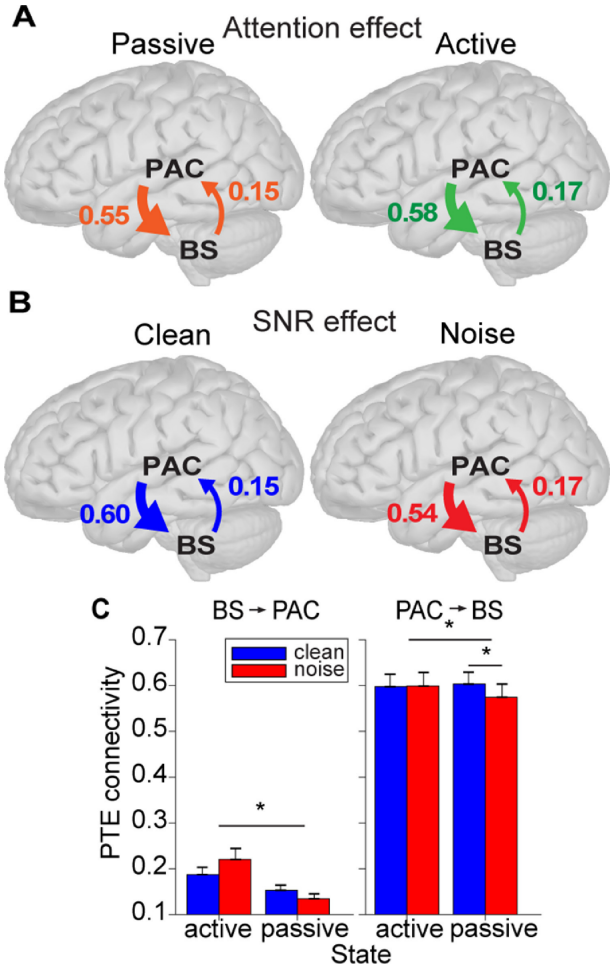

Figure 1.7: Modulation of subcortical responses to speech through corticofugal pathways. In Price et al. 2021, the authors used phase transfer entropy (PTE) to measure nonlinear, directed (causal) signal dependency between the sources in the primary auditory cortex (PAC) and in the brainstem (BS). They applied this methodology to investigate the effects of attention and background noise in the EEG study, which involved attending (or not) to a stream of repeated vowels. The significant bidirectional communication was found in all considered conditions of the experiment, when the participants attended the task or were passively listening (**A**, active vs passive), as well as when the target stimulus was clean or masked with background noise (**B**, clean vs noise). Bottom-up (BS $\longrightarrow$ PAC) connectivity was modulated by the listeners' attention (**C**, left). The top-down connectivity (PAC $\longrightarrow$ BS), reflecting corticofugal pathways, significantly decreased only during passive listening (**C**, right). Error bars represent 1 standard error of the mean, and asterisks denote significant differences ($p < 0.05$). Reproduced with permission from Price et al. 2021.

voiced (with pitch) and unvoiced (without pitch) parts of speech to fit their decorrelation models. As a result, the two methods might emphasize different portions of the brainstem response to speech (see cABR in Fig. 1.5), which might be differently affected by top-down attentional modulation. Specifically, since the phase locking of subcortical responses tends to be weaker for high-frequency stimuli (Joris et al. 2013; Saiz-Alía et al. 2020), the tracking of high-frequency, unvoiced parts of speech and its attentional modulation might be accordingly smaller and thus difficult to detect in non-invasive scalp recordings. Notably, Etard et al. 2021 considered the

two-talker cocktail party with continuous musical pieces played by different instruments and did not find a significant attentional modulation of the neural response tracking the instruments' pitch.

While the debate on the exact mechanisms of efferent top-down attentional modulation of the subcortical responses is still ongoing, studies such as Forte et al. 2017; Price et al. 2021 provide convincing evidence for the existence of functional feedback between cortical and subcortical structures through corticofugal pathways. However, what other types of top-down modulation might be implemented through analogous feedback mechanisms remains unknown. Studies in an animal model found the correlates of prediction errors in the subcortical structures (Parras et al. 2017). A similar study found the significant modulation of the human frequency following response by the expectancy of the stimulus (Slabu et al. 2012). However, this effect could not be replicated in Font-Alaminos et al. 2021. While inconclusive, the above findings might suggest the active role of the auditory brainstem in the predictive coding, which is hypothesized to be crucial for the neural processing of language (Lewis et al. 2015). To date, no other studies attempted to study early neural correlates of language processing in the subcortical responses, or very early high-frequency cortical response, to continuous speech.

1.2 Aims and thesis outline

Although decades of research contributed to great progress in understanding the neural processing of sound in humans, the neural mechanisms underlying speech perception are still not fully understood, especially those related to top-down cognitive modulation. In this thesis, we aimed to develop computational models characterizing neural mechanisms of speech processing across a broad spectrum of neural dynamics, including slower and lower-frequency cortical responses, as well as rapid high-frequency activity, of predominantly subcortical origin. To achieve that, we used tACS and EEG to, respectively, perturb and record the neural activity of young normal-hearing volunteers listening to speech. The proposed modelling frameworks involve theoretical models, based on the leading theory on how the brain processes speech, and data-driven models, optimized directly from the experimental data. In particular:

Chapter 2 introduces a tACS study in which we aimed to investigate whether neural oscillations in the theta and/or delta frequency ranges play a causal role in speech-in-noise comprehension. Despite many previous neuroimaging studies, it is still not clear whether oscillations, in either or both frequency ranges, actively facilitate speech in noise comprehension. During the experiment, young and healthy participants were listening to spoken sentences masked by background noise. At the same time, we applied tACS stimulation derived from the envelope of the target talker over their auditory cortices. By using stimulation waveform filtered in theta and delta frequency ranges, we investigated whether either of them produced a consistent modulation of speech-in-noise comprehension scores across the population of participants. We found that only stimulation in the theta frequency range yielded significant modulation of speech comprehension scores in our cohort of volunteers. The modulation of the listeners' comprehension

was dependent on the phase of the stimulation waveform, with respect to the acoustic stimulus. Importantly, the modulation was consistent across participants and significantly improved the speech-in-noise comprehension above the unrelated *sham* stimulation.

Although tACS is a popular tool for studying neural mechanisms of speech perception, the mechanisms through which weak electrical current interacts with neural circuits for speech processing and give rise to behavioural effects are not understood. In **chapter 3**, we proposed a spiking neural network model for studying the effects of tACS on the cortical encoding of speech in noise. The preliminary model simulations were used to define hypotheses for the experiment from chapter 2, as well as to optimize the stimulation protocol to maximize its efficacy. Following the conclusion of the study from chapter 2, we simulated the experimental setup in the model to compare its prediction and the experimental results. The speech in noise encoding of sentences in the model decayed in a sigmoidal fashion across SNRs, which reflected speech in noise comprehension of normal-hearing adults. Furthermore, the effects of the external tACS stimulation on the model's encoding capability matched the experimental findings from not only ours (Keshavarzi et al. 2020a) but also other recent tACS studies attempting to modulate speech comprehension (Kadir et al. 2019; Keshavarzi et al. 2020b; Wilsch et al. 2018). These findings support the claim that tACS directly influences cortical oscillations, which are indeed actively involved in cortical speech processing.

While chapters 2 and 3 investigated cortical mechanisms of speech processing through cortical oscillations, **chapter 4** proposed a complex computational modelling method for detecting early high-frequency neural response, of predominantly subcortical origin. The methodology extends the framework proposed in Forte et al. 2017 to high-density EEG setups commonly used for studying cortical mechanisms of speech processing. The response detected by the model had high frequency (above 70 Hz) and low latency (approx. 10 ms), which suggests the subcortical origin, in agreement with previous studies. Unlike previous methods using sparse ABR recording setups, the high-density EEG allows studying the topography of the detected response. We employed the proposed model to decode participants' auditory attention in a two-talker *cocktail party* from their EEG recordings. We found that the response exhibited a stronger phase locking to the pitch of the attended talker. Comparing the models reflecting the response to attended and ignored voices, we did not find significant attention-based gain modulation, but rather changes in the response phase. We have shown that the proposed model-based attention decoder used smaller decision windows while maintaining comparable accuracy to the conventional decoders based on the cortical responses. These results support the claims of top-down attentional modulation of the subcortical or early cortical response through feedback loops in the human auditory system. Moreover, our results shed light on how top-down-modulated subcortical speech processing may contribute to effectively solving the *cocktail party* problem.

In **chapter 5** we extended the modelling framework developed in chapter 4 to study early high-frequency responses to individual words in a continuous narrative. In particular, we used the high-density EEG dataset from Weissbart et al. 2020 to investigate whether the neural

activity in question is modulated by acoustic and linguistic word-level features. The acoustic features were derived from the fluctuations in the speaker's pitch. The linguistic features were derived from the text read by the speaker and included context-independent word frequency (i.e., how common the word is) and context-dependent word surprisal and precision. The latter features were based on the conditional probability of the word given the past context and represented how unexpected the word is (surprisal) and the confidence about predicting it given previous words (precision). We found that word-level neural response at the fundamental frequency was predominantly modulated by the acoustic features and, to a lesser extent, by the context-independent word frequency. Context-dependent linguistic features did not modulate the responses. Our results illustrate that the early neural responses are modulated not only by different voices (Saiz-Alía et al. 2019, 2020; van Canneyt et al. 2021d) but also individual words produced by the same talker. The significant modulation of the response through one of the linguistic features supports the existence of top-down linguistic modulation of neural activity of predominantly subcortical origin, or early, high-frequency cortical responses.

Finally, **chapter 6** summarizes the work presented in this thesis, discusses its broader impact on the field and proposes future work directions.

Chapter 2

Transcranial alternating current stimulation in the theta band but not in the delta band modulates the comprehension of naturalistic speech in noise

The work presented in this chapter has been previously published as Keshavarzi et al. (2020a). This work was tightly coupled with the development of the computational model introduced in Chapter 3. In particular, the preliminary results from the model simulations were used to formulate the experimental hypothesis, and optimize the tACS stimulation protocol to maximize its efficacy. In turn, the experimental data collected here were used to validate the model. This work was conducted in collaboration with Dr Mahmoud Keshavarzi, who led the collection of behavioural data presented in this study.

2.1 Introduction

Speech is a complex signal that unfolds over several temporal scales, from phonemes to syllables, words, and phrases. The neural activity in the auditory cortex entrains to the amplitude modulations in speech, as well as to more specific speech structures such as phonemes, the onset of words, and to higher-level linguistic information such as surprisal of word sequences and syntactic structure (Brodbeck et al. 2018a; Broderick et al. 2018; Ding et al. 2016, 2012; Giraud et al. 2012; Lakatos et al. 2005; Weissbart et al. 2020). This cortical entrainment has recently been shown to play a functional role in speech processing. In particular, transcranial alternating current stimulation, paired to rhythmic speech, modulated neural responses that correlated with behaviour when speech was intelligible, but not when it was unintelligible (Zoefel et al. 2018). Moreover, transcranial alternating current stimulation with the speech envelope was found to modulate the comprehension of degraded speech (Kadir et al. 2019; Riecke et al. 2018; Wilsch et al. 2018). However, it remains unclear which more specific aspects of the cortical speech entrainment underlie the modulation of speech comprehension.

Two main frequency bands dominate the neural speech entrainment. First, cortical activity in the theta frequency band (4–8 Hz) tracks the onset of syllables which may aid the parsing of a speech stream (Di Liberto et al. 2015; Ding et al. 2014). A computational model of theta oscillations coupled to gamma oscillations showed indeed that the entrainment of theta activity to a

speech signal can act as an efficient parser of syllables, and that the connected gamma network can encode speech efficiently (Hyafil et al. 2015). Second, cortical activity in the delta band (1–4 Hz) entrains to the onset of words in natural speech and has been found to encode both syntactic as well as semantic information (Broderick et al. 2018; Ding et al. 2016; Weissbart et al. 2020).

Much effort has been devoted to tease apart the roles of cortical entrainment in the delta and in the theta band for speech processing (Ding et al. 2014; Kösem et al. 2017). In particular, an MEG investigation into speech with a degraded spectro-temporal fine structure found that the neural entrainment in the delta, but not in the theta, band correlated with speech comprehension (Ding et al. 2014). We have recently employed an experimental paradigm with native and foreign speech in different levels of background noise that allowed us to tease apart the effects of lower-level acoustics and higher-level comprehension, demonstrating that speech acoustics related mostly to theta-band activity and comprehension to delta-band entrainment (Etard et al. 2019b). These findings agree with a role of the theta band in tracking lower-level acoustical structures such as syllable onsets, and a role of the delta band in entraining to higher-level linguistic features such as semantic and syntactical structures (Brodbeck et al. 2018a; Broderick et al. 2018; Ding et al. 2016; Weissbart et al. 2020). However, the distinct roles of both frequency bands to the modulation of speech comprehension through neurostimulation have not yet been investigated.

Here we combined transcranial alternating current stimulation with a behavioural task of speech-in-noise comprehension to tease apart the individual contributions of the delta- and theta band entrainment to speech processing. In particular, we presented young adult participants without hearing impairment with semantically unpredictable sentences that were embedded in speech-shaped noise, such that subjects understood roughly 50% of the words correctly (Fig. 2.1). Simultaneously to the sound presentation, we stimulated both their left and right auditory cortex symmetrically through small alternating electric currents that were applied through scalp electrodes (transcranial alternating current stimulation or tACS). The current signal was obtained from the envelope of the simultaneously-presented speech signal. To distinguish between the roles of delta- and theta-band entrainment, we filtered the speech envelope in both frequency bands. We hypothesized that the theta-band and delta-band stimulation would modulate speech comprehension in different ways, since the theta-band stimulation would act on the lower-level acoustic processing while the delta-band stimulation would relate to higher-level linguistic information.

Previous investigations of the modulation of speech comprehension through neurostimulation have partly employed speech that was artificially produced to exhibit a rhythm at a particular frequency (Riecke et al. 2018; Zoefel et al. 2018). These studies then employed an alternating current at the same frequency, and investigated how phase differences between the current and the speech affected comprehension. Alternatively, previous studies used naturalistic speech, the envelope of which had a broad spectrum, and then considered a current waveform that mimicked the speech envelope, but was shifted by different temporal delays (Riecke et al. 2018; Zoefel et al.

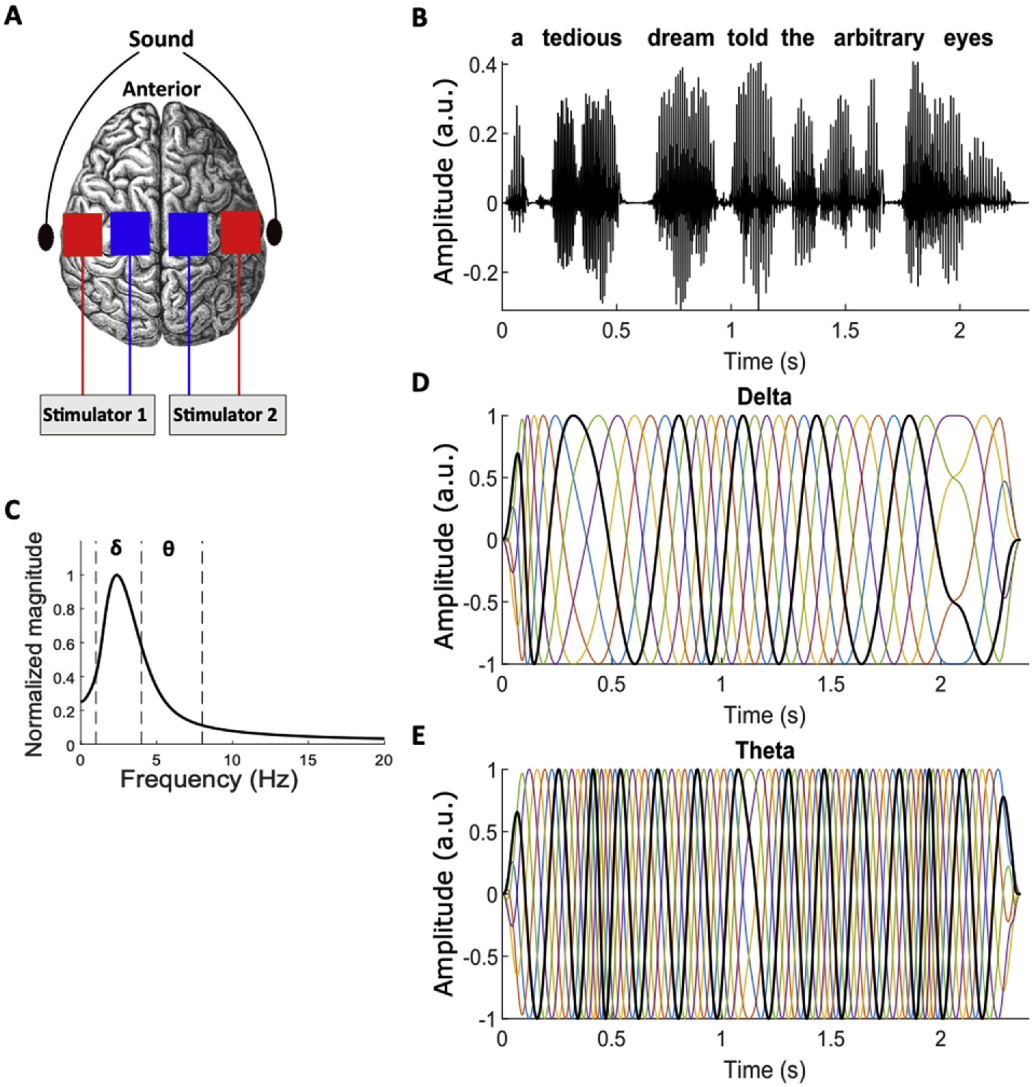

Figure 2.1: The experimental design. (**A**) Participants listened to a sentence embedded in speech-shaped noise. Transcranial alternating electrical current was simultaneously applied symmetrically to both hemispheres through electrodes located over the temporal areas (T7, T8, red) as well as adjacently left and right of the vertex (Cz, blue). (**B**) Each sentence lasted around 2 s. (**C**) The spectrum of the envelope of the sentences (computed from averaging over 1000 sentences) was dominated by the delta frequency band, but also contained significant contributions from the theta band. (**D**, **E**) We employed current waveforms that followed the speech envelope but were filtered in the delta band (**D**) or the theta band (**E**). The resulting waveforms were then shifted by different phases (different colours, black corresponds to no phase shift). The waveforms were further processed so their maxima all had the same value, and such that the values of the minima were all equal as well, except those near the beginning or end of the sentence.

2018).

Because we sought to investigate the influence of the neurostimulation in the delta and theta band on speech comprehension, we presented subjects with naturalistic sentences that had significant amplitude modulation in both the delta and the theta frequency range (Fig. 2.1). We

concurrently applied transcranial alternating current stimulation with a waveform that corresponded either to the delta-band portion of the speech envelope or to the theta-band portion of the speech envelope. However, particular care needs to be taken in the analysis of the resulting effects on speech comprehension to avoid analytical bias and false positive results (Asamoah et al. 2019). To avoid such analytical bias, we used various phase shifts instead of temporal shifts of the current signal. In contrast to temporal changes, phase shifts lead to circular changes of the signal that allowed us to employ powerful methods from Fourier analysis to determine the modulation of speech comprehension.

2.2 Methods

2.2.1 Participants

Eighteen native English speakers took part in the experiment (9 females, 8 males, aged between 18 and 29 years, mean age 23 years, standard deviation 3.3 years). All reported normal hearing, had no history of mental health problems or neurological disorders, and were right-handed according to their own assessment. All participants gave informed consent. The experiment was approved by the Imperial College Research Ethics Committee. One female participant did not complete the study due to problems with the electrode attachment.

2.2.2 Hardware setup

A PC with a Windows 7 operating system was used to generate the acoustic stimuli and the current waveforms digitally. Both signals were synchronized on the PC, and were then converted to analogue signals using a USB-6212 BNC device that kept the temporal alignment between the two signals (National Instruments, U.S.A.). The current waveforms were fed to a splitter connected to two neurostimulation devices (DC-Stimulator Plus, NeuroConn, Germany). The acoustic stimuli were passed through a soundcard (Fireface 802, RME, Germany) connected to earphones (ER-2, Etymotic Research, U.S.A.). The temporal alignment of the resulting sound signal to the current waveform was verified by measuring both signals simultaneously, which showed that the timing of both signals differed by less than 1 ms.

2.2.3 Acoustic stimuli

The acoustic stimuli used in the experiment were single sentences presented in speech-shaped noise. The sentences were semantically unpredictable and were generated using Python's Natural Language Toolkit (Beysolow II 2018; Bird et al. 2009). Each sentence (e.g. "*The current months solve the important trial.*") consisted of seven words, including five key words used to evaluate the participant's level of comprehension. The sentences were converted to an audio stimulus using the TextAloud software with a male voice and with the sampling rate of 44,100 Hz. The speech signal was presented at an intensity of 65 dB SPL which provided a comfortable

sound level.

The speech-shaped noise was generated by determining the average Fourier transform of the different sentences. The phases of the spectral components were then randomized while the magnitude was kept. The noise was then obtained by the inverse Fourier transform of the resulting randomized signal.

2.2.4 Neurostimulation waveforms

We presented subjects with speech signals and concurrently applied transcranial alternating current stimulation. For the latter we employed 15 different waveforms. One waveform was designed to provide a sham stimulus. This current started at the beginning of the speech signal but lasted only 500 ms. Smooth onsets and offsets were produced through ramps of a duration of 100 ms. This sham stimulation was used to mimic the current delivery, in particular the attachment of the scalp electrodes. It could in principle also control for a brief skin sensation resulting from the current, although, as described below, we adjusted the current magnitude such that subjects did not experience a skin sensation.

The other 14 waveforms were all based on the speech envelope. The latter was computed as the absolute value of the analytical signal of the speech. The speech envelope was then band-pass filtered into the delta frequency band (zero phase IIR filter, low cutoff (-3 dB) 1 Hz, high cutoff (-3 dB) 4 Hz, order 6). The envelope was also band-pass filtered into the theta frequency band (zero phase IIR filter, low cutoff (-3 dB) 4 Hz, high cutoff (-3 dB) 8 Hz, order 6). The band-pass filters implied that both waveforms had a mean of 0.

To enhance the influence of the current signal on the neural entrainment, the waveforms were then processed to boost all maxima and minima in the waveform to the maximal (minimal) value that was encountered in the signal. This was done by computing the analytical (complex) signal through the Hilbert transform, by subsequently setting the amplitude to unity, and by then taking the real part of the obtained function.

The waveforms in both the delta and theta frequency band was then shifted by the six phases 0°, 60°, 120°, 180°, 240° and 300°. A shift by a phase ϕ was implemented by first computing the analytical signal of the band-pass filtered envelope, followed by multiplication by $e^{i\phi}$ (where ϕ has been converted to radians) and by taking the real part of the obtained signal. Because $e^{i(\phi+2\pi)} = e^{i\phi}$, this procedure ensured the circularity of the phase shifts, despite the broad frequency range of the speech envelope. In particular, a shift by a phase of $\phi + 360°$ (where ϕ is measured in degrees again) yielded the same signal as a shift by phase ϕ.

The six phase shifts of both the delta- and the theta-band envelope yielded twelve waveforms. We furthermore employed a delta-band and a theta-band envelope that were obtained from an unrelated sentence, yielding two more current waveforms.

2.2.5 Experimental setup and procedure

The participants were seated in a soundproof room. The sound was presented diotically through earphones (ER-2, Etymotic Research, U.S.A.). Two rubber electrodes were placed adjacently left and right of the location Cz of the subject's head, and the remaining two at the locations T7 and T8 of the International 10-10 system (Fig. 2.1A). One electrode near Cz and the one at T7 were connected to one neurostimulation device, and the remaining electrodes to the second device. Based on simulations of electrical field distribution in a standard human head model and previous experimental studies, such a configuration of electrodes induces strong modulation of the auditory cortices (Herrmann et al. 2013; Riecke et al. 2018; Wilsch et al. 2018; Zoefel et al. 2018). The electrodes at the temporal areas served as the anodes and the electrodes at Cz as the cathodes. The electrodes were covered by 35 cm^2 sponge pads moistened by a 0.9% saline solution (about 5 ml per side). After placing them on the participant's head, the impedance between electrodes of each device was set to below 10 kΩ.

To measure the maximum magnitude of the stimulation current to be used for a participant, a pure sinusoidal signal at a frequency of 3 Hz and with a duration of 5 s was presented to the subject. The signal amplitude was increased from 0.1 mA to a maximum of 1.5 mA in step sizes of 0.1 mA. To minimize the transcutaneous effects of tACS, the procedure was stopped when the participant reported a skin sensation, and the amplitude of the previous step was selected as the maximum threshold for the stimulation current for that participant. The maximal currents that we thereby estimated for the different participants were in the range of 0.7–1.3 mA, with a mean of 1.1 mA and a standard deviation of 0.3 mA.

For each participant, we first measured the sentence reception threshold (SRT) of 50%, that is, the signal-to-noise ratio at which speech comprehension was 50%. During the measurement the participants were subjected to sham stimulation at the onset of each sentence. To estimate the SRT, we employed an adaptive procedure (Kaernbach 2001; Kollmeier et al. 1988). We started with an initial SNR that was randomly selected between 0 dB and -3 dB. If the subject understood at least three key words in the sentence correctly, the SNR value was decreased by 1 dB for the subsequent sentence. The SNR was increased by 1 dB otherwise. The adaptive procedure was stopped after seven reversals in the SNR or after 17 sentences. The adaptive procedure was carried out four times for each subject. The subject's SRT was computed as the average of the last three SNRs that were employed in each of the different runs of the adaptive procedure, with the exception of those of the first run. The so-established SRT was then used as the SNR for the subsequent measurements.

We then measured subjects' speech comprehension during concurrent transcranial alternating current stimulation with 15 different waveforms. For each waveform we therefore presented each subject with a total of 25 sentences in speech-shaped noise, at the SNR corresponding to the personalized SRT that was measured earlier, and applied the current stimulation simulta-

neously. After listening to each sentence, the subject repeated what he or she understood. The response was recorded through a microphone and manually graded by the experimenter for the percentage of correctly understood words. A total of 375 sentences was presented in two different testing sessions that took place on two different days. Which of the 15 different waveforms was used for the current stimulation varied randomly from sentence to sentence and was unknown to both the experimenter and the subject (double blind design). After every 50 sentences the subject took a 2-min break.

2.2.6 Statistical analysis

To investigate the modulation of speech comprehension through both the delta- and the theta-band neurostimulation, we shifted the envelope in each of the two frequency bands by six different phases (0°, 60°, 120°, 180°, 240° and 300°). Each phase shift can modulate the cortical entrainment in the respective frequency band differently: a particular phase shift may, for instance, increase the cortical entrainment whereas another one may diminish it (Riecke et al. 2018; Zoefel et al. 2018). Importantly, although the band-pass filtered envelopes did contain a range of frequencies, the phase shifts were applied in such a way that they were nonetheless cyclical. In particular, a phase change of 360° corresponded to no phase change at all (0°).

If the current stimulation affected speech comprehension, the latter would depend in a cyclical manner on the phase of the current stimulation. In contrast, a finding of no dependence of speech comprehension on the neurostimulation phase would signal that there is no influence of the stimulation, and hence no impact of the neural entrainment on speech processing. We therefore measured the comprehension scores of volunteers and analyzed their dependence on the phase of the current stimulation.

We performed this analysis separately for the current waveforms filtered in the delta and in the theta frequency band. Because we measured the comprehension scores at different phase shifts, the circularity of the phase, and the resulting circularity of the dependence of the speech comprehension on the stimulation phase, meant that the data could be analyzed using the Discrete Fourier Transform. In particular, the data could be written as a discrete sum of cosine functions, each with a particular period that was the either the largest-possible period of 360° or a fraction of 360°. Because we measured speech comprehension at six different phases $\{\phi_k\}_{k=1}^{6}$, the Discrete Fourier Transform implied that the dependence of the speech comprehension score $CS(\phi_k)$ on the phase ϕ_k of the current stimulation could be written as:

$$CS(\phi_k) = \sum_{n=0}^{5} a_n e^{in\phi_k} \tag{2.1}$$

with the complex Fourier coefficients a_n. Four of these coefficients are related through complex conjugation: $a_4 = a_2^*$ and $a_5 = a_1^*$. Let $\frac{A_1}{2}$ be the magnitude of the complex coefficient a_1, and $-\Phi_1$ its phase: $a_1 = \frac{A_1}{2}e^{-i\Phi_1}$. The coefficient a_5 follows via complex conjugation. The coefficient a_2 can be expressed analogously through its amplitude and phase as $a_2 = \frac{A_2}{2}e^{-i\Phi_2}$.

The coefficient a_4 follows as the complex conjugate. The two coefficients a_0 and a_3 are real: they denote a constant offset respectively a contribution that alternates at +1 and -1. They are therefore entirely defined by their magnitudes A_0 and A_3, respectively: $a_0 = A_0$ and $a_3 = A_3$. Because the six discrete phase values $\{\phi_k\}_{k=1}^{6}$ at which we have assessed speech comprehension lead to $e^{i6\phi} = 1$, we have $e^{in\phi} = e^{i(n-6)\phi}$ and can therefore write Equation 2.1 as:

$$CS(\phi) = A_0 + A_1 cos(\phi - \Phi_1) + A_2 cos(2\phi - \Phi_2) + A_3 cos(3\phi) \tag{2.2}$$

The model parameters A_1, A_2 and A_3 hereby denote the amplitude of the variation at the periods 360°, 180° and 120°, respectively. The phases Φ_1 and Φ_2 are the phase shifts at the two longer periods. Because the shortest period corresponds to the Nyquist frequency, it does not allow the inference of a phase shift. A_0 denotes a constant offset. The resulting number of parameters is six, matching the number of phase shifts at which comprehension scores are measured.

We determined the offset A_0 from the mean comprehension score. The modulation amplitudes A_1, A_2 and A_3 as well as the phase shifts Φ_1 and Φ_2 were computed through the Discrete Fourier Transform. We then wondered which of the amplitudes would be statistically significant. Significance of either of these amplitudes would mean that there was a significant dependence of speech comprehension on the stimulation phase at the corresponding period. This would therefore show a significant modulation of speech comprehension through the current stimulation.

The significance of the modulation amplitudes was determined in two independent ways. First, we kept the two phase shifts Φ_1 and Φ_2 as well as the constant offset A_0 fixed, and estimated the amplitudes A_1, A_2 and A_3 from multiple linear regression. We then determined the associated p-values and corrected for multiple comparisons through the FDR correction.

Second, we employed a permutation-based method to test the significance of the modulation amplitudes A_1, A_2 and A_3. We therefore computed null models for these amplitudes. The null models were obtained from random permutations of the speech comprehension scores across the six different phases. The permutations were done separately for each subject. For each set of permutations, the parameters in Equation 2.1 were then determined from the Discrete Fourier Transform, as for the actual data. We performed this procedure 10,000 times, resulting in 10,000 null models. We therefrom obtained the null distributions of the modulation amplitudes A_1, A_2 and A_3. We determined the amplitude threshold such that the probability to have a higher amplitude in a null model was 1.7%. This corresponded to a probability of 5% with a Bonferroni correction for the three comparisons. The Bonferroni correction was employed instead of the FDR correction since the latter requires p-values and could not be employed to obtain an amplitude threshold. The null models further allowed us to compute p-values for the amplitudes. The p-value of a particular amplitude followed as the probability of observing a larger value in a null model.

The phase dependence of speech comprehension may differ from subject to subject. We

therefore also analyzed the data when aligning the phase to the '*best phase*' per subject, that is, to the phase that yielded the highest speech comprehension score for that particular subject. We then analyzed the speech comprehension scores $CS(\tilde{\phi})$ at the phases $\tilde{\phi}$ that were aligned to the best phase through the Discrete Fourier Transform. Because the alignment with respect to the best phase left us with only five phases, the Discrete Fourier Transform had only five instead of the previous six parameters:

$$CS(\tilde{\phi}) = A_0 + A_1 cos(\tilde{\phi} - \Phi_1) + A_2 cos(2\tilde{\phi} - \Phi_2) \tag{2.3}$$

In particular, a modulation of speech comprehension could arise through a modulation at either the period of 360° or 180°, with the modulation amplitude of A_1 and A_2, respectively.

We determined the statistical significance of the two amplitudes A_1 and A_2 as for the case of the non-aligned data described above. In particular, we used two independent methods, multiple linear regression and the permutation-based test.

2.3 Results

2.3.1 Relation between time-shifted and phase-shifted waveforms

We first sought to investigate the effect of the phase shifts on the neurostimulation waveforms. Both the neurostimulation signal in the delta frequency band as well as that in the theta frequency band contained a range of frequencies and therefore differed from purely sinusoidal signals (Fig. 2.1D and E). Because the same phase shift was applied to all frequency components, the phase shift did not change the group delay, which follows as the derivative of the phase with respect to frequency. However, the phase delay is defined as the ratio of the phase to the angular frequency, and was therefore altered by the phase shift, in a manner that varied with the frequency. This effect led to a phase-shifted signal that had a different shape from the original one. Moreover, the phase-shifted signal differed from a time-shifted waveform as well.

However, because both the delta-band portion and the theta-band portion of the speech envelope are comparatively narrow-band signals, phase shifts translated approximately to temporal shifts as long as the latter were not too long. To quantify this correspondence, we computed the cross-correlation between the delta-band signal shifted by different phases and temporal delays with the unshifted version, that is, with the signal with neither a time shift nor a phase shift (Fig. 2.2A). We found that for latencies around 0 the maximal correlation values were close to 1. As an example, a maximal correlation value of 0.5 (across phases) was observed for delays between -210 ms and 210 ms. If we consider a correlation value of at least 0.5 to denote a reasonable correspondence between two signals, then this shows that time delays between -210 ms and 210 ms could be approximately represented by phase shifts. We carried out the same analysis for the speech envelope filtered in the theta band (Fig. 2.2B). We obtained maximal correlation (across the different phase shifts) of at least 0.5 for temporal shifts between -150 ms

and 150 ms, evidencing that such temporal delays could partly be captured by phase shifts.

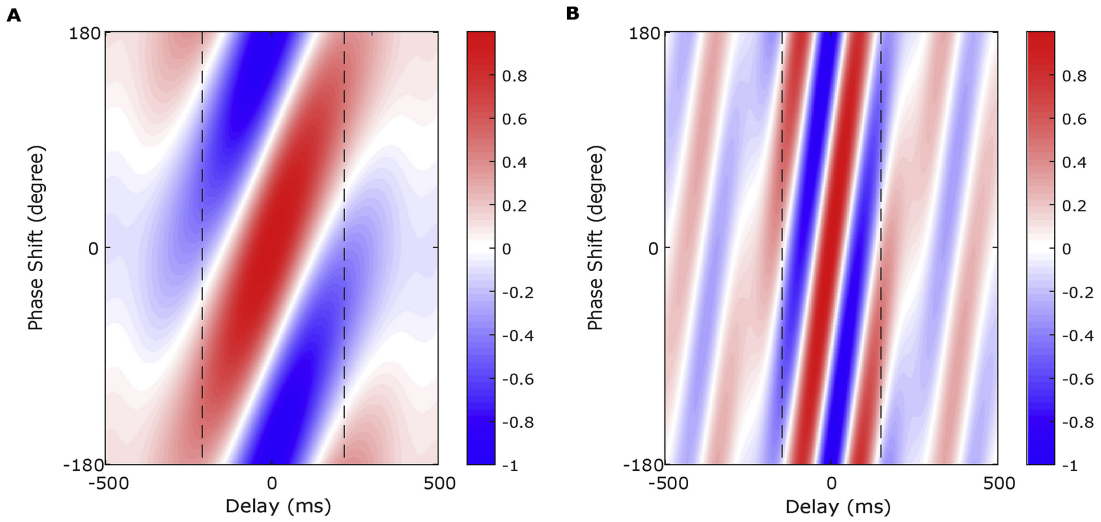

Figure 2.2: The relation between phase and time shifts. (**A**) The correlation of the speech envelope filtered in the delta band, to this signal shifted by different delays and phases. A temporal shift can be compensated by a certain shift in phase. In particular, the maximal correlation (across the different phases) is at least 0.5 between a delay of -210 ms and 210 ms (dashed lines). (**B**) The correlation of the theta-band filtered speech envelope with a version shifted in phase and time. A time shift can be compensated by a certain phase shift: the maximal correlation (across phase) exceeds 0.5 for delays between -150 ms and 150 ms (dashed lines). For larger temporal shifts there is less correspondence between time and phase shifts.

The cross-correlation analysis also verified the cyclical nature of the phase changes. In particular, in the absence of a temporal delay, a signal at a phase change of -180° or of 180° was anti-correlated to the signal without a phase change. The phase change of -180° or of 180° did indeed yield a signal that corresponded to the original one, but with the opposite polarity (Fig. 2.1D and E). Other phase shifts led to a cross-correlation with the unshifted waveform that changed cyclically from -1 (perfect anti-correlation) for a phase shift of -180° to 0 (no correlation) for a phase shift of -90°, to 1 (perfect correlation) for no phase shift (0°), and then back to 0 (no correlation) for a phase shift of 90° and to -1 (perfect anti-correlation) for a phase shift of 180°.

These results confirm that phase shifts and temporal delays are two different ways to manipulate the neurostimulation waveform. Although phase changes relate approximately to temporal delays as long as these are not too long, both manipulations yield in general different results and can therefore have different effects on speech comprehension. In this study we employed phase shifts since this type of manipulation allowed us, due to the cyclical nature of the phase shifts, to use circular statistics for the investigation of the resulting speech comprehension.

2.3.2 Modulation of speech comprehension through theta- but not delta-band neurostimulation

We measured speech comprehension scores while participants experienced transcranial alternating current stimulation with a waveform that was derived from the speech envelope, but band-pass filtered into either the delta or the theta band. To explore the effect of the two types of current stimulation on speech comprehension, we then employed current waveforms that were shifted by six different phases (0°, 60°, 120°, 180°, 240° and 300°). As set out in the Methods section, due to the cyclical nature of the phase, the dependence of the speech comprehension score on the phase of the current stimulation can be written as a linear combination of sinusoidal variations at periods of 360°, 180° and 120° (Equation 2.1). We computed the amplitudes A_1, A_2 and A_3 of these variations through the Discrete Fourier transform. We then assessed the statistical significance of each modulation amplitude through two independent methods, multiple linear regression as well as a permutation-based test.

For the current stimulation with the speech envelope filtered in the delta band, the multiple linear regression showed that none of the amplitudes were statistically significant ($df = 3$; $A_1 = 0.01$, $t = 2.9$, $p = 0.2$; $A_2 = 0.01$, $t = 1.8$, $p = 0.2$; $A_3 = 0.005$, $t = 1.0$, $p = 0.3$; $R^2 = 0.064$; FDR correction for multiple comparisons, Fig. 2.3A). This was confirmed by the permutation-based method (A_1, $p = 0.3$; A_2, $p = 0.1$; A_3, $p = 0.2$; Fig.2.4A–C). There was accordingly no modulation of speech comprehension through the delta-band current stimulation.

For the stimulation in the theta band, however, the multiple linear regression revealed the statistical significance of the modulation amplitude A1, although the others were insignificant ($df = 3$; $A_1 = 0.02$, $t = 2.9$, $p = 0.01$; $A_2 = 0.01$, $t = 1.5$, $p = 0.2$; $A_3 = 0.0002$, $t = 0.03$, $p = 0.97$; $R^2 = 0.097$; FDR correction for multiple comparisons, Fig. 2.3B). The permutation test corroborated this finding (A_1, $p = 0.01$; A_2, $p = 0.3$; A_3, $p = 0.3$; Fig. 2.4D–F). This showed that the theta-band current stimulation had a significant influence on speech comprehension, namely at the longest period of 360°.

2.3.3 Consistent phase dependencies across subjects

The above analysis was performed on the population level, and the phase of the neurostimulation was not adjusted per subject. However, prior studies found that the effect of neurostimulation on speech comprehension may depend on the parameters of the current stimulation, such as phase delay or time shift, in a manner that is not consistent across subjects (Riecke et al. 2018; Wilsch et al. 2018; Zoefel et al. 2018). We therefore investigated whether we had significant subject-to-subject variation in the dependence of the comprehension scores on the stimulation phase. To this end, we determined for every subject, and separately for the delta and for the theta band, the phase that yielded the highest comprehension score. We referred to this phase as the 'best phase' for that subject, and aligned the phase relative to this best phase (Fig. 2.5A and B). We performed the analysis of the dependence of the comprehension scores on the relative phase through the model given by Eq. 2.3. This model described the dependence of the

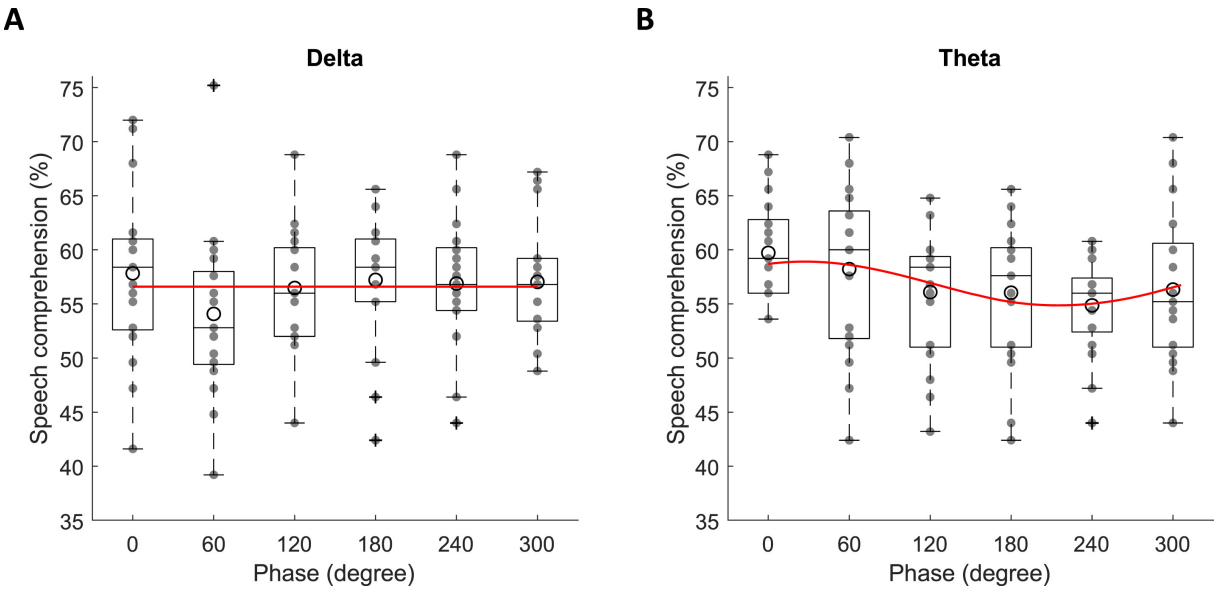

Figure 2.3: Modulation of speech comprehension through theta-band but not delta-band current stimulation. Speech comprehension scores at different phases are shown as box plots, the circles indicate the mean values and crossed denote outliers. Results from individual subjects are indicated as grey disks. The red line denotes the fit obtained from the model given by Eq. 2.1, but using only those terms that are statistically significant. (**A**) Speech comprehension during delta-band stimulation is not influenced by the phase of the stimulation. (**B**) Theta-band stimulation leads to a significant modulation of speech comprehension, at the longest possible period of 360°.

speech comprehension scores on the aligned phases through variations at only two periods, 360° and 180°, with the corresponding amplitudes A_1 and A_2, reflecting that only five phases remain after the alignment to the best phase.

For the stimulation with the delta-band filtered speech envelope, the multiple linear regression revealed no significant modulation of speech comprehension ($df = 2$; $A_1 = 0.009$, $t = 1.0$, $p = 0.3$; $A_2 = 0.005$, $t = 1.2$, $p = 0.3$; $R^2 = 0.03$; FDR correction for multiple comparisons, Fig. 2.5A), which was confirmed by the permutation-based test (A_1, $p = 0.3$; A_2, $p = 0.2$). Likewise, the multiple linear regression showed no significant impact of the theta band stimulation either ($df = 2$; $A_1 = 0.008$, $t = 0.35$, $p = 0.7$; $A_2 = 0.007$, $t = 0.8$, $p = 0.7$; $R^2 = 0.01$; FDR correction for multiple comparisons, Fig. 2.5B). This was corroborated by the permutation test (A_1, $p = 0.6$; A_2, $p = 0.2$). The alignment with respect to the best phase per subject accordingly rendered the previously-obtained modulation with speech comprehension through the theta-band current insignificant.

To investigate the potential inter-subject variability of the phase dependence further, we computed the distribution of the subjects' best phases (Fig. 2.5C and D). We found that, for neurostimulation in the delta band, the distribution was not significantly different from a uniform one ($p = 0.4$, Rayleigh test). This accorded with our finding that delta-band stimulation did not have a significant influence on speech comprehension, since the best phase is then dis-

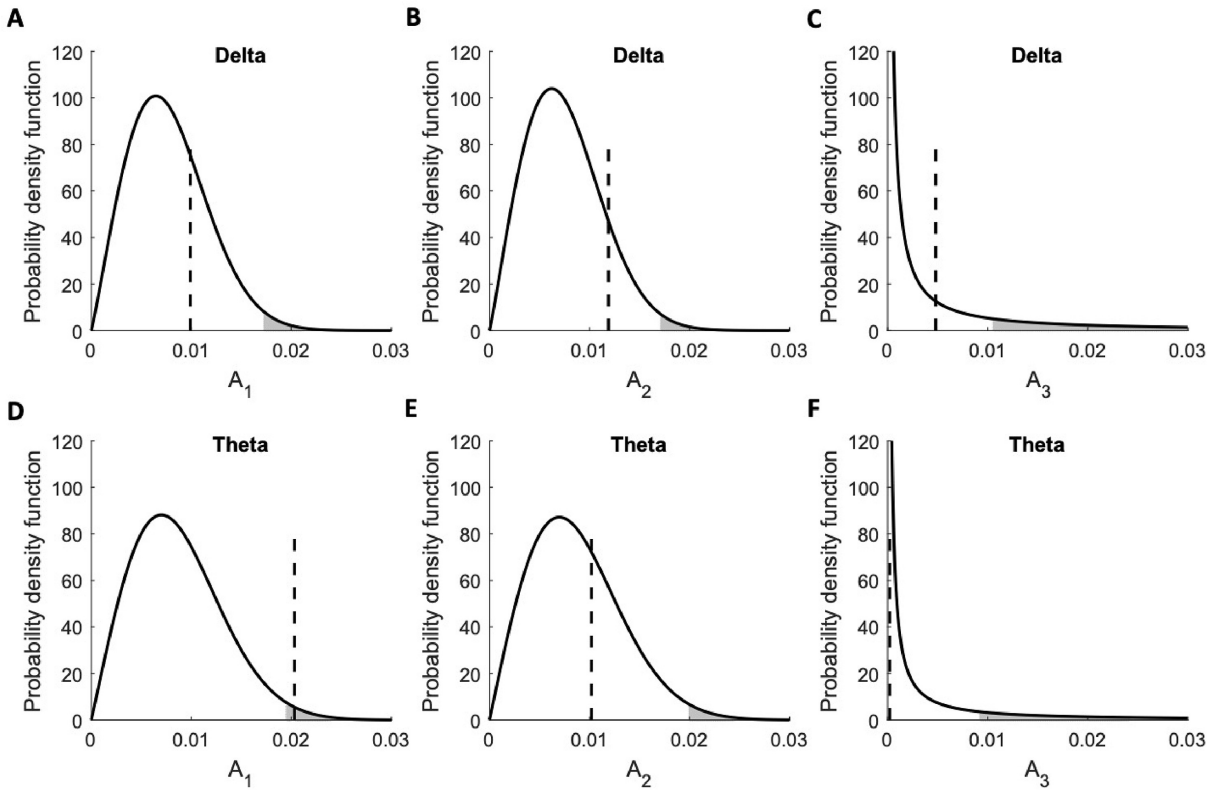

Figure 2.4: Significant dependence of speech comprehension on the stimulation phase for theta-band but not delta-band current stimulation. We used permutations of the speech-comprehension scores to compute null models of the modulation amplitudes, and therefrom their probability distributions (black lines). The grey areas show the largest amplitudes that were observed in the null models with a probability of less than 1.7%, which corresponded to a probability of 5% adjusted for the three comparisons with the Bonferroni correction. The modulation amplitudes computed from the actual data are shown as dashed lines (**A**-**C**) The dependence of speech comprehension on the stimulation phase for delta-band stimulation is insignificant at all three periods. (**D**-**F**) The dependence of speech comprehension on the stimulation phase for the theta-band stimulation is significant for the longest period (A_1 is significant) but not at the two others (A_2 and A_3 are insignificant).

tributed randomly. The current stimulation in the theta band, however, showed a distribution of the best phases that differed significantly from uniformity ($p = 0.02$, Rayleigh test). The mean phase was $36° \pm 30°$. This provided additional evidence that the best phase for the theta-band stimulation was consistent across subjects.

2.3.4 Enhancement of speech comprehension through theta-band neurostimulation

Furthermore, we wondered whether current stimulation could not only modulate but actually enhance the comprehension of speech in noise. We therefore also measured the comprehension scores when subjects experienced a sham stimulus. As an additional control, we stimulated volunteers with a current that followed the envelope of an unrelated sentence, filtered either in the delta or in the theta frequency band. These currents obtained from unrelated sentences

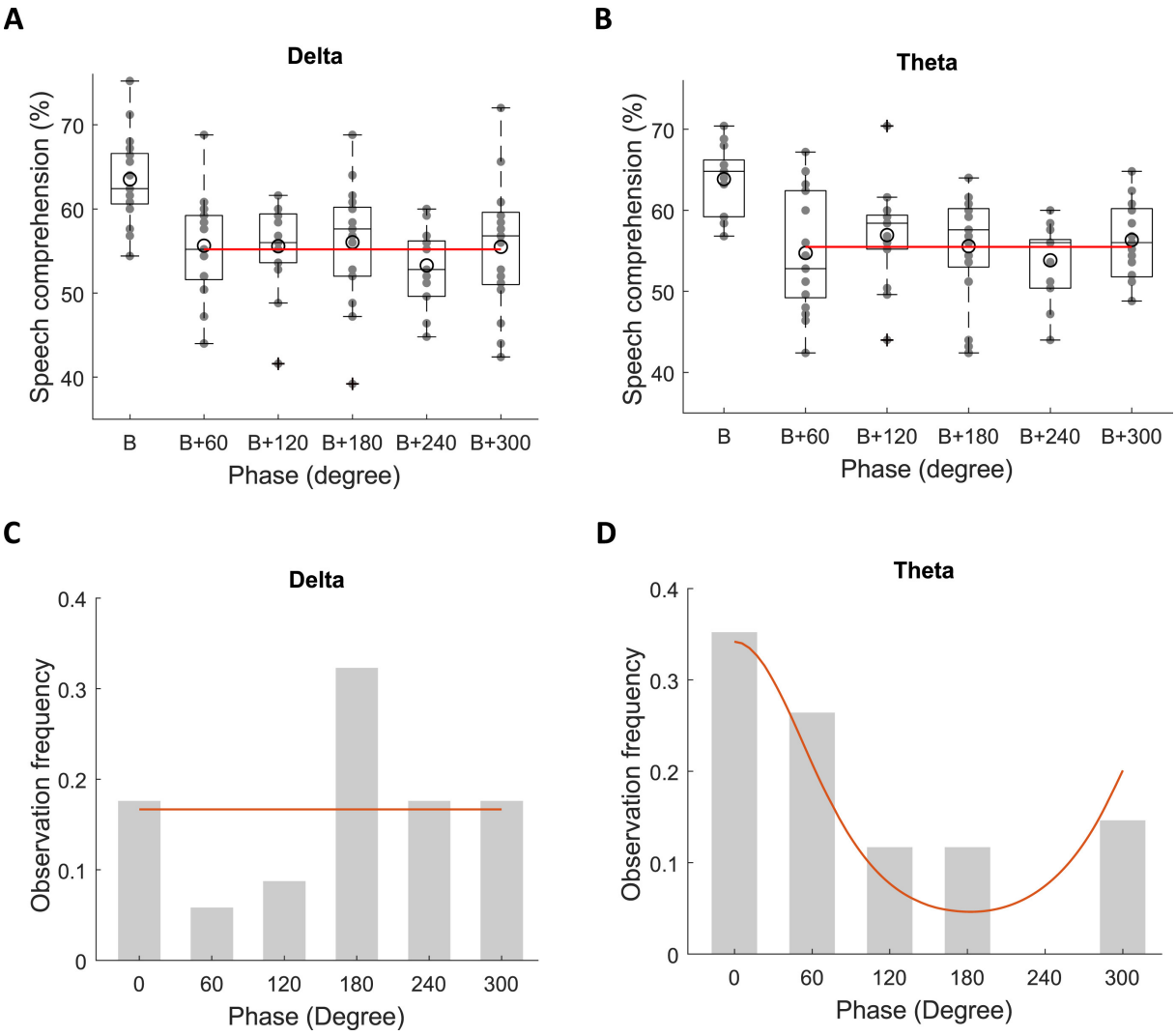

Figure 2.5: Consistent phase dependency across subjects. (**A**, **B**) Results on the population level are shown through box plots, with crossed denoting outliers. Open circles denote the mean values, and grey disks indicate the results from individual subjects. When adjusting the phase relative to the best phase ('*B*') per volunteer, neither the delta-band stimulation (**A**) nor the theta-band entrainment (**B**) lead to significant effect of phase on speech comprehension. The red line denotes the fit obtained from the significant parts of the model given in Eq. 2.1. (**C**) For delta-band stimulation the distribution of the best phases (grey bars) is not significantly different from a uniform distribution (red line). (**D**) Theta-band stimulation leads to a distribution (grey bars) that differs significantly from uniformity. The best phases occur around the mean phase of 36°±30°. The distribution can be approximated well by a von Mises distribution (red line).

should not facilitate speech comprehension, but, if anything, hinder it.

We compared the comprehension scores that we obtained for the delta- and theta-band stimulation at the phase that yielded the highest comprehension across subjects — the phase of 0° in either case — to the different control conditions (Fig. 2.6). We found that there was statistically significant variation between the different comprehension scores (One-way ANOVA, $df = 4$, $F = 3.1$, $p = 0.02$, $\eta^2 = 0.1$). Post-hoc tests showed that the only two types of neurostimulation that yielded significantly different speech comprehension were the theta-band stimulation and

the sham stimulation ($p = 0.03$), Tukey-Kramer method (Driscoll 1996; Tukey 1949). In particular, transcranial alternating current stimulation with the theta-band filtered speech envelope, and without a phase shift, yielded speech comprehension that was significantly better than the one obtained under sham stimulation. Speech comprehension improved by 6%, which is comparable to the efficacy of some noise-reduction algorithms for hearing aids and suggests that this type of neurostimulation may have practical applications in auditory prosthetics (Chung 2004; Healy et al. 2019).

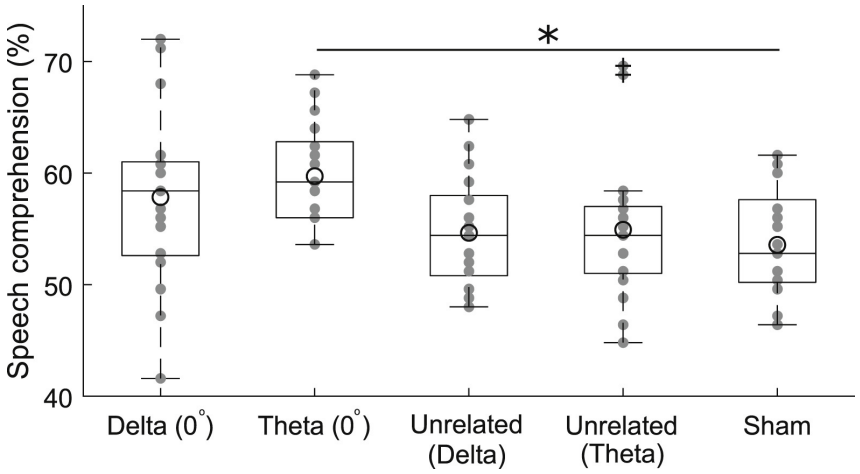

Figure 2.6: Enhancement of speech comprehension through current stimulation. We compared current stimulation with the best phase of the delta- and theta-band waveforms (0° for both), stimulation with the envelope of an unrelated sentence filtered either in the delta or in the theta frequency band, as well as sham stimulation. Theta stimulation without phase shift leads to significantly better comprehension scores than sham stimulation. Box plots denote results on the population level, with open circles showing the population mean and crosses indicating outliers. Grey disks show the results from individual subjects. The asterisk indicates a statistically-significant difference.

We also wondered if the variances of the speech comprehension scores differed between the various conditions. Although the variance was largest for the delta-band stimulation, we did not find a statistically-significant difference between the five conditions (Bartlett's test, $k = 5$, $\xi^2 = 7.1$, $p = 0.13$).

2.4 Discussion

We showed that neurostimulation with the theta-band but not the delta-band portion of the speech envelope impacts comprehension. This finding ties in with previous studies that have identified different roles of these two frequency bands for speech processing. In particular, entrainment in the theta band has been shown to relate to acoustic properties of speech, including the clarity of a speech signal in background noise, whereas the delta-band entrainment can inform on higher-level linguistic aspects of speech such as syntactic features, semantics, and thereby comprehension (Broderick et al. 2018; Di Liberto et al. 2015; Ding et al. 2014; Hyafil

et al. 2015; Weissbart et al. 2020). Our study suggests that the theta-band entrainment plays a functional role, perhaps through aiding the acoustic parsing of speech. Our observed lack of modulation of speech comprehension through delta-band stimulation may reflect that, although the neural speech tracking in the delta band relates to higher-level linguistic information in speech and to speech comprehension, this relationship originates in only a small portion of the delta-band entrainment (Broderick et al. 2018; Ding et al. 2016; Etard et al. 2019b). Transcranial alternating current stimulation with the delta-band portion of the speech envelope may not be efficient in modulating this small neural response. Alternatively, the effect may have been too small to observe in the comparatively small number of 17 subjects that we assessed here, or the delta-band speech entrainment may be an epiphenomenon of other neural processes.

Cortical activity entrains to speech rhythms at different temporal lags, in particular at an early latency of 150 ms and a longer latency of 250 ms, suggesting that the timing of the neurostimulation signal with respect to the sound may affect how comprehension is modulated (Ding et al. 2014; Horton et al. 2013). Previous studies on the effects of neurostimulation on speech processing have partly investigated different temporal lags between the speech signal and the transcranial alternating current, and found best lags that were distributed broadly among participants between -400 and 400 ms (Riecke et al. 2018; Wilsch et al. 2018). While our approach employed no temporal delay between the envelope-based current and the speech, our analysis showed that the phase shifts that we used partly corresponded to time lags of about 200 ms in magnitude, such that our approach effectively captures a significant range of temporal delays.

We found evidence of a consistent phase, across volunteers, at which the theta-band current stimulation modulated speech comprehension. Moreover, when considering a subject-specific phase alignment, we no longer obtained a significant effect of phase on speech comprehension. This may indicate that the alignment of the phase according to the best phase per subject increased the noise in the data, which may in turn follow from uncertainty in determining the best phase for each individual. However, our finding of a consistent influence of phase on speech comprehension across the subjects differed from previous studies that found broad variability in how certain temporal lags or phase shifts modulated speech comprehension (Riecke et al. 2018; Wilsch et al. 2018; Zoefel et al. 2018). These studies employed either the broad-band speech envelope, mostly between 1 and 15 Hz, or speech that was artificially altered to follow a single rhythm, which may have increased the variability across participants.

Because the theta-band entrainment plays a functional role in speech comprehension, we expected that current stimulation with an unrelated envelope would worsen speech comprehension compared to a sham stimulus. However, we found that neither stimulation with an unrelated delta band envelope nor with an unrelated theta-band envelope rendered significantly lower comprehension scores. This may indicate that, perhaps due to the relatively high background noise, the theta-band entrainment in the absence of current stimulation was already rather low and did not decrease significantly further upon stimulation with an unrelated envelope.

In summary, our results show that the modulation of speech comprehension through transcranial alternating current stimulation stems from the theta but not from the delta band. We have further demonstrated that the theta-band stimulation modulates speech comprehension in a manner that is consistent across subjects. In particular, there exists an optimal phase shift across subjects at which speech comprehension is aided. Importantly for potential practical applications, our results evidence that current stimulation within the theta frequency band can enhance speech comprehension with respect to sham stimulation, a result that had not been possible with the use of broad-band current stimulation (Riecke et al. 2018; Wilsch et al. 2018).

Chapter 3

Modelling the effects of transcranial alternating current stimulation on the neural encoding of speech in noise

The work presented in this chapter has been previously published as Kegler et al. (2021). Implementation of the model introduced in this chapter, as well as all the associated analysis tools, are openly available at `https://github.com/MKegler/SpeechTACSmodel`. The development of the model was tightly coupled with the tACS experiment presented in Chapter 2. The model was employed to generate hypotheses for the experimental study and optimize the stimulation protocol, while the data collected during the experiment were used to validate the model.

3.1 Introduction

Naturalistic listening environments are often noisy. Talking to a friend in a busy pub or restaurant, for instance, means that we need to ignore other distracting sounds around us. However, humans excel at this challenging task: we can still understand speech even when the background noise becomes louder than the target signal itself (Anderson et al. 2010; Drullman 1995; Hutcherson et al. 1979; Soli et al. 2008)

This remarkable performance partly involves the tracking of amplitude fluctuations in speech by cortical activity (Han et al. 2019; Hickok et al. 2007; Mesgarani et al. 2014; Morillon et al. 2012). In particular, the neural oscillations in the delta (1 - 4 Hz) and theta (4 - 8 Hz) frequency ranges become correlated with the acoustic envelope of a speech stimulus (Brodbeck et al. 2020b; Kubanek et al. 2013; Lalor et al. 2010; Molinaro et al. 2018). They can thereby track the rhythm set by words (in the delta range) and by syllables (in the theta range). When a speech stimulus is obscured by background noise, such as a competing speaker, this low-frequency cortical tracking can predict speech discrimination performance (Luo et al. 2007), selective attention (Golumbic et al. 2013; O'Sullivan et al. 2015; O'Sullivan et al. 2017), speech intelligibility (Lesenfants et al. 2019a; Vanthornhout et al. 2018) and comprehension (Etard et al. 2019b; Iotzov et al. 2019). The delta and theta frequency band thereby play different roles: cortical tracking in the theta band is linked to lower-level acoustic processing of the speech stimulus, while delta-band tracking can inform on higher-level aspects such as the processing of semantic and syntactic information (Broderick et al. 2018; Ding et al. 2016; Etard et al. 2019b).

Neural tracking of speech features has also been demonstrated in a higher frequency band, the gamma band. It contains activity above 25 Hz and can encode phonemes, the basic units of speech (Gross et al. 2013; Shamir et al. 2009). A recent hypothesis postulates that speech processing occurs through a cross-frequency coupling of cortical oscillations (Giraud et al. 2012; Gross et al. 2013). According to this hypothesis, the cortical activity in the theta band parses speech into smaller units, presumably syllables (Ghitza 2011; Giraud et al. 2007). The theta activity then modulates the cortical responses in the gamma range, thus providing temporal frames for the phonemic encoding.

Transcranial alternating current stimulation (tACS) provides a non-invasive means to influence cortical activity in humans, in particular at the frequency of the stimulation (Helfrich et al. 2014; Krause et al. 2019; Reato et al. 2013; Ruhnau et al. 2016; Zaehle et al. 2010). Sinewave tACS combined with the rhythmic presentation of a speech stimulus has indeed been shown to affect the cortical responses to speech (Zoefel et al. 2018). Moreover, tACS with the speech envelope impacts behaviour as well: the comprehension of speech in noise can be modulated through concurrent neurostimulation (Kadir et al. 2019; Keshavarzi et al. 2020a, 2020b; Wilsch et al. 2018). The modulation is modest, up to a few percent in the comprehension scores. It results from the theta but not the delta portion of the speech envelope, indicating that the stimulation may act on the syllable parsing (Keshavarzi et al. 2020a). Moreover, the current stimulation in the theta band can boost the comprehension of speech in background noise beyond that observed during sham stimulation (Keshavarzi et al. 2020a, 2020b).

The experimental data regarding the effect of tACS with the speech envelope on speech comprehension show, however, certain inconsistencies. Two key variables that have been explored when applying tACS simultaneous to speech in noise are the delay between the current waveform and the speech envelope, as well as a potential phase shift between these two signals. Some studies found that the value of the stimulation parameter, either of the delay or of the phase shift, that yielded the highest speech comprehension varied considerably between subjects (Riecke et al. 2018; Wilsch et al. 2018). These results suggest that the current stimulation acts on a cortical source that is highly variable from subject to subject. In contrast, other studies found that the optimal delay and phase shift of the current waveform with respect to the speech signal were similar across different study participants (Kadir et al. 2019; Keshavarzi et al. 2020a, 2020b). The inconsistencies between these different investigations provide additional motivation for better understanding the functional mechanisms by which tACS influences speech comprehension.

Computational modelling offers a promising route to investigate the effects of non-invasive brain stimulation (Bestmann et al. 2015; Bonaiuto et al. 2015; Fröhlich 2015; Frohlich et al. 2013; Fröhlich et al. 2015). Well-established finite-element models that are based on structural imaging data are, for instance, used to estimate the distribution of electrical current in the brain (Datta et al. 2009; Huang et al. 2019a). They allow to optimize the placement of elec-

trodes on the scalp and can explain some inter-subject variability (Huang et al. 2019b; Kasten et al. 2019). They do, however, not provide information on the functional mechanisms by which the current stimulation influences the neural network activity underlying the behavioural effects.

The functional influence of current stimulation can be addressed through biophysically-plausible spiking neural network models combined with a model of how each neuron's activity is affected by a weak current (Ali et al. 2013; Cakan et al. 2020; Herrmann et al. 2016; Reato et al. 2010). Recent effort in this direction has, for instance, uncovered that tACS can act on cortical oscillations through periodic forcing (Cakan et al. 2020; Fröhlich et al. 2010; Herrmann et al. 2016; Reato et al. 2010) as known from other nonlinear dynamical systems (Pikovsky et al. 2001). However, the functional mechanisms of current stimulation in relation to sensory processing have not yet been investigated computationally.

Here, we introduce a framework for modelling the effects of external electrical stimulation, similar to tACS, on the neural encoding of speech in background noise. Our computational work is based on a recently introduced model of speech encoding through coupled cortical oscillations in the theta and in the gamma frequency ranges (Hyafil et al. 2015). We show that the model can be used to describe the encoding of speech in background noise. We then extend it to include the effects of alternating current stimulation and employ it to investigate the mechanism by which current stimulation affects the speech encoding.

3.2 Methods

3.2.1 Computational model of speech encoding

We employed a computational model for speech encoding in a spiking neural network (Hyafil et al. 2015). The model consisted of two modules of spiking neurons that generated endogenous oscillations in the theta (4 - 8 Hz) and in the gamma (25 - 40 Hz) frequency ranges (Fig. 3.1). The gamma oscillations resulted from a Pyramidal Interneuron Gamma (PIN-G) module (Jadi et al. 2014). In this well established and experimentally validated model, a group of excitatory neurons and another group of inhibitory neurons are reciprocally connected to each other to generate oscillations (Brosch et al. 2002; Cardin et al. 2009; Ray et al. 2011; Sohal et al. 2009). Since the mechanisms of the neural activity in the theta frequency range remain unknown (Ainsworth et al. 2011), the theta-generating module was designed analogously to the gamma module, but with adjusted parameters such as slower time scales, and was referred to as PIN-TH model.

The spiking neural network model contained 84 leaky integrate-and-fire neurons of four distinct types: gamma excitatory neurons (Ge, $N_{Ge} = 32$ cells), gamma inhibitory neurons (Gi, $N_{Gi} = 32$ cells), theta excitatory neurons (Te, $N_{Te} = 10$ cells) and theta inhibitory neurons (Ti, $N_{Ti} = 10$ cells). The first two types of neurons formed the PIN-G module, and the second two types belonged to the PIN-TH module.

The temporal evolution of the membrane potential V_i of neuron i is described by the following

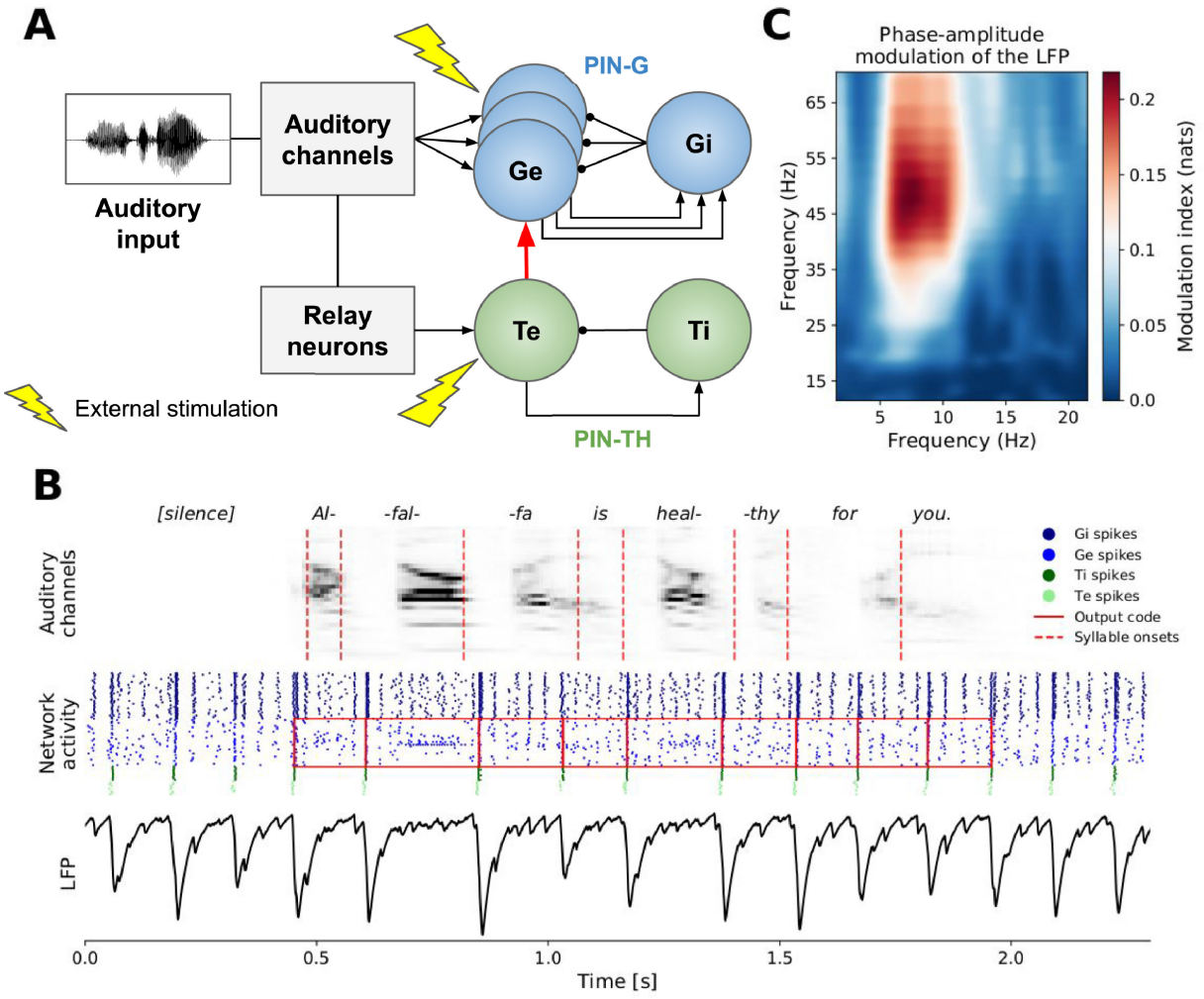

Figure 3.1: Architecture of the spiking neural network and its dynamics. (**A**) Network architecture. A PIN-TH module (green) consisted of 10 excitatory neurons (Te) and 10 inhibitory neurons (Ti) to generate self-sustained oscillations in the theta frequency band. Analogously, a PIN-G module (blue) with 32 excitatory cells (Ge) and 32 inhibitory cells (Gi) produced faster gamma-range activity. Both modules were coupled unidirectionally through all-to-all connections from the Te to the Ge cells. The auditory input to the model was firstly decomposed into 32 frequency-specific auditory channels, using a model of the auditory periphery. The resulting signals were projected to Ge neurons. They were also convolved with a spectrotemporal filter that mimicked the action of relay neurons and then fed into the Te neurons. The application of transcranial current stimulation (yellow) was simulated as a current injection to all excitatory cells in the model. (**B**) The network's response to the example sentence *'Alfalfa is healthy for you'*, preceded by silence. The model of the auditory periphery decomposed the sound into 32 auditory channels (top). The resulting neural spikes from the theta module (middle, green) allowed to infer syllable boundaries, and to group the neural output of the gamma module (middle, red boxes) according to the individual syllables, enabling the decoding of the syllable identity. The local field potential (LFP, bottom) followed as the sum of the synaptic currents delivered to the excitatory neurons. (**C**) The coupling from the theta module to the gamma module resulted in phase-amplitude modulation. In particular, the phase-amplitude modulation index was high for phases in the theta range, around 5 – 12 Hz, and for amplitudes between 35 – 70 Hz, in the gamma range.

equation:

$$C\frac{dV_i}{dt} = g_L(V_L - V_i) + I_i^{SYN} + I_i^{INP} + I_i^{EXT} + I_i^{DC} + \eta, \tag{3.1}$$

in which C is the capacitance of the cellular membrane, g_L and V_L are the conductance and the reversal potential of the leak current; I_i^{SYN}, I_i^{INP}, I_i^{EXT} and I_i^{DC} are the synaptic, stimulus-induced, exogenous and constant currents delivered to the cell, and η is a Gaussian noise with variance σ_i. When the membrane potential of the ith neuron reached the threshold V_{THR} a spike was generated and V_i returned to the reset potential V_{RESET}.

The dynamics of synaptic currents between neurons were modelled as follows:

$$\frac{dx_{ij}^R}{dt} = -\frac{x_{ij}^R}{\tau_j^R} + \delta(t - t_j^{SPK}), \tag{3.2}$$

$$\frac{ds_{ij}}{dt} = \frac{x_{ij}^R - s_{ij}}{\tau_j^D}, \tag{3.3}$$

where s_{ij}, x_{ij}^R are activation variables of the synapse at neuron i for a connection coming from neuron j, $\delta(t - t_j^{SPK})$ indicates a spike generation in the presynaptic neuron at the time t_j^{SPK}, and τ_j^R, τ_j^D are time constants that describe the rise and decay of the activation from neuron j, respectively. The synaptic current I_i^{SYN} is then the sum of all synaptic inputs to neuron i from the remaining cells:

$$I_i^{SYN}(t) = \sum_j g_{ij} s_{ij}(t)(V_j^{SYN} - V_i(t)), \tag{3.4}$$

where g_{ij} is the synaptic conductance of the synapse from neuron j to i, and V_j^{SYN} is the equilibrium potential of the presynaptic neuron j.

Because we only modelled a small and local neural network, we employed all-to-all connections between the different neurons of each subtype. The PIN-G and PIN-TH modules were then created by reciprocally coupling the corresponding excitatory and inhibitory neurons, that is, those of type *Ge* and *Gi* respectively those of type *Te* and *Ti*. In addition, the *Ti* neurons were all-to-all connected to facilitate sparse synchronous spiking within this population. The cross-frequency coupling in the model was implemented by connecting the PIN-G module to the PIN-TH module through unidirectional all-to-all connections from the *Te* to the *Ge* neurons.

The values of the model parameters were obtained from the study that introduced the model (Hyafil et al. 2015), and are listed in Table 3.1. Equations 3.1, 3.2, 3.3, 3.4 were solved numerically using the Euler method with a time step of 10 μs. The local field potential (LFP) was obtained by summing the absolute values of all synaptic currents delivered to the excitatory cells *Ge* and *Te* in the network (Mazzoni et al. 2008).

Table 3.1: Model parameters.

Parameter	Description	Value
	Neuron model	
C	Cell membrane capacitance	1 pF
V_{THR}	Spiking threshold	-40 mV
V_{RESET}	Resting potential	-87 mV
V_L	Equilibrium potential of leak	-67 mV
V_E^{SYN}	Equilibrium potential of excitatory neurons	0 mV
V_I^{SYN}	Equilibrium potential of inhibitory neurons	-80 mV
	PIN-G network	
g_{LE}, g_{LI}	Leak conductance in Ge, Gi neurons	0.1 nS
τ_{Ge}^R	Synaptic rise constant of Ge neurons	0.2 ms
τ_{Gi}^R	Synaptic rise constant of Gi neurons	0.5 ms
τ_{Ge}^D	Synaptic decay constant of Ge neurons	2 ms
τ_{Gi}^D	Synaptic decay constant of Gi neurons	20 ms
I_{Ge}^{DC}	Constant current delivered to Ge neurons	3 pA
I_{Gi}^{DC}	Constant current delivered to Gi neurons	1 pA
σ_{Ge}, σ_{Gi}	Variance of the noise term in Ge, Gi neurons	2.028 pA$\cdot\sqrt{ms}$
	PIN-TH network	
g_{LE}	Leak conductance in Te neurons	0.0264 nS
g_{LI}	Leak conductance in Ti neurons	0.1 nS
τ_{Te}^R	Synaptic rise constant of Te neurons	4 ms
τ_{Ti}^R	Synaptic rise constant of Ti neurons	5 ms
τ_{Te}^D	Synaptic decay constant of Te neurons	24.3150 ms
τ_{Ti}^D	Synaptic decay constant of Ti neurons	30.3575 ms
I_{Te}^{DC}	Constant current delivered to Te neurons	1.25 pA
I_{Ti}^{DC}	Constant current delivered to Ti neurons	0.0851 pA
σ_{Te}	Variance of the noise term in Te neurons	0.282 pA$\cdot\sqrt{ms}$
σ_{Ti}	Variance of the noise term in Ti neurons	2.028 pA$\cdot\sqrt{ms}$
	Connectivity	
$g_{Ge,Gi}$	$Gi \rightarrow Ge$ synaptic conductance strength	$5/N_{Gi}$ nS
$g_{Gi,Ge}$	$Ge \rightarrow Gi$ synaptic conductance strength	$10/N_{Ge}$ nS
$g_{Ge,Te}$	$Te \rightarrow Ge$ synaptic conductance strength	$1/N_{Te}$ nS
$g_{Te,Ti}$	$Ti \rightarrow Te$ synaptic conductance strength	$2.07/N_{Ti}$ nS
$g_{Ti,Te}$	$Te \rightarrow Ti$ synaptic conductance strength	$6.66/N_{Te}$ nS
$g_{Ti,Ti}$	$Ti \rightarrow Ti$ synaptic conductance strength	$4.32/N_{Ti}$ nS

3.2.2 Simulation of alternating current stimulation in the model

Following recent computational models for the effects of tACS on neural oscillations, we simulated the neurostimulation as a current injected to all excitatory neurons in the network (Ali et al. 2013; Herrmann et al. 2016; Negahbani et al. 2018) (Fig. 3.1, yellow). Experimental evidence suggests indeed that pyramidal neurons, the excitatory ones, are significantly more susceptible to external electric fields than the inhibitory interneurons (Radman et al. 2009).

To calibrate the intensity of the exogenous stimulation, a constant stimulation current I^{EXT} was applied to an isolated Ge pyramidal neuron. Specifically, the synaptic current I^{SYN}, the stimulus input current I^{INP}, as well as the constant current I^{DC} were all set to 0 with the

remaining parameter values unchanged. The external current was applied 10 s after the start of a simulation, for a duration of 10 s. Its intensity was varied from 0.01 pA to 1 pA in steps of 0.01 pA. For each intensity of the external current, we ran 100 simulations. We thereby identified the spiking threshold of an isolated *Ge* neuron as 0.71 pA. This intensity of stimulation led, just below the spiking threshold, to an average membrane depolarization over 7 mV, comparable to the levels observed in previous computational models for the effects of tACS (Negahbani et al. 2018). Since non-invasive transcranial electrical stimulation in humans is not powerful enough to directly cause spiking in cortical neurons, in the following simulations we considered sub-threshold stimulation at three intensities: 0.1 pA, 0.2 pA and 0.5 pA. These led to an average membrane depolarization of 1 mV, 2 mV and 5 mV, respectively.

3.2.3 Auditory stimuli and network simulations

Spoken English sentences from the TIMIT dataset (Garofolo et al. 1993) at a sound-pressure level of 76 dB SPL were used as input to the neural network model. To investigate speech-in-noise encoding in the model, we chose a random subset of 100 sentences. We added four-talker babble noise to each sentence at signal-to-noise ratios (SNRs) that ranged from -25 to 25 dB, in steps of 5 dB. The SNR was thereby determined from the ratio of the root-mean-square amplitudes of the signal and of the background noise.

For each SNR and each sentence, we simulated the neural network response 100 times. Because the theta module generated intrinsic oscillatory activity, we wanted to prevent an accidental alignment between this theta activity and the onset of the speech. Each sentence was therefore preceded by a silent period whose duration varied randomly between 380 ms and 550 ms. Each simulation was terminated 100 ms after the end of the presented sentence.

To investigate the effect of the neural coupling between the PIN-TH module and the PIN-G module, we employed a simpler simulation setup: we computed the LFP in response to the exemplary sentence '*Alfalfa is healthy for you.*'. The model responses were simulated 30 times, and in each simulation the sentence was preceded by a random period of silence that ranged from 500 ms to 1,000 ms.

3.2.4 Input of the acoustic signal to the neural network

Following the previously introduced model for speech processing through a coupled PIN-TH and PIN-G modules (Hyafil et al. 2015), the auditory input was processed through a model of the auditory periphery (Chi et al. 2005). This model firstly decomposed the auditory stimulus through a cochlear filter bank into 128 channels. The signals in the different channels were then subjected to nonlinear transformations that reflected neural processing in the auditory nerve and the subcortical nuclei. First, mimicking the action of hair cells, the filtered signals were high-pass filtered, nonlinearly compressed and then low-pass filtered (Yang et al. 1992). Second, a first order derivative across frequency channels was taken, followed by a half-way rectification,

which reflected the lateral inhibition in the cochlear nucleus (Shamma 1989). Third, the signal in each channel was integrated over a short temporal duration of 8 ms, reflecting the decay of temporal precision in the midbrain. The obtained signals were interpreted as currents measured in pA, and approximated the tonotopically organized input to the primary auditory cortex.

The auditory stimuli processed through the model of the auditory periphery were projected to both the PIN-G and the PIN-TH module. First, regarding the PIN-G module, each of the 32 Ge neurons received input from one auditory channel, in a tonotopic fashion. To this end, the number of auditory channels was reduced to 32 by selecting every fourth auditory channel from all 128 available.

Second, the sound stimuli were used as input to the slower PIN-TH module as well. In particular, the Te neurons were stimulated in a way that tracked syllable onsets as faithfully as possible. To this end, the Te neurons received an input current $Y(t)$ that was the convolution of the 32 auditory channels described above with a spectrotemporal filter at auditory channel c and delay τ:

$$Y(t) = \sum_{c=1}^{32} \sum_{i=1}^{6} B(c, \tau) X(c, t - \tau_i), \tag{3.5}$$

in which X is the signal in the auditory channel c at time $t - \tau_i$. The 6 temporal delays τ_i were uniformly distributed between -50 ms and 0 ms. The convolution of the auditory input with the filter B modelled the effect of a population of relay neurons with delays of up to 50 ms, and with weights that represent the strength of synaptic connections (Pillow et al. 2008). Unlike the tonotopically organized Ge neurons, all Te cells received the same current input $Y(t)$.

The spectrotemporal filter B was computed from 1,000 randomly chosen sentences from the TIMIT corpus to optimize the predictions of the syllable onsets (Hyafil et al. 2015). These sentences differed from the ones that were used for subsequent investigations of speech coding in the neural network. The audio signals were preceded by a silent part whose duration varied randomly between 500 ms and 1,000 ms. The signals were processed by the model of the auditory periphery, downsampled to 100 Hz and concatenated to obtain the signals X.

The onsets of syllables were obtained from the TIMIT transcription, and were used to compute a binary vector. The syllable onsets in this vector were shifted forward by 20 ms such that they occurred after the actual onsets. The filter coefficients B were then computed through sparse bilinear logistic regression to predict this syllable vector, with the syllable onset vector replacing the current input $Y(t)$ in Eq. 3.5 (Adam et al. 2020; Shi et al. 2014).

3.2.5 Stimulation waveform design

We explored stimulation waveforms that were based on the envelope of the speech stimuli (Fig. 3.2). The envelope of a sentence was computed by determining the analytic representation of the speech signal using the Hilbert transform, and by calculating its absolute value.

The obtained signal was further band-pass filtered between 1 - 4 Hz, between 4 - 8 Hz, or between 1 - 20 Hz, yielding the delta portion of the speech envelope, the theta portion of the envelope, or the broadband envelope, respectively (2nd order, zero-phase Butterworth bandpass filter).

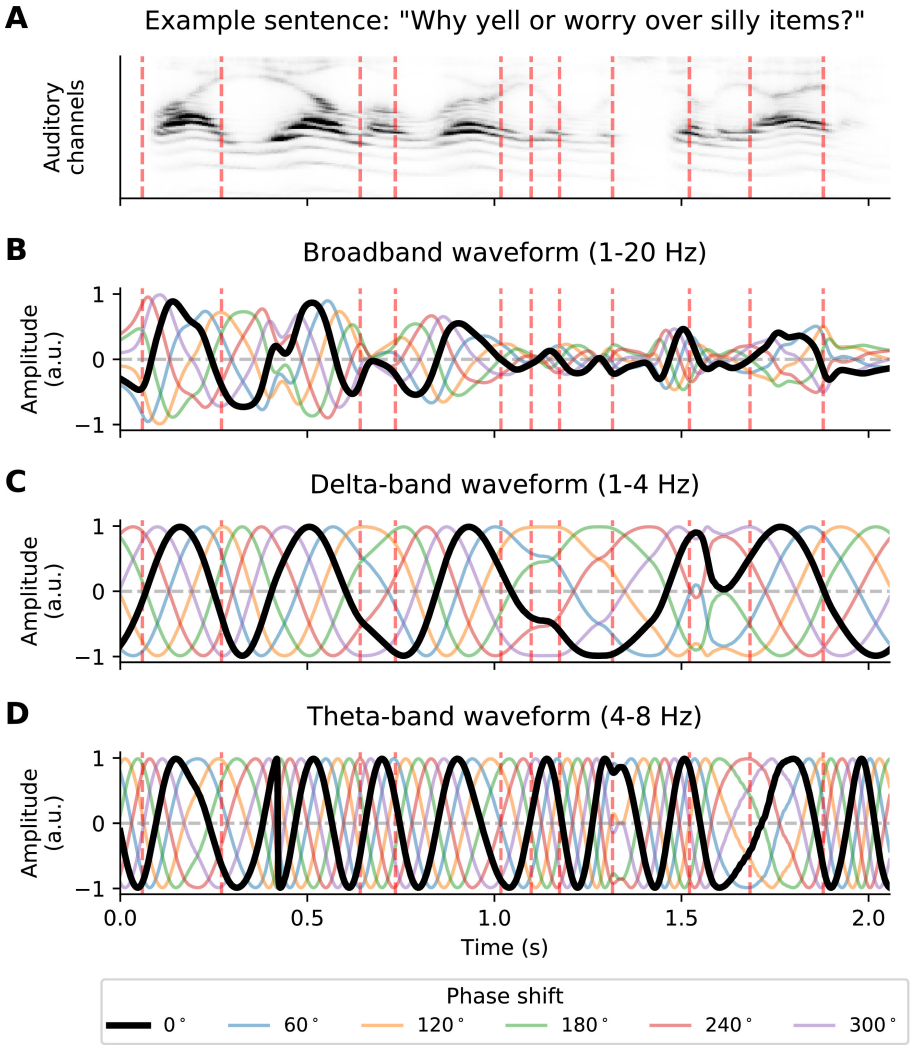

Figure 3.2: Envelope-shaped stimulation waveforms. Stimulation waveforms derived from the exemplary TIMIT sentence '*Why yell or worry over silly items?*'. (**A**) Exemplary sentence decomposed into 128 auditory channels and its syllable boundaries obtained from the TIMIT's phonetic transcription (dashed red lines). (**B**) - (**D**) Waveforms for the neurostimulation were derived from the speech envelope, filtered into a broadband frequency range (**B**), into the delta range (**C**), or into the theta range (**D**). The waveforms in the delta and in the theta band were altered so that the maxima and minima occurred at the values of 1 and -1, respectively. All waveforms were then shifted by six different phases (coloured).

We then shifted the obtained envelopes by six different phases, ranging from 0° to 300°, in steps of 60°. In particular, the shift of an envelope $e(t)$ by a phase ϕ was implemented through the Hilbert transform $H[e(t)]$, yielding the analytical representation $E(t)$ of the envelope:

$$E(t) = e(t) + i \cdot H[e(t)], \tag{3.6}$$

in which i denotes the imaginary unit. The phase-shifted envelope $e_\phi(t)$ then followed as

$$e_\phi(t) = |E(t)|Re(e^{i\{arg[E(t)]+2\pi\phi/360^\circ\}}) \tag{3.7}$$

For the two narrowband stimulation signals, the ones that were filtered in the delta and in the theta ranges, we processed the waveforms further such that all the maxima had the same value, and that the minima had the opposite value. We recently employed such signals in an experimental investigation on the effects of current stimulation on speech comprehension (Keshavarzi et al. 2020a, 2020b). To obtain these waveforms, the amplitude of the analytical envelope, $|E(t)|$, was set to 1 in Eq. 3.7.

For the broadband stimulation waveform, 1 - 20 Hz, we kept its original, non-fixed, amplitude, since this enabled comparison with previous experimental work (Kadir et al. 2019; Wilsch et al. 2018). In addition, processing these waveforms to achieve maxima and minima at equal amplitudes would have introduced major distortion to the signals.

Each phase-shifted envelope $e_\phi(t)$ was then normalized such that no value of the waveform either exceeded 1 or fell below -1. The neurostimulation was simulated in the model by multiplying a particular stimulation waveform by the desired stimulation intensity.

In order to investigate how the temporal alignment of the envelope-shaped stimulation waveform with the acoustic input influenced the speech processing, we employed stimulation waveforms without phase shift (i.e. with a phase shift of 0°) but with different temporal delays. We employed time lags ranging from -250 ms to 250 ms in steps of 50 ms step, with positive lags representing a stimulation waveform that preceded the acoustic stimulation.

3.2.6 Analysis of the phase-amplitude modulation

The spiking neural network was designed such that the phase of the theta oscillations influenced the amplitude of the gamma oscillations. To quantify this coupling, we computed the phase-amplitude modulation index from the LFP (Tort et al. 2010). In particular, we computed the LFP in response to the exemplary sentence '*Alfalfa is healthy for you.*'. The model responses were computed independently 30 times, and the sentence was each time preceded by a random period of silence that ranged from 500 to 1,000 ms. The simulated LFP was then downsampled to 1,000 Hz. It was further subjected to the complex Morlet wavelet transform with frequencies between 1 and 80 Hz, in steps of 0.1 Hz. For each frequency, the extracted amplitudes were binned into 18 bins according to their instantaneous phases. The phase-amplitude modulation index was computed as the Kullback–Leibler divergence (Kullback et al. 1951) of the amplitude distribution across the phase bins from a uniform distribution.

3.2.7 Analysis of syllable parsing

The theta module PIN-TH produced only sparse spiking activity (Fig. 3.1). Spikes that occurred synchronously across different neurons emerged rhythmically in silence and followed syllable boundaries in response to speech. We accordingly employed the model to infer syllable onsets by detecting spike bursts. We thereby defined a spike burst as the spiking activity of at least two inhibitory neurons, which had sparser spiking activity than the excitatory neurons, within a 20 ms window. The precise timing of the syllable onset was assigned according to the maximal firing rate of the Ti neurons, computed using sliding 20-ms-long gaussian window with a standard deviation of 3 ms.

The performance of the resulting syllable parser was assessed by computing the distance, or dissimilarity, between the actual syllable boundaries and those inferred from the activity of the PIN-TH module. We thereby measured the dissimilarity through the non-normalized Victor-Purpura spike distance with a cost parameter of 50 ms (Victor 2005). We only included those inferred syllable onsets that occurred within the duration of the presented sentence, but not those that occurred before the start of the sentence or after it had ended.

For each simulation, the performance of the network was compared to predictions obtained from a simple constant rhythm. The rate of the constant rhythm for this control model was matched to the frequency of the syllable predictions that the theta network generated during the presentation of a sentence. The onset of the constant rhythm was randomly chosen from the same range of 380 - 550 ms, as in the case of model simulations. The performance of the syllable predictions achieved by this constant rhythm was quantified through the dissimilarity of the rhythmic predictions from the actual syllable boundaries, in the same way as for actual syllable predictions. Because the syllable prediction generated by this constant rhythm were not influenced by the simulated speech stimulus, they served to estimate the chance-level performance of predicting the syllable onsets.

A non-dimensional parsing score was then computed by subtracting the distance of the inferred to the actual syllable onsets from the analogous measure achieved by the constant rhythm. A parsing score of 0 accordingly reflected no difference from the prediction performance of the constant rhythm, whereas a positive score indicated a prediction of the syllable onsets from the model that was better than in the control. As an additional control measure, we assessed the syllable parsing when only babble noise was presented to the network, with the actual sentence removed from the acoustic stimulus. The obtained parsing scores provided an additional empirical estimate of the chance level.

3.2.8 Syllable decoding

The excitatory gamma neurons Ge received acoustic input that was pre-processed through a model of the auditory periphery, which decomposed the sound into different frequency bands. The activity of the Ge neurons therefore partly reflected the spectrotemporal information in the

incoming sound. We investigated how well the neural activity encoded the identity of a syllable.

To this end we simulated the neural network response to speech in quiet, as well as to speech in background noise, at various SNRs. We segmented the obtained neural data into subsequent chunks, according to the syllable onsets as inferred from the theta activity described above. Each chunk was assigned the identity of that syllable in the presented sentence during which the corresponding onset was inferred. Moreover, the neural activity in each chunk was characterized by a matrix of pairwise spike distances, for which we employed the non-normalized Victor-Purpura distance with a cost parameter of 60 ms (Victor 2005) (Fig. 3.3A). As a result, each single syllable encoded by the model (Fig. 3.1B, red boxes) was characterized by a particular dissimilarity matrix.

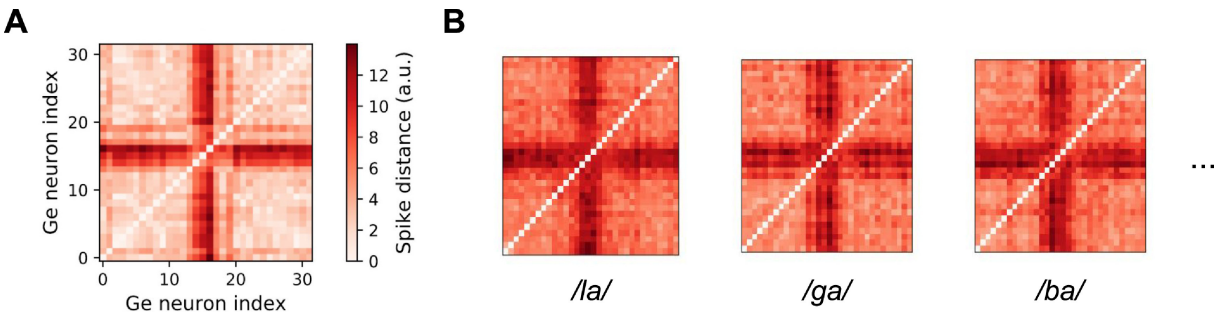

Figure 3.3: Syllable decoding. (**A**) The identity of a syllable was decoded from the corresponding chunk of the neural response of the gamma module. To this end, the outputs of the 32 *Ge* neurons in that segment were characterized by their pairwise dissimilarity matrix. One such matrix was obtained for each syllable parsed by the PIN-TH module of the network (Fig. 3.1B, red boxes). (**B**) The pairwise dissimilarity matrices of the *Ge* neuronal responses differed for different syllables. To decode the identity of an unknown syllable from the neuronal response, its dissimilarity matrix was compared to the averaged dissimilarity matrices for the different syllables, obtained from simulations employing clean speech. The unknown syllable was then assigned the identity of the nearest clean-speech dissimilarity matrix.

Decoding the identity of a syllable then meant to infer the syllable identity assigned to the chunk from its pairwise spike distance matrix. We performed this decoding in two steps. First, we established the neural responses to speech in quiet as the reference neural activity. For each syllable, this reference neural activity was computed by averaging the pairwise spike distances from all chunks of neural data that were associated to that particular syllable (Fig. 3.3B).

Second, we employed a nearest centroid algorithm to decode the identity of a syllable associated with a particular chunk of neural data, which could correspond to speech in noise. The reference pairwise spike distances thereby served as centroids. A chunk of neural data was thus assigned that syllable identity to whose reference pairwise spike distance its own pairwise spike distance was closest to. The distance between two matrices of pairwise spike distance was computed as the root mean square of their difference.

3.2.9 Determining the syllable decoding accuracy

We measured the accuracy of the syllable decoding from the output of the neural network for speech in various levels of background noise, at 11 different SNRs. To this end we performed a large number of trials, in each of which we sought to decode the identity of certain syllables from the network response to speech in noise, using the network response to speech in quiet as a reference.

Neural responses to the speech material were computed as described in Section 3.2.3. In each classification trial, we chose a random subset of ten syllables (classes) amongst which the neural data was to be classified. For each of the ten syllables we gathered all the neural network's responses to that syllable in a given sentence spoken by a particular speaker, at a particular SNR (testing data) as well as without background noise (training data). For each of the ten syllables, we obtained 100 chunks of corresponding neural data, each characterized by its own dissimilarity matrix.

However, due to inaccuracies in the syllable parsing by the PIN-TH module, the chunks of neural data associated to a particular syllable were sometimes more than 100 and sometimes less. In particular, such deviations are expected for shorter syllables or faster speech production rates (Ghitza 2011; Hyafil et al. 2015). To balance the classification problem and to prevent biases, in the former case, we selected a random subset of 100 neural data chunks. In the latter case we selected another subset of 10 syllable labels to be classified, until 100 associated neural data chunks were found for each syllable in the classification trial.

The neural data associated with each syllable, from presenting the sentences in quiet, was then used to establish the reference neural activity. Each chunk of neural data from stimulations employing speech in noise was classified according to the nearest centroid as described above. These predictions were subsequently compared to the actual syllable identities and were averaged to determine the classification accuracy in the decoding trail. Due to the ten different syllables (i.e. classes) that were considered in each trial, the chance level accuracy was 10%.

We performed 200 of such 10-way syllable decoding trials for each of the 11 SNRs for which we simulated the neural network response. The subset of 10 syllables to be classified was chosen at random in each of the 200 trials, but was then kept for each of the SNRs to enable fair comparison between the corresponding syllable decoding accuracies.

3.2.10 Analysis of the effect of SNR on the speech encoding

The dependency of the syllable decoding accuracy A on the different SNRs could be modelled using a four-parameter sigmoid function:

$$A = \frac{A_{max} - A_{min}}{1 + e^{-k(SNR - SNR_0)}} + A_{min}, \tag{3.8}$$

in which A_{min} is the minimal decoding accuracy, achieved for a very small SNR, and A_{max} is the maximal decoding accuracy, resulting from a very high SNR. SNR_0 is the SNR at which the decoding accuracy is the average of the maximal and the minimal value, that is, the SNR at which the decoding accuracy is halfway between A_{min} and A_{max}. SNR_0 may therefore be related to the 50% speech reception threshold (SRT) that is commonly used to quantify the level of speech-in-noise comprehension in behavioural experiments. k determines the slope of the curve at SNR_0.

To obtain the model parameters of Eq. 3.8, as well as their confidence intervals, we employed a bootstrapping procedure (Davison et al. 1997). The 200 trials of syllable decoding, performed for the eleven different SNRs, resulted in 2,200 datapoints. We resampled these 10,000 times with replacement, and each time computed the parameters of the sigmoidal fit through non-linear least squares (Levenberg-Marquardt algorithm (Marquardt 1963)). We thereby obtained empirical distributions for each model parameter. The mean value of each parameter followed as the mean of the corresponding distribution, and the associated $(100-n)\%$ confidence interval was computed as the range between the distribution's $(\frac{n}{2})^{th}$ and the $(100-\frac{n}{2})^{th}$ percentile. The optimal curve fitted to the data and its confidence bands were computed from these values.

We modelled the effect of background noise on the syllable parsing score through a sigmoidal function as well. The parameters of the sigmoidal fit and their confidence intervals were determined analogously to dependence of the syllable decoding accuracy on the SNR set out above.

3.2.11 Quantifying the contributions of spectral cues to the speech encoding in the model

To identify the contributions of spectral cues to the syllable parsing and encoding in the model, we repeated the simulations of speech in background noise, but with randomly shuffled auditory channels. Specifically, for each simulation of the model, the 32 auditory channels that contained the auditory input were randomly re-ordered. The time course of each channel remained unchanged, so that the net acoustic input to the model remained the same as for the original stimulus.

The shuffled acoustic inputs were then processed in the model as specified in Section 3.2.4. In particular, the randomly shuffled auditory channels were projected to the Ge neurons and to the population of relay neurons, which provided input to the population of Te neurons. The model simulations employed the same sentences and SNRs as in the previous experiment (see Section 3.2.3 for details). Syllable parsing and decoding were analysed as described in Sections 3.2.7 - 3.2.10. In particular, the model simulations of the original, unshuffled, clean sentences were used as a reference to evaluate the syllable decoding accuracy of the shuffled input.

3.2.12 Modelling the effects of external electrical stimulation on the speech encoding

To investigate the effects of external electrical stimulation on the encoding of speech in the model, we ran the same model simulations as specified in Section 3.2.3, but this time simulating the application of external alternating current as well. The stimulation waveforms used and their alignment with respect to the acoustic input were specified in Section 3.2.5.

The analysis of the syllable parsing and decoding was the same as described in Sections 3.2.7 - 3.2.10. Importantly, for syllable decoding, the stimulation waveform was applied also when speech without background noise was simulated in the model. This meant that the centroids of the syllable classifier were computed from speech in quiet, but with added current stimulation. We chose this approach because the neural encoding of speech, including speech in quiet, was likely affected by the applied current. Our goal was, however, to assess the impact of current stimulation on the network's encoding of speech in noise, and not on speech in quiet. We therefore employed the neural responses to speech in quiet during current stimulation as a reference to assess how the applied stimulation influences the consistency of the neural code across SNRs for a given type of stimulation.

3.3 Results

3.3.1 Intrinsic network activity

The PIN-G and the PIN-TH modules in the network generated self-sustained rhythmic activity in the gamma (25-40 Hz) and in the theta (4-8 Hz) frequency range, respectively (Fig. 3.1B). Through the unidirectional coupling from the Te to the Ge neurons (Fig. 3.1A, red), the theta rhythm modulated the faster gamma activity. In particular, each burst of spikes generated in the PIN-TH module reset the phase of the faster gamma oscillations (Fig. 3.1B). We quantified this coupling through computing the phase-amplitude modulation index between the LFP of the PIN-TH module and the LFP of the PIN-G module, when processing an exemplary sentence preceded by a period of silence and without additional current stimulation (Fig. 3.1C). We found that the neural activity between 5 - 12 Hz modulated the faster activity in the gamma band, above 25 Hz.

3.3.2 The neural network's encoding of speech in noise

When the network was presented with speech, the theta rhythm aligned to the syllable onsets (Fig. 3.1B). We quantified this alignment by computing a syllable parsing score, and used it to systematically quantify how well the network parsed syllables when speech was presented in different levels of babble noise. To estimate the empirical chance level, we presented the neural network with babble noise alone, and computed the syllable parsing score that would have been associated with the missing speech signal.

For the lowest SNR that we considered, -25 dB, the syllable parsing score attained a very low value of -0.03 ± 0.05 (mean and 95% CI, Fig. 3.4A, blue). This was comparable to the empirically estimated chance level of 0.01 ± 0.05 (mean and 95% CI, Fig. 3.4A, red). The syllable parsing at this low SNR was therefore insignificant. However, SNRs of -10 dB or higher led to syllable parsing score that exceeded the chance level. For the highest SNR of 25 dB that we simulated, the score reached 0.88 ± 0.05 (mean and 95% CI).

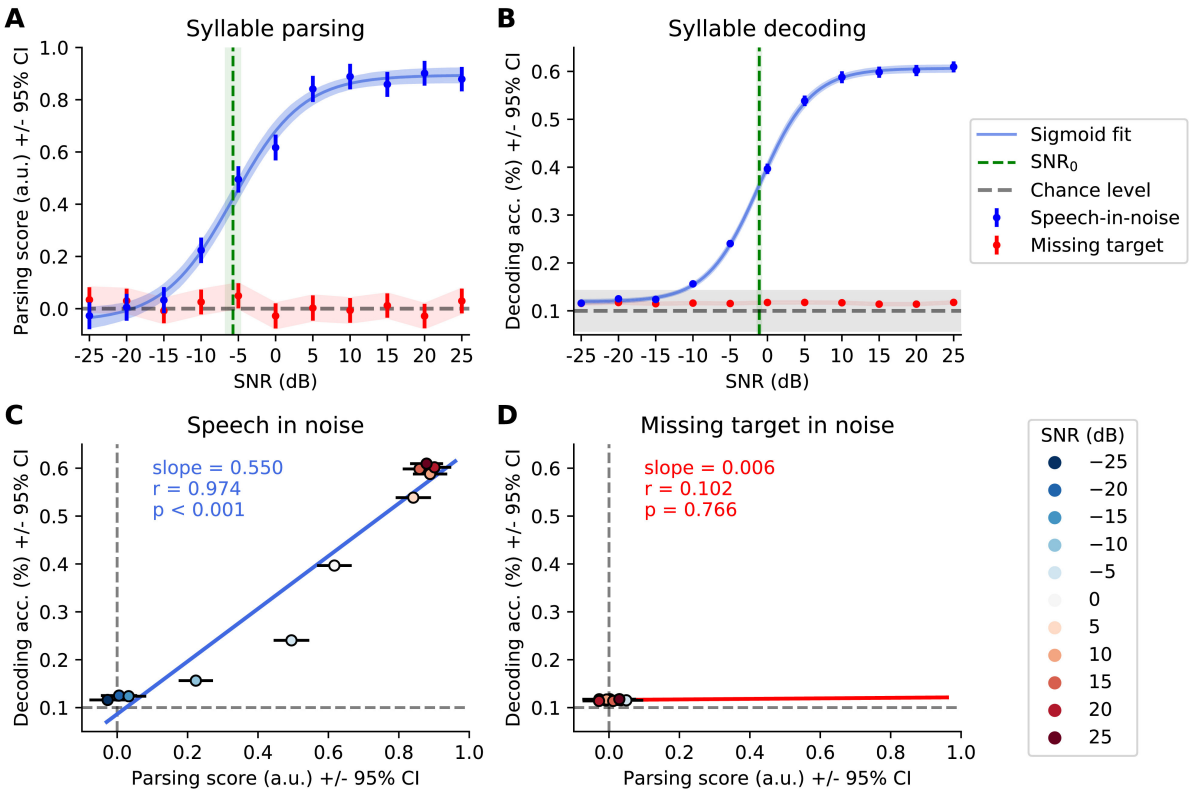

Figure 3.4: Speech-in-noise encoding in the model. (**A**) The syllable parsing by the theta module (blue) was at chance level (grey) for high levels of background noise (low SNR), but exceeded chance level for SNRs above -15 dB. It saturated at a value of around 0.9 for high SNRs, following a sigmoidal relationship with an inflection point at the SNR of -5.7 dB (green). Syllable parsing did not exceed the chance level when the speech signal was absent from the acoustic input (red). (**B**) The accuracy of the syllable decoding (blue) from the neural response of the gamma module exhibited a sigmoidal dependence on the level of the background noise as well. The decoding accuracy was above the chance level (grey) when the SNR was -10 dB or higher. The inflection point of the sigmoidal fit occurred at an SNR of -1.1 dB (green). No significant syllable decoding could be achieved when the speech signal was removed from the background noise (red). (**C**) The syllable decoding accuracy increases monotonously with the syllable parsing score, with increasing SNR. The correlation between the two measures is statistically highly significant ($p = 4{\cdot}10^{-7}$). (**D**) A control computation in which syllable parsing and syllable decoding are obtained from sound mixtures in which the target speech signal has been removed shows performance that is only at the chance level (grey).

To interpret the magnitudes of the parsing scores, we computed the maximal parsing score,

which followed from the true syllable onsets. We obtained a maximal parsing score of 9.21 ± 0.02 (mean and 95% CI). Likewise, the parsing score of 0 reflected an insignificant parsing that was equal to that of the null model (Fig. 3.4A, grey dashed). The maximal parsing scores obtained from the spiking neural network were therefore only about 10% of the maximal possible value, that is, the one that would result from perfect alignment of the predicted and actual syllable onsets.

The dependence of the parsing score on the SNR could be fitted well by a sigmoidal curve (Fig. 3.4A, blue). The inflection point of the sigmoid, that is, the SNR at which the syllable parsing score was midway between the minimal and the maximal value, occurred at -5.7 dB $\pm$ 1.0 dB (mean and 95% CI).

The excitatory neurons of the PIN-G module, the *Ge* neurons, were influenced by the PIN-TH module. At the same time, the *Ge* neurons were stimulated by the sound as well, in a tonotopic fashion (Fig. 3.1A). While the PIN-TH module could parse syllables, the neuronal activity of the faster PIN-G module could therefore encode the identity of the corresponding syllable. We determined the accuracy of the syllable encoding by assessing how well syllables could be decoded from the spiking activity of the *Ge* neurons.

Because we decoded syllable identities out of ten possible choices, the chance level for the decoding accuracy was 10%. We verified this chance level by assessing the syllable decoding when only background noise was presented to the neural network. This yielded a decoding accuracy of $11.6\% \pm 0.3\%$ (mean and 95% CI), approximately in line with the chance level (Fig. 3.4B, red).

We found that the accuracy of the decoding of syllables in background noise, as a function of the SNR, followed a sigmoidal curve (Fig. 3.4B, blue). For the lowest considered SNR of -25 dB, the decoding was poor, with an accuracy of $11.6\% \pm 0.6\%$ (mean and 95% CI). This low accuracy exceeded the chance level of 10% only slightly.

The largest SNR that we simulated, 25 dB, led, in contrast, to a high decoding accuracy of $61.0\% \pm 1.1\%$ (mean and 95% CI). Indeed, the decoding accuracy exceeded the chance level already for the comparatively low SNR of -10 dB, as well as for higher SNRs. Fitting a sigmoid to the dependence of the decoding accuracy on SNR showed that the inflection point of the curve was at a SNR of -1.1 $\pm$ 0.2 dB (mean and 95% CI).

We also investigated the relationship between syllable parsing scores and syllable decoding accuracies (Fig. 4C). We found a strong positive correlation between the two measures (Pearson's $r = 0.97$, $p < 10^{-6}$). Low parsing scores were accordingly associated to low accuracies of syllable decoding and vice versa. The slope of a linear fit was 0.55.

However, the relation between the two scores was not exactly linear. Instead, intermediate SNRs led to relatively higher syllable parsing scores than the syllable decoding accuracies. This

behaviour reflected our earlier finding that the inflection point of the sigmoidal dependence of the syllable parsing scores on the SNR occurs at a lower SNR, -5.7 dB, than that of the decoding accuracy, -1.1 dB.

As a control, we also assessed the correlation between the syllable parsing score and the decoding accuracy when both were obtained from the background babble noise (Fig. 3.4D). As expected, the resulting scores were low and not significantly correlated (Pearson's $r < 0.1$, $p = 0.8$).

3.3.3 Quantifying the contributions of spectral cues to the speech encoding in the model

To investigate the contributions of frequency-specific cues to the model's speech encoding, we shuffled the auditory channels of the acoustic inputs in model simulations. We compared the obtained syllable parsing scores and decoding accuracies with the case when the network was encoding the original acoustic input (Fig. 3.5).

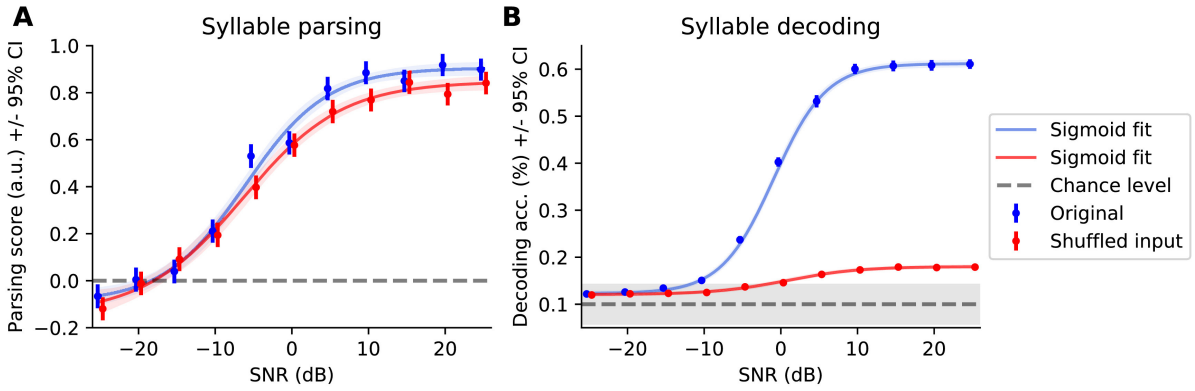

Figure 3.5: Encoding of speech with shuffled auditory channels. The figure depicts the syllable parsing scores (**A**) and the decoding accuracies (**B**) obtained for the original acoustic input (blue) or when the auditory channels were randomly shuffled (red). The dashed grey lines represent the chance level for each score, and error bars depict 95% confidence intervals. (**A**) The syllable parsing scores remained at approximately the same level when auditory channels were shuffled, especially for lower SNRs (between -25 and -5 dB). A discrepancy between the shuffled and the original inputs occurred for SNRs above -5 dB. (**B**) In contrast to the syllable parsing, syllable decoding accuracy decreased substantially when the auditory channels were shuffled. In particular, the syllable decoding of the shuffled input remained at or only slightly above chance level for all SNRs. The syllable decoding of the original speech input, however, was significantly higher than the chancel level for SNRs above -10 dB.

Shuffling the auditory channels influenced the syllable parsing and decoding differently (Fig. 3.5). Syllable parsing was not affected strongly by the shuffling, and its dependence on the SNR of the spectrally-shuffled input was comparable to that of the original acoustic signal (Fig. 3.5A). In particular, for low SNRs, below -5 dB, the results were almost identical. For

SNRs above -5 dB, shuffling of the auditory channels led to a slight decrease in performance. The largest difference in the parsing scores between the shuffled and the original acoustic input, a difference of 0.092 a.u., was observed for a SNR of approximately 5 dB . For a SNR of 25 dB, the parsing scores from the two conditions remained different, but the discrepancy between them was smaller (0.063 a.u.).

For syllable decoding, however, the shuffling of auditory channels led to a major deterioration of the classification accuracy (Fig. 3.5B). Similarly to the syllable parsing scores, for the very low SNRs below -10 dB, the decoding accuracy for both the shuffled and the original input was similar and did not exceed chance level. For SNRs above -10 dB, the results obtained from the two types of input started to diverge. Notably, the syllable decoding accuracy for the shuffled input (Fig. 3.5B, red) did not exceed the chance level below approximately 0 dB SNR. Even at a SNR of 25 dB it remained substantially below that of the original, non-shuffled, input, reaching only 18.0% $\pm$ 0.5% accuracy (mean and 95% CI).

3.3.4 The effects of the external current stimulation on speech processing in the model

We assessed the effects of the external current stimulation with the speech envelope on the network's encoding of speech stimuli. We investigated three main types of current waveforms: one type that was based on the broad-band speech envelope, a second type that was based on the delta-band portion, and a third type that was based on the theta-band portion of the speech envelope (Fig. 3.2). For each of these three types, we then considered six different phase shifts. Because the waveforms of each type encompassed more than a single frequency, these phase shifts differed from temporal delays.

In addition, we considered eleven time delays that ranged from -250 ms to 250 ms with 50 ms step. Positive time lags thereby meant that the stimulation onset preceded the sentence that was presented to the model. The phase of the time-shifted waveforms was not manipulated, such that their phase shift was 0°.

Each waveform was applied at three different intensities of 0.1 pA, 0.2 pA and 0.5 pA.

The effects of the external current stimulation on syllable parsing

To quantify the influence of the applied stimulation waveforms on the syllable parsing, governed by the slower theta rhythm in the model, we assessed the obtained parsing scores at the SNR of 0 dB. For this SNR, model simulations without additional current stimulation yielded a parsing score of 0.65 $\pm$ 0.05 (mean and 95% CI) (Fig. 3.4A).

For each applied stimulation waveform, we obtained the parsing score at 0 dB SNR and compared them with the case when no stimulation was applied to the model. The Wilcoxon

signed rank test (Wilcoxon 1992) was used to assess whether the difference between the two was significant. We then applied the Benjamini-Yekutieli correction for false discoveries from multiple comparisons to the obtained p-values (Benjamini et al. 2001). The significance threshold for hypothesis testing was set to $p = 0.05$.

For the stimulation waveforms shifted in phase, the theta-band stimulation provided the largest difference to the case without stimulation, consistently across all considered stimulation intensities (Fig. 3.6A, C, E). The theta-band stimulation notably outperformed the other stimulation waveforms for the phase shifts of 180° and 240°, and provided the largest improvement of the parsing scores. Delta-band stimulation yielded slightly larger improvement than the broadband waveform for the phase shifts of 180° and 240°. Overall, the effects of the applied stimulation provided phase-dependent modulation, which remained consistent across stimulation intensities. For all types of stimulation waveforms phase shifts ranging from 0° to 120° typically led to the decrease in the parsing scores. In turn, phase-shifts ranging from 180° to 300° facilitated the syllable parsing. While the phase-dependent modulation was observed for all the waveforms, the strength of the modulation varied depending on the frequency of the stimulation waveform.

For the stimulation waveforms shifted in time, the effect on the syllable parsing depended on the frequency of the stimulation waveform (Fig. 3.6B, D, F). For theta-band stimulation, the largest improvement was observed for a delay of 50 ms, that is, when the onset of the stimulation preceded the onset of the acoustic input by 50 ms. Additional significant improvements in parsing scores were observed for time lags of -100 ms, 100 ms and 250 ms. Interestingly, the difference between the two pairs of beneficial lags was 150 ms, corresponding to a frequency of approximately 6.67 Hz. In turn, the negative effects of the theta-band stimulation on the parsing scores were observed for -200, -50, 0 and 150 ms. As for the beneficial time lags, the difference between two successive delays were therefore 150 ms as well.

For the delta-band stimulation, the time lag that led to the largest improvement of the parsing score was between 50 and 100 ms, depending on the stimulation intensity. Similarly to the stimulation with different phase shifts, the effects of the delta-band stimulation at the best time lag were smaller than for the theta-band stimulation, but were larger than the broadband stimulation, across all stimulation intensities. Delta-band stimulation that proceded the acoustic input, that is, at negative time lags, led to a decrease of the parsing scores. The size of this decrease depended on the stimulation intensity and was comparable to that of the theta-band stimulation.

Stimulation with the broadband waveforms shifted in time influenced syllable parsing the least. Notably, only the time lag of 100 ms led to a consistent improvement of parsing scores across all three stimulation intensities. For the two higher stimulation intensities, the time lags of 50 ms (at 0.2 pA, 0.5 pA) and 250 ms (at 0.5 pA) also facilitated syllable parsing, but not as strongly as at the delay of 100 ms.

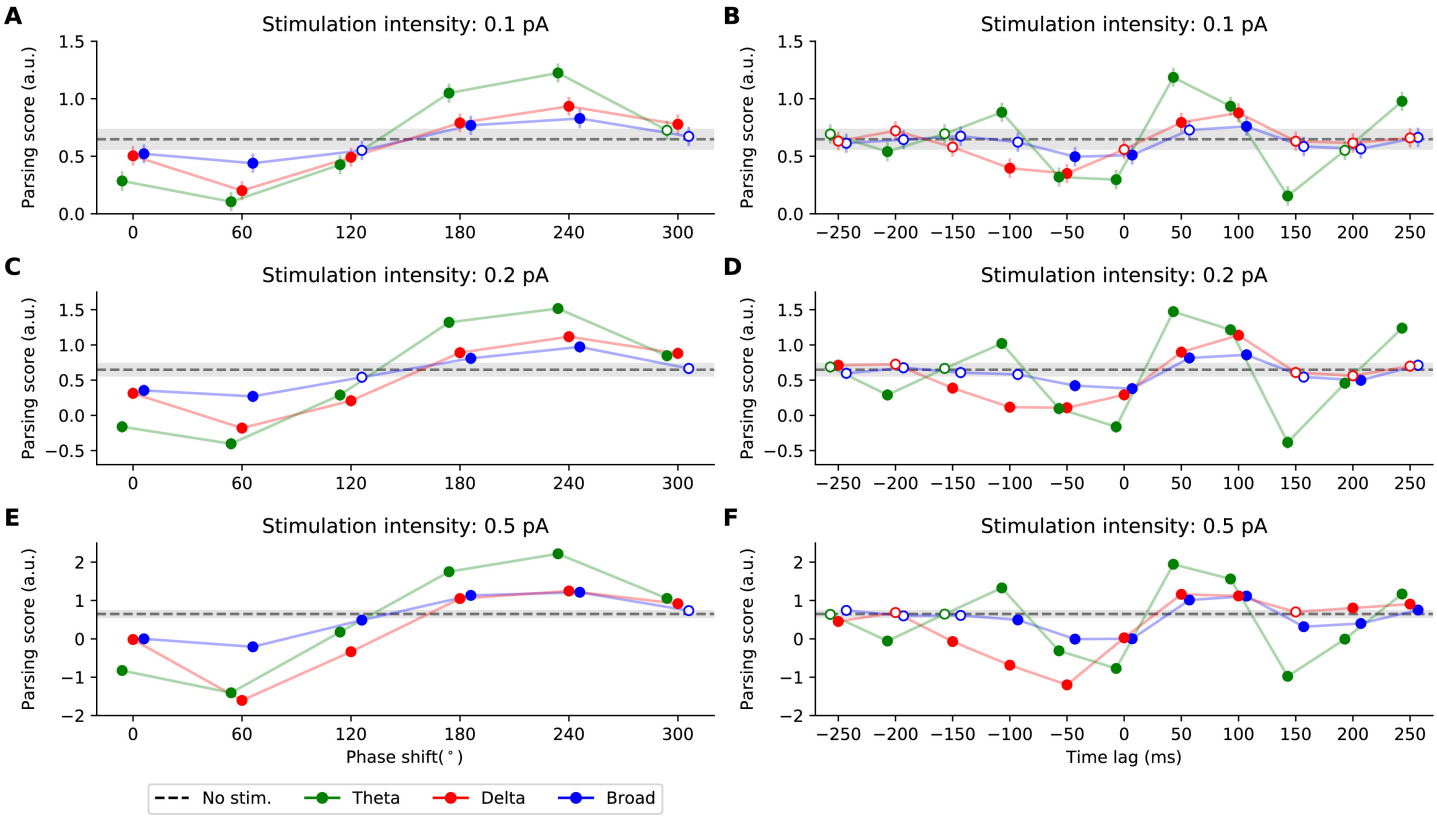

Figure 3.6: The effects of the external current stimulation on the syllable parsing. The syllable parsing scores during current stimulation were computed for speech in background noise, at a SNR of 0 dB. We computed the syllable parsing scores for broad-band stimulation (blue), delta-band stimulation (red), and theta-band stimulation (green), and compared the results to the case of no stimulation (black dashed line). The stimulation waveforms were either shifted in phase (**A**, **C**, **E**) or in time (**B**, **D**, **F**) with respect to the acoustic input. Positive time lags represent stimulation onset preceding the neural processing. Each row of panels shows results for a different stimulation intensity, and the error bars and shaded areas represent 95% confidence intervals. Parsing scores that differ significantly from those obtained without stimulation are indicated by coloured disks ($p < 0.05$, FDR correction for multiple comparisons). Stimulation at phase shifts of about 240° as well as at time shifts of about 50 ms typically enhance the syllable parsing, whereas phase shifts of 60° as well as time shifts of about -50 ms lead to a worsening of the syllable parsing.

The effects of the external current stimulation on syllable decoding

To assess the neural network's speech encoding during stimulation, we measured the syllable decoding accuracies at the SNR of -1.1 dB. This SNR yielded, without current stimulation, a decoding accuracy of 36.4% ± 0.7% (mean and 95% CI) that was halfway between the minimal and the maximal accuracy (Fig. 4B).

For each type of stimulation waveform, we then established whether the obtained syllable decoding accuracy was significantly different from the one that resulted in the absence of current

stimulation. To this end, we obtained the empirical distribution of the syllable decoding accuracies without current stimulation at -1.1 dB SNR through a bootstrapping procedure as described in Section 3.2.10. This empirical distribution represented the lack of the effects of the applied stimulation. For each stimulation waveform, it was then compared to the syllable decoding accuracy that resulted when current stimulation was applied, to establish an empirical p-value (two tailed) for the obtained decoding accuracy. We then applied the Benjamini-Yekutieli correction for false discoveries from multiple comparisons to the obtained p-values. The significance threshold for hypothesis testing was set to $p = 0.05$.

The lowest stimulation intensity that we considered was 0.1 pA, leading to a depolarization of the membrane potential of a stimulated isolated neuron by 1 mV. Stimulation at such a low intensity led only to significant change in the decoding accuracy for the delta-band stimulation (Fig. 3.7A, B). In particular, delta-band stimulation at certain phase shifts and time shifts worsened the syllable decoding: a phase shift of 180° resulted in a lower decoding accuracy of 35.0% ± 0.7%, a phase shift of 240° reduced the decoding accuracy to 35.2% ± 0.7%, a time lag of -150 ms yielded a decoding accuracy of 34.9% ± 0.6%, and a time shift of 250 ms lowered the decoding accuracy to 35.2% ± 0.7%.

We then investigated an intermediate stimulation intensity of 0.2 pA, leading to a 2-mV depolarization of an isolated neuron. This led to significant differences in the syllable decoding accuracy, as compared to no stimulation, for theta- and delta-band stimulation, but not for the broadband current waveforms (Fig. 3.7C, D). In particular, theta-band stimulation yielded significant improvement for phase shifts of 0°, resulting in a syllable decoding accuracy of 37.8% ± 0.7%, and 60°, increasing the accuracy to 37.8% ± 0.8%, as well as at a 0 ms time lag, yielding a decoding accuracy of 37.8% ± 0.7%. In turn, delta-band stimulation yielded significant improvement only for 300° phase shift, increasing the decoding accuracy to 37.5% ± 0.7%. It also led to a significant decrease in decoding accuracy, namely for a 180° phase shift that yielded an accuracy of 35.3% ± 0.7% as well as for a 150 ms time lag that lowered the accuracy to 34.9% ± 0.7%.

At the highest considered stimulation intensity of 0.5 pA, the narrow-band stimulation waveforms had, at most phase shifts, a significant impact on the network's speech coding, while the broadband stimulation yielded significant effects only for a couple of phase and time shifts (Fig. 3.7E, F). The largest improvement of about 5% in the syllable decoding accuracy emerged for the theta-band stimulation aligned with the sentence onset, that is, without a shift in phase or time. This improvement was slightly larger than that observed for the delta-band stimulation without phase- or time shift, and was substantially larger than that resulting from the broadband stimulation.

3.4 Discussion

We investigated the influence of alternating current stimulation with the speech envelope on the neural processing of speech in background noise, using a computational model of a spiking

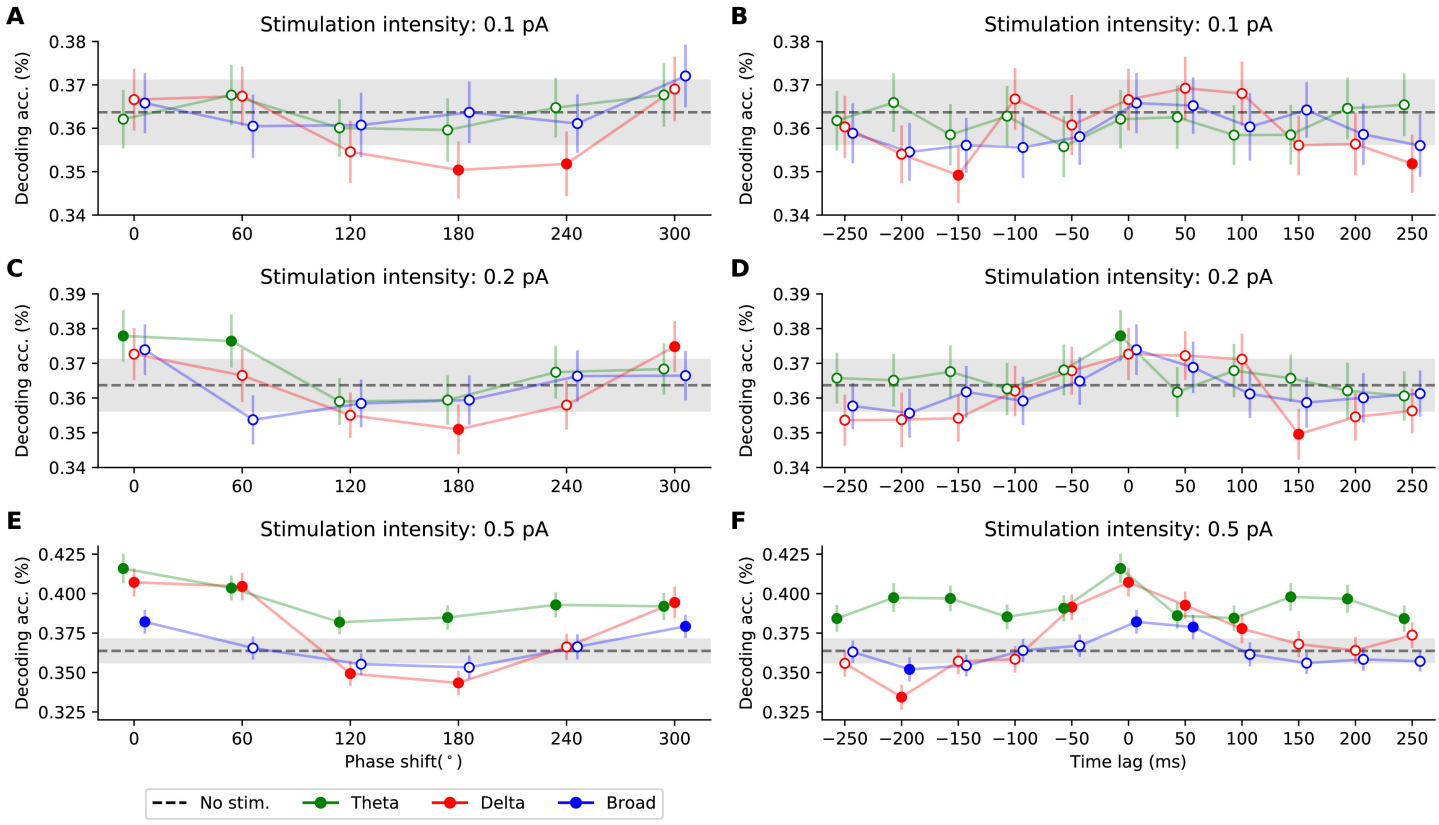

Figure 3.7: The effects of the external current stimulation on the syllable encoding. Syllable decoding accuracies during current stimulation were computed for speech in background noise, at a SNR of -1.1 dB. We computed the syllable decoding accuracies for broad-band stimulation (blue), delta-band stimulation (red), and theta-band stimulation (green), and compared the results to the case when no stimulation was applied to the model (black dashed line). The stimulation waveforms were either shifted in phase (**A**, **C**, **E**) or in time (**B**, **D**, **F**). Positive time lags represent stimulation waveform preceding the acoustic signal. The error bars and shaded areas represent the 95% confidence intervals. Decoding accuracies that differed significantly from the case when no stimulation was applied to the model are indicated by coloured disks ($p < 0.05$, FDR correction for multiple comparisons).

neural network. We characterized the network's speech encoding through two measures, the syllable parsing score and the accuracy with which the syllable identity could be decoded from the neural activity. We found that both measures increased with increasing SNR, following a sigmoidal curve. This behaviour resembled psychometric curves of speech comprehension measured behaviourally (Nilsson et al. 1994; Plomp et al. 1979; Spyridakou et al. 2020). An important characteristic of each sigmoidal curve was the inflection point, that is, the SNR at which the corresponding measure—the syllable parsing score or the syllable decoding accuracy—was midway between the lowest and the highest value. This inflection point occurred, for the syllable decoding accuracy, at an SNR of -1.1 dB (Fig. 3.4B). At this SNR, the human comprehension of speech in babble noise is about 50%, suggesting that the neural network's speech encoding may capture certain aspects of the neural mechanisms through which humans understand speech in noise.

The inflection point of the sigmoidal curve occurred at a lower SNR of -5.7 dB for the syllable parsing scores, suggesting that syllable parsing was somewhat more robust in the presence of background noise than syllable decoding. Our further investigation employing spectrally-shuffled versions of the acoustic input showed that the syllable parsing depended mainly on the slow amplitude fluctuations in the acoustic input rather than on frequency-specific features (Fig. 3.5A). On the contrary, the accuracy in the syllable decoding task deteriorated almost completely when the frequency channels of the acoustic input were randomly shuffled (Fig. 3.5B). Syllable decoding therefore relied mostly on the frequency-specific information in the acoustic input. These differences between the syllable parsing and the syllable decoding highlight the two distinct mechanisms by which both tasks are accomplished in the model, either through the neural output of the PIN-TH module, or through that of the PIN-G network (Fig. 3.1).

We investigated the effects of the stimulation waveforms that were derived from the speech envelope on the speech encoding in the model. The stimulation waveforms were either narrow- or broadband, and were simulated with different stimulation intensities. In order to quantify the impact of the alignment of the stimulation waveform and the acoustic input, we shifted the stimulation waveform in either phase or time.

With increasing stimulation intensity, the effects on syllable parsing and syllable decoding increased as well (Figs. 3.6, 3.7). For syllable parsing, the phase-shifts of the stimulation waveforms led to a modulation pattern that was periodic in the phase. Phase shifts between 0-120° led to a decrease of parsing scores, and phase shifts of 180-300° improved them (Fig. 3.6A, C, E). Importantly, this behaviour was consistent for the delta-band, theta-band and broadband stimulation, although the size of the effect depended on the frequency band. Specifically, we observed the largest phase-dependent modulation of syllable parsing for the theta-band stimulation, a moderate one for the delta-band waveforms, and the smallest one for the broadband type stimulation.

These findings parallel recent experimental results on a phase-dependent modulation of speech-in-noise comprehension through transcranial current stimulation with the speech envelope (Kadir et al. 2019; Keshavarzi et al. 2020a; Riecke et al. 2018; Zoefel et al. 2018). These experiments reported significant effects of tACS on speech comprehension only for theta-band stimulation, which may correspond to our finding of the theta-band stimulation having the largest effect on the speech encoding in the model (Keshavarzi et al. 2020a).

In contrast, time shifts of the stimulation waveform led to modulations of the syllable parsing that depended on the frequency of the stimulation waveform (Fig. 3.6B, D, F). All of the considered stimulation waveforms yielded improved syllable parsing scores when the stimulation preceded the acoustic input by 50 ms to 100 ms. This time lag matches the neural delay of the early cortical processing of speech (Brodbeck et al. 2020b; Kubanek et al. 2013).

To interpret our results regarding the temporal delays, it is important to note that our model

did not include a temporal delay between the acoustic input and the neuronal modules, with the exception of the relay neurons that fed into the theta network and that had delays between 0 and 50 ms. However, the neural responses in the human auditory cortex exhibit delays between approximately 20 ms and hundreds of ms (Pickles 1998). In particular, the primary auditory cortex responds largely at delays between 50 to 100 ms. If we assume that our model of a spiking neural network corresponds to a part of the primary auditory cortex, we therefore need to account for an additional delay of 50 ms to 100 ms in the neural response with respect to the acoustic signal. The maximal enhancement of the syllable parsing score would then occur for neurostimulation waveforms that had no time shift with respect to the acoustic signal. This interpretation of our model prediction agrees with recent behavioural experiments that found that theta-band stimulation with no temporal delay led to the largest improvement in speech comprehension (Keshavarzi et al. 2020a, 2020b).

The magnitude of the enhancement of syllable parsing at the time shifts of 50 ms to 100 ms depended on the type of stimulation waveform. Consistent with the results obtained for the phase-shifted waveforms, the largest improvement was obtained for theta-band stimulation, followed by its delta-band counterpart, while broadband stimulation yielded the smallest improvements.

Notably, only theta-band stimulation waveform led to substantial improvement of parsing scores for the two further temporal delays, at approximately -100 ms and at approximately 250 ms. Both time lags differ from the time lags that led to the highest syllable parsing score, 50 ms to 100 ms, by 150 ms to 200 ms. This apparent periodicity in the modulation of the syllable parsing score at a period of 150 ms to 200 ms may reflect the periodicity in the theta-band waveform. The period of 150 ms to 200 ms corresponds indeed to a frequency between 5 Hz and 7 Hz, so entirely in the theta frequency range.

Our observation of theta-band stimulation yielding the largest enhancement of syllable parsing presumably reflects the fact that the theta-band stimulation had a frequency range similar to the intrinsic activity of the PIN-TH network, about 5-10 Hz. The theta-band stimulation could therefore efficiently entrain the oscillations in the theta module on the per-cycle basis (Herrmann et al. 2016). Delta- and broadband stimulation waveforms could entrain theta oscillations as well, and consequently influence syllable parsing, but to a smaller extent, especially regarding an enhancement. For the delta-band stimulation, this was likely due to the mismatch between the frequency bands: the delta-band frequencies, at 1 - 4 Hz, were subharmonics of those that occurred in the intrinsic activity of the PIN-TH module. As a result, one cycle of the applied stimulation affected, on average, two cycles of the theta-band oscillations tracking syllable onsets, leading to a weaker entrainment and associated improvement in syllable parsing. Because the broadband stimulation included both the theta and the delta band, it presumably led to interferences between the two and therefore to a further weakening of the effect. Moreover, the delta- and theta-band waveforms but not the broadband waveforms had been processed so that their maxima and minima occurred at the same values, which might have further increased the

efficacy of the delta- and theta stimulation.

Syllable decoding was overall affected by the current stimulation to a lesser degree than syllable parsing (Fig. 3.7). In particular, the lower stimulation intensities of 0.1 pA and 0.2 pA yielded barely a significant modulation of the syllable decoding. At the highest stimulation intensity of 0.5 pA, the effect depended on the phase and time shifts. The largest improvement in the syllable decoding accuracy was achieved when the applied waveform was aligned to the speech signal, without an additional phase shift, whereas the opposite phase shift of 180° yielded the worst syllable decoding accuracy. This parallels recent experimental findings that have found the phase-dependent modulation of speech-in-noise comprehension due to current stimulation with the significant improvement for a phase shift of 0° and the worst performance obtained for a phase shift of 180° (Keshavarzi et al. 2020a, 2020b).

Somewhat unexpectedly, at the highest stimulation intensity, theta-band stimulation consistently improved the decoding accuracy for all the phase and time lags that we considered. This result presumably reflected the matching of the theta stimulation to the intrinsic rhythm of the theta module that parsed the syllables. Because the syllable decoder was established using the model's response to clean sentences under a certain type of stimulation, the decoding scheme emphasized the consistency of the neural code across SNRs under certain stimulation condition. Since the effects of the theta-band stimulation on the parsing of syllables in the model were overall the strongest, the encoding of speech could therefore benefit from theta-band stimulation, for different time and phase shifts. However, no such effect was observed for the delta- and broadband stimulation waveforms, whose influence on the syllable parsing was notably weaker.

The syllable decoding under the strongest theta-band stimulation depended nonetheless on both phase and time shifts. The largest enhancement of the decoding accuracy was obtained in the absence of either phase or time shifts, that is, at shifts of 0° and 0 ms. Regarding phase shifts, the worst performance was obtained when the waveform was shifted by 120° - 240°. Regarding time shifts, the performance decreased symmetrically for both negative and positive shifts up to ± 100 ms, and then increased again towards peaks between ±150 ms to ±200 ms. The emergence of these peaks that differed from the largest peak at 0 ms by 150 ms to 200 ms was reminiscent of the dependence of the syllable parsing score on the time shifts, for the theta-band stimulation (Fig. 3.6F). As in that case, the dependence of the syllable encoding on the time lags likely reflected periodicity in the intrinsic rate of the syllable-parsing PIN-TH module, between 5-10 Hz, yielding a period between 100 ms to 200 ms.

For the strongest stimulation intensity of 0.5 pA, and without phase or time lag, the delta-band stimulation yielded an enhancement of the syllable decoding accuracy that was only slightly below that of the theta stimulation. However, as opposed to the stimulation with the theta-band portion of the speech envelope, the delta-band stimulation could also decrease the accuracy, such as at phase shifts of 120° and of 180° as well as at a time lag of -200 ms. The significant decrease in syllable decoding at these phase and time lags reflects a substantial deterioration of

the consistency of the neural code across SNRs during those stimulations.

Broadband stimulation at the intensity of 0.5 pA yielded substantially smaller effects on syllable decoding accuracy than both the delta- and the theta-band stimulation. We found significant improvements of the accuracy only for the phase shifts of 0° and of 300°, as well as for time shifts of 0 ms and 50 ms. Similar to the delta-band stimulation, the only significant negative effect was observed for a time lag of -200 ms, although the effect was smaller.

Surprisingly, the best and worst stimulation phases and time lags for the syllable parsing differed from those of the syllable decoding accuracies. The best syllable parsing was obtained for a phase shift of 240°, yielding a phase advance of 120° with respect to the unshifted waveform. We obtained similarly the worst syllable parsing at a phase shift of 60°, also at a phase advance of 120° as compared to the phase at which the worst syllable decoding accuracy occurred. Regarding the time delays, the largest improvement in the parsing scores were obtained for either 50 ms (theta-band stimulation) or for 100 ms (delta- and broadband stimulation). The best syllable decoding resulted in the absence of a time delay.

These systematic phase and time differences between the influence of the current stimulation on the syllable parsing and on the syllable decoding were unexpected. They reiterate that the parsing of syllables and the encoding of the syllable content in the network activity were governed by distinct mechanisms, implemented by the two modules constituting the network (Fig. 3.1). First, the activity in each module was likely influenced by the external current in a different way. In particular, the frequency of the theta rhythm was similar to that of the exogenous, envelope-shaped, stimulation waveform, and could therefore be entrained by the latter (Herrmann et al. 2016). The gamma activity, in contrast, had higher intrinsic frequency and therefore the substantially slower external alternating current stimulation waveform could only temporally modulate, rather than directly entrain, the activity of neurons making up the PIN-G network (Fröhlich et al. 2010).

Second, the influence of the current stimulation on the two tasks of syllable parsing and syllable decoding differed as well. In particular, the speech envelope reflects mainly the voiced parts of speech, which generally have a larger amplitude than the voiceless parts (Biesmans et al. 2016; Grant et al. 1985; Shannon et al. 1995). Syllables begin, however, often with a voiceless part. Even in syllables beginning with a voiced part, their onsets precede the majority of their energetic content. The speech envelope, shifted to have a phase or time advance, therefore aligns better with the syllable onsets than the unshifted envelope. A phase or time advance of the current stimulation could accordingly lead to better syllable parsing in the model. In contrast, the syllable decoding from the model output relied mostly on the voiced parts of speech, which yield larger activations of auditory channels than the voiceless parts (Chi et al. 2005). The peaks of the speech envelope aligned with the acoustic input therefore coincided with the stimulus-driven current delivered to the PIN-G module, what in turn facilitated the syllable encoding.

Our model therefore suggests that the stimulation waveforms that are optimal for syllable parsing are not optimal for syllable decoding, and vice versa. However, syllable decoding partially depends on syllable parsing. The different influence of the current stimulation on both processes accordingly implies that the neurostimulation's effect on speech encoding in the model is partly inhibited by these interferences.

An important limitation of our model is that it operates only in a feed-forward fashion. The acoustic stimuli and the current stimulation serve as input to the model. The PIN-TH module parses syllables and feeds forward to the PIN-G module, the neural activity of which allows to decode the syllable identity. The brain, in contrast, employs many feedback loops. In particular, attention to one of several acoustic streams as well as linguistic predictions likely act as top-down effects on speech coding (Etard et al. 2019b; Golumbic et al. 2013; O'Sullivan et al. 2015; Weissbart et al. 2020). Incorporating such higher-level cognitive processes as feedback mechanisms in the model will likely influence the neural network's capability for speech encoding. Importantly, incorporation of these mechanisms in the model will allow us to simulate how they may be influenced by tACS and determine their contribution to the neural processing underlying speech-in-noise comprehension.

Because our model is based on the hypothesis of speech encoding through coupled neural oscillations (Giraud et al. 2012; Hyafil et al. 2015), it might be used in the future to generate further predictions of how speech processing can be impacted by neurostimulation. Experimental verification or falsification of such predictions may allow to further establish the neural mechanisms of speech processing, and in particular to further investigate the role of coupled cortical oscillations. Moreover, our modelling framework may also be adapted to assess the effects of neurostimulation on the neural processing of other sounds, such as on music perception that may involve coupled oscillations as well.

Further developments of the model may also integrate it with structural modelling that seeks to estimate the current intensity in different brain regions for a certain placement of the electrodes and applied current (Huang et al. 2019a; Thielscher et al. 2015). In particular, such modelling may be based on subject-specific data like MRI images that might allow to obtain subject-specific outcomes, for instance regarding current intensities and neural delays. Integrating structural and functional modelling might therefore facilitate the understanding of inter-subject variations as well as allow to optimize stimulation parameters for an individual subject.

Chapter 4

Decoding of selective attention to continuous speech from the human auditory brainstem response

The work presented in this chapter has been previously published as Etard et al. (2019a). Implementations of the complex forward and backward models discussed in this chapter are openly available at `https://github.com/MKegler/cTRF_toolbox`. I would like to express my gratitude to Dr Octave Etard, who contributed to this work by sharing his EEG dataset collected in Etard et al. 2019b.

4.1 Introduction

Humans have an extraordinary capability to analyse crowded auditory scenes. We can, for instance, focus our attention on one of two competing speakers and understand her or him despite the distractor voice (Middlebrooks et al. 2017). People with hearing impairment such as sensorineural hearing loss, however, face major difficulty with understanding speech in noisy environments, and this difficulty persists even when they wear auditory prosthesis such as hearing aids or cochlear implants (Armstrong et al. 1997). Auditory prosthesis could potentially aid with understanding speech in noise through selectively enhancing a target speech, for instance based on its direction, using algorithms such as beam forming (Kidd Jr et al. 2015). However, such selective enhancement requires knowledge of which sound the user aims to attend to. Current research therefore attempts to decode an individual's focus of selective attention to sound from non-invasive brain recordings (Biesmans et al. 2016; Fuglsang et al. 2017; Mirkovic et al. 2015; O'Sullivan et al. 2015). If such decoding worked in real time, it could inform the sound processing in an auditory prosthesis. It could also form the basis of a non-invasive brain-computer interface for motor-impaired patients with brain injury, for instance, who may not be able to respond behaviourally. Moreover, such decoding of selective attention could be employed clinically for a better understanding and characterization of hearing loss.

Neural activity in the cerebral cortex, especially in the delta (1–4 Hz) and theta (4–8 Hz) frequency bands, tracks the amplitude envelope of a complex auditory stimulus such as speech (Ding et al. 2012, 2014; Giraud et al. 2012; Power et al. 2012). The tracking is shaped by selective attention to one of several sound sources and can be measured from electrocorticography

(ECoG) (Mesgarani et al. 2012), and noninvasively from magnetoencephalograpy (MEG) (Ding et al. 2012), as well as from the clinically more applicable electroencephalography (EEG) (Horton et al. 2013; Kerlin et al. 2010). Attention to one of two competing voices has been successfully decoded from single trials of 1 min in duration using MEG (Ding et al. 2012) as well as EEG (Fiedler et al. 2017; Mirkovic et al. 2015; O'Sullivan et al. 2015). Further optimization of the involved statistical modelling led to an accurate decoding of the focus of selective attention from still shorter recordings lasting less than 30 s (Biesmans et al. 2016; Van Eyndhoven et al. 2016). Moreover, a subject's changing focus of attention could be detected within tens of seconds from EEG data, and even faster from MEG data, when combined with additional sparse statistical modeling (Miran et al. 2018).

Recently we showed that subcortical neural activity is consistently modulated by selective attention as well (Forte et al. 2017). To this end we developed a method to measure the response of the auditory brainstem to natural non-repetitive speech. We employed empirical mode decomposition (EMD) to extract a waveform from the speech signal that, at each time instance, oscillates at the fundamental frequency of the voice. We then correlated this fundamental waveform to the neural recording obtained from a few scalp electrodes. We observed a peak in the cross-correlation at a latency of 9 ms, evidencing a neural response at the fundamental frequency with a subcortical origin. This method determined the brainstem response to the voiced parts of speech, and in particular to its pitch. When volunteers listened to two competing speakers, we observed that the brainstem response to the fundamental frequency of each speaker was larger when the speaker was attended than when she or he was ignored.

Because the brainstem response to speech that we measured occurs at the fundamental frequency of speech, typically between 100 and 300 Hz, it is ten-to hundredfold faster than the cortical tracking of the speech envelope. The rapidness of the brainstem response could imply a high information rate, despite the small magnitude of the response that is below that of cortical responses. We therefore wondered if the brainstem response to natural speech can be detected from high density EEG, that is typically used to capture the cortical activity, and whether it can be used to efficiently decode auditory attention.

4.2 Methods

4.2.1 Participants

18 healthy adult English native speakers (aged 22.8 $\pm$ 1.9 year, four females), with no history of auditory or neurological impairments participated in the study. All participants provided written informed consent. The experimental procedures were approved by the Imperial College Research Ethics Committee.

4.2.2 Experimental design and statistical analysis

We employed the same experimental design that we used previously to measure the brainstem response to non–repetitive speech and its modulation through selective attention (Forte et al. 2017). In particular, approximately 10-min long continuous speech samples from a male and female speaker were obtained from publicly available audiobooks (`librivox.org`). For the female voice excerpts from "*The Children of Odin*" (chapters 2 and 4) and "*The Adventures of Odysseus and the Tale of Troy*" (part 2, chapter 8), all by Pádraic Colum and read by Elizabeth Klett, were selected. For the male voice excerpts from "*Tales of Troy: Ulysses the Sacker of Cities*" by Andrew Lang (section 11) and "*The Green Forest Fairy Book*" by Loretta Ellen Brady (chapter 10), all read by James K. White, were used. The first story from the female speaker was employed when presenting speech in quiet. The two other female speech samples were used to generate two stimuli with two competing speakers by mixing each with one sample from the male speaker, at equal root-mean-square amplitude.

Participants first listened to the stimulus with a single speaker without background noise. They then listened to the two stimuli with two competing speakers each. They were instructed to exclusively attend either the male or female voice in the first stimulus, and to attend to the speaker they previously ignored in the second one. Whether a subject was instructed to first attend the male speaker and then the female speaker or vice versa was determined randomly for each subject. Each stimulus was presented in four parts of approximately equal duration ($\sim$ 2.5 min), and comprehension questions were asked after each part. All stimuli were delivered diotically, that is, the same waveforms were delivered to the right and left ears, at 76 dB(A) SPL (A-weighted frequency response) using Etymotic ER-3C insert tube earphones to minimise artifacts. The sound intensity was calibrated with an ear simulator (Type 4157, Brüel & Kjaer, Denmark). EEG recordings were obtained during the stimuli presentation and their statistical analysis was performed using custom Matlab and Python code and functions from the MNE toolbox (Gramfort et al. 2013, 2014) as described below.

4.2.3 Neural data acquisition and processing

Neural activity was recorded at 1 kHz through a 64-channel scalp EEG system using active electrodes (actiCAP, BrainProducts, Germany) and a multi-channel EEG amplifier (actiCHamp, BrainProducts, Germany). The electrodes were positioned according to the standard 10-20 system and referenced to the right earlobe. The EEG recordings were band-pass filtered offline between 100 and 300 Hz (low pass: linear phase FIR filter, cutoff (-6 dB) 325 Hz, transition bandwidth 50 Hz, order 66; high pass: linear phase FIR filter, cutoff (-6 dB) 95 Hz, transition bandwidth 10 Hz, order 364; both: one-pass forward and compensated for delay) and then referenced to the average. When only using three channels for the decoding, all channels except the two mastoids TP9 and TP10 and the vertex Cz were discarded, and the filters described above were applied. The audio signals were simultaneously recorded by the amplifier at a sampling rate of 1 kHz through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts, Germany), and were used to align the neural responses to the stimuli. A 1 ms

delay of the acoustic signal introduced by the earphones was taken into account by shifting the audio signal forward by 1 ms with respect to the neural response.

4.2.4 Computation of the fundamental waveform of speech

We employed Empirical Mode Decomposition (EMD) to extract a waveform from each speech signal that, at each time instance, oscillates at the fundamental frequency of the voice; we refer to it as the fundamental waveform (Forte et al. 2017). EMD is indeed well suited to analyze data that results from non-stationary and nonlinear processes such as speech production, and has been successfully used for pitch detection (Huang et al. 2006). The fundamental waveform was downsampled to 1 kHz, the sampling rate of the neural recordings, and filtered between 100 and 300 Hz as described above. Silent or unvoiced parts of the speech produced some segments where the fundamental waveform was equal to zero. For the stimuli with a single speaker, we excluded such segments from the further analysis. For the stimuli with two competing speakers we excluded the few segments where the fundamental waveform of one of the two voices was entirely zero as attention could not be decoded in this case.

We also computed a proxy of the fundamental waveform by band-pass filtering the audio signal in the range of the fundamental frequency. We thereby employed FIR filters with corner frequencies of 100 Hz and 200 Hz for the male voice (linear-phase FIR filter, lower cutoff (-6 dB): 90 Hz, transition bandwidth 17.5 Hz, higher cutoff (-6 dB): 210 Hz, transition bandwidth 17.5 Hz, order 237, one pass forward and compensated for delay), as well as corner frequencies of 150 Hz and 250 Hz for the female voice (linear-phase FIR filter, lower cutoff (-6 dB): 135 Hz, transition bandwidth 25 Hz; higher cutoff (-6 dB): 275 Hz, transition bandwidth 25 Hz, order 157, one pass forward and compensated for delay). We employed the band-pass filtered audio signals to obtain the results on attention reported in Fig. 4.7-B. All other results presented here were obtained from waveforms extracted by EMD.

4.2.5 Backward model

We first used a linear spatio-temporal backward model to reconstruct the fundamental waveform of speech from the neural recordings. Specifically, at each time instance t_n, the fundamental waveform $y(t_n)$ was estimated as a linear combination of the neural recordings $x_j(t_n + \tau_k)$ as well as their Hilbert transform $x_j^h(t_n + \tau_k)$ at a delay τ_k:

$$\hat{y}(t_n) = \sum_{j=1}^{N} \sum_{k=1}^{T} [\beta_{j,k}^{(r)} x_j(t_n + \tau_k) + \beta_{j,k}^{(i)} x_j^h(t_n + \tau_k)] \tag{4.1}$$

The index j refers hereby to the recording channel, and $\beta_{j,k}^{(r)}$, $\beta_{j,k}^{(i)}$ are a set of real coefficients to determine. We used a set of $T = 25$ possible delays ranging from -5 ms to 19 ms with an increment of 1 ms. The Hilbert transform of each recording channel was included in Equation 4.1, denoted with the upper index h, to allow the reconstruction of the fundamental waveform from

these signals as well. The Hilbert transform of a sinusoid results in a phase shift of $\frac{\pi}{2}$, which equates to a temporal shift of a quarter period. Even narrow-band signals such as our band-pass filtered EEG recordings contain, however, a range of frequencies. While the Hilbert transform of these signals can still be interpreted as a phase shift of $\frac{\pi}{2}$, it can no longer be obtained by a temporal shift. The Hilbert transforms therefore add another set of predictors in Equation 4.1 that are independent of the time-shifted EEG signals, and that thereby aid the reconstruction of the fundamental waveform.

The model's coefficients can be assembled into complex coefficients $\beta_{j,k} = \beta_{j,k}^{(r)} + i\beta_{j,k}^{(i)}$ that encode accordingly the amplitude of the brainstem response, the temporal delay as well as the phase difference between stimulus and response. We thus obtained $T = 25$ temporal delays that, together with the $N = 64$ recording channels, led to 1,600 complex model coefficients.

The model coefficients were then estimated for each subject using a regularised ridge regression as $\beta = (X^tX + \lambda I)^{-1}X^ty$, in which X is the design matrix of dimension $n_p \times 2NT$ with the number of samples available in the recording, and λ is a regularisation parameter (Friedman et al. 2001). In particular, the columns of the design matrix are the neural recordings $x_j(t_n+\tau_k)$ at the different time points t_n as well as their Hilbert transforms $x_j^h(t_n + \tau_k)$. To normalise for differences between datasets, λ can be written as $\lambda = \lambda_n e_m$, where e_m is the mean eigenvalue of X^tX and λ_n is a normalised regularisation coefficient (Biesmans et al. 2016).

A five-fold cross-validation procedure was implemented to evaluate the model. In each of five iterations, and for each participant, four folds of the 10-min data were used to compute the model coefficients, yielding about 8 min of training data. The remaining fifth fold, 2 min of testing data, served to estimate the fundamental waveform and to compute the performance of the model. The performance was quantified by dividing the reconstructed ($\hat{y} = X\beta$) and the actual (y) fundamental waveforms obtained on the testing data in 10-s long segments and computing Pearson's correlation coefficient between these waveforms for each segment. The correlation values obtained over the five testing folds were pooled to determine the mean and standard error of the reconstruction performance. This performance was determined for 50 different normalised regularization parameters with values ranging from 10^{-15} to 10^{15}. For each subject, the regularization parameter that yielded the largest reconstruction performance was chosen as the optimal regularization parameter.

The procedure above, including the use of the Hilbert transform of the EEG data, was employed both when reconstructing the fundamental waveform obtained from EMD as well as when estimating the fundamental waveform obtained from band-pass filtering the speech signal (see below).

4.2.6 Significance of the fundamental waveform reconstruction

To determine if the linear backward models showed a significant brainstem response to the fundamental frequency, we also computed, for each subject, one noise model as a linear backward model that attempted to reconstruct the fundamental waveform of an unrelated speech segment from the same female speaker. The noise models were computed using the same methodology we employed for determining the actual brainstem response, including the same cross-validation procedure and the same determination of the optimal regularization parameter per subject.

We then assessed whether the correct linear backward model outperformed the noise model, or the opposite, by comparing the correlation coefficients obtained on the 10-s segments through a two-tailed Wilcoxon signed rank test. The results of the statistical tests are indicated for each subject in Fig. 4.1-A through asterisks: no asterisk is given when results are not significant ($p > 0.05$), one asterisk when results are significant (*, $0.01 < p \leq 0.05$), two asterisks when significance is high (**, $0.001 < p \leq 0.01$), and three asterisks when significance is very high (***, $p \leq 0.001$).

4.2.7 Estimation of the neural response (forward model)

To gain further information about the neural origin of the response we also computed a linear forward model that estimated the EEG responses from the fundamental waveform. The coefficients of the forward model, as opposed to those of a backward model, allow for a neurobiological interpretation of their spatio-temporal characteristics (Haufe et al. 2014). The forward model relates the EEG recording $x_j(t_n)$ at time t_n to the fundamental waveform $y(t_n - \tau_k)$ as well as its Hilbert transform $y^h(t_n - \tau_k)$ at a delay τ_k:

$$x_j(t_n) = \sum_{k=1}^{T} [\alpha_k^{(r)} y(t_n - \tau_k) + \alpha_k^{(i)} y^h(t_n - \tau_k)], \tag{4.2}$$

in which $\alpha_k^{(r)}$ and $\alpha_k^{(i)}$ are the model's real coefficients. They can be interpreted as real and imaginary parts of the complex coefficients $\alpha_k = \alpha_k^{(r)} + i\alpha_k^{(i)}$. To investigate the temporal dynamics of the neural response, we considered a broader range of time lags than for the backward model. Specifically, we employed a set of T = 201 possible delays τ_k that ranged from -50 ms up to 150 ms with an increment of 1 ms. Although we did not expect a neural response at negative delays or at delays larger than 20 ms, we included those nevertheless to verify the absence of a significant response there. The model coefficients were estimated by concatenating the data from all subjects that showed a significant brainstem response to the speech signal as assessed earlier (generic or subject-averaged model) and using a regularised ridge regression as previously described.

4.2.8 Significance of the auditory brainstem response

We sought to investigate at which latencies significant neural responses emerged. We therefore compared the obtained forward model to noise models. One thousand forward noise models were computed analogously to the forward model, except that the fundamental waveform of the actual speech signal was replaced with a fundamental waveform of an unrelated speech stimulus, from the same female speaker. We constructed these unrelated speech stimuli by randomly picking four parts, each with a duration of 2.5-min, from the eight parts that constituted the female speech material used in the competing speaker condition. This procedure was repeated to create 1,000 surrogate waveforms (out of all 1,680 possible combinations). We then employed a mass-univariate analysis to identify the significant time delays (Groppe et al. 2011). In particular, we computed the average magnitude of the responses over the EEG channels, yielding a single real time-varying function for the actual neural response and of the noise responses. We then pooled the values from the 1000 noise responses over the time lags to establish a single empirical null-distribution. We used this distribution to determine a critical value corresponding to a p-value of 0.05 to which the actual neural response was compared at each time lag from -50 ms to 150 ms (Bonferroni correction for multiple comparison).

In addition, we analysed the topography of the forward model at the peak latency τ_0 of the average magnitude of the responses over the EEG channels. To this end, the forward noise models were used to build an empirical null distribution for each channel. For each noise model, the peak latency of the average magnitude was determined, and the magnitude of each channel's response at this latency was used to establish the null distribution of that channel. Finally, the forward model at time τ_0 was compared to the corresponding null empirical distribution at the respective channel at a significance level of $p = 0.05$, with FDR correction for multiple comparison over channels.

4.2.9 Stimulus artifacts

We also computed the cross-correlation between the EEG responses to speech in quiet and the corresponding broad-band speech signal, with the purpose of checking for stimulus artifacts. To this end the speech stimulus was resampled from 44,100 Hz to 1,000 Hz, the sampling frequency of the EEG data. The cross-correlation functions were then analysed for statistically significant peaks at delays between -200 ms and 200 ms following the same procedure as described above for the forward model. Briefly, the cross-correlations were first averaged over subjects, and the absolute value of the resulting functions were then averaged over electrodes, yielding the average neural response as a function of latency. To establish a chance level, the same calculations were reproduced when replacing the speech stimulus by a different one from the same speaker. This procedure was repeated 1,000 times, yielding 1,000 noise responses. These stimuli were constructed as described above. These noise responses were pooled over time lags to build a single null distribution that was then used to assess the significance of the actual averaged neural responses as described above for the forward model ($p = 0.05$, Bonferroni-corrected for multiple comparison over time lags between -200 ms and 200 ms).

4.2.10 Attentional modulation of the auditory brainstem response

To analyse the attentional modulation of the brainstem response to one of two competing speakers, we computed two pairs of backward models for each subject. The first pair of models reconstructed the fundamental frequency of the male voice while it was either attended (MA model) or ignored (MI model). The second pair of models reconstructed the fundamental waveform of the female voice when the subject attended it (FA model) or when the subject ignored it (FI model). The computation of the backward models, and the assessment of their performance, was done through five-fold cross-validation as explained above.

For each speaker, the performances of the attended and ignored models were then compared using a two-tailed Whitney-Mann rank test at the subject level. The results are indicated in Fig. 4.4 through asterisks as described above. We further employed a two-tailed Wilcoxon signed rank test to investigate whether the population-average ratios of the performances were, for each speaker, significantly different from unity. Finally, we used a two-tailed Wilcoxon signed rank test to check if the population-average ratios obtained from the responses to the male voice and to the female voice were significantly different.

4.2.11 Differences between brainstem responses to attended and to ignored speech

We sought to determine whether the difference in the brainstem response to attended and to ignored speech reflected merely a difference in the strength of the response, or if there were other changes as well. To this end, we compared the magnitudes and the phases of the complex coefficients of the forward model for an attended voice to those for an ignored voice. Because the forward models for the male and for the female voice reflected the different fundamental frequencies of both speakers, we performed this analysis separately for the male and for the female voice. Regarding the magnitude, we computed the ratio of the amplitude of the attended and of the unattended model, at the peak delay of their average amplitude (9 ms, for both the male and female voices). We then employed a two-tailed Wilcoxon signed rank test to determine for which electrodes the ratio was significantly different from unity ($p < 0.05$, FDR-corrected for multiple comparison over electrodes). To compare the phase, we computed the phase difference between the attended and the ignored model at each electrode at this same peak latency. We considered the wrapped phase differences that were mapped to the range of $(-\pi, \pi)$. We then determined the statistical significance of the phase difference through the Rayleigh test for non-uniformity of circular data ($p < 0.05$, FDR-corrected for multiple comparison over electrodes). The Rayleigh test assesses the null hypothesis that the phase differences are uniformly distributed around the circle. However, it does not inform on the value of the phase differences. Therefore, we derived 95% confidence intervals for the mean phase difference by pooling the data across all electrodes that exhibited significant phase clustering. All circular statistics were performed using the Circular Statistics Toolbox for Matlab (Berens 2009). Finally, we compared the latency of peak

amplitude between the attended and ignored models using a Wilcoxon signed rank test.

In order to enable a direct comparison with our previous related work, we also computed the difference between the TRF at electrode CPz and the average TRF of the two mastoids to produce one dipolar response (Forte et al. 2017). CPz was selected due to its central location, similar to the one used in our previous study, and because it emerged in our present study as one of the central electrodes that displayed a significant response to speech in quiet (Fig. 4.1-C). We then computed the ratio of this dipolar response between the attended and the ignored condition.

4.2.12 Decoding of auditory attention

We investigated how attention could be decoded from short segments of neural data that were obtained in response to competing speakers. We first trained and assessed the performances of the two pairs of speaker-specific linear backward models (MA, MI, FA, FI, as described above) using five-fold cross-validation. For all the attention decoding procedures presented hereafter, the normalised regularisation coefficient of the backward models was fixed to the value that yielded the best reconstruction for speech in quiet, $\lambda_n = 10^{-0.5}$.

The testing fold was divided into testing segments with a duration of 0.5, 1, 2, 4, 8, 16 and 32 s. For each testing segment we therefore obtained four different correlation coefficients: the correlation coefficient r_{MA} between the fundamental waveform of the male speaker and its reconstruction based on the MA model, the correlation coefficient r_{MI} between the fundamental waveform of the male speaker and its reconstruction based on the MI model, as well as the correlation coefficients r_{FA} and r_{FI} between the fundamental waveform of the female speaker and its reconstruction based on the FA and FI model, respectively. The computed correlation coefficients were then employed to decode attention on each segment. We thereby explored two different avenues (Fig. 4.6-A).

First, we based the decoding on the attended models MA and FA only. To this end, we compared the correlation coefficients from both models. If r_{MA} exceeded r_{FA} we concluded that the male speaker was attended, and otherwise that the female speaker was the focus of attention. Second, we considered the ignored models MI and FI only. If r_{MI} was larger than r_{FI} attention was decoded as having been directed at the female speaker, and *vice versa* if r_{MI} was smaller than r_{MI}.

The decoding of attention using these two different methods was performed using all 64 EEG channels as well as based on three EEG channels only (vertex and mastoids: Cz, TP9, TP10). The decoding of attention based on the attended models was also performed using the fundamental waveform obtained by band-pass filtering.

We sought to compare the performance of the obtained attentional decoding to that of a random classifier. A random binary classifier can achieve a high accuracy by chance. This is

especially true when the number of testing data is small, which in our case occurs when the duration of the testing segments is long. To account for this effect, we determined the 95% chance level, that is, the highest accuracy that a random classifier would achieve in at least 95% of cases. This 95% chance level was computed using a binomial distribution (Combrisson et al. 2015).

4.2.13 Subject-independent attention decoding

In real-life situations, the decoding of auditory attention may be required for a subject for whom training data is not available. This situation requires to train a decoder on other people for whom training data is at hand, and to then apply it to the subject under consideration. We refer to such decoders as out-of-the-box models since, once trained on the data from a set of volunteers, they can be readily applied to other subjects. To assess how well these out-of-the-box models decode auditory attention, we trained linear backward models on the pooled data from all subjects and quantified their performances using a leave-one-subject-out cross-validation coupled with a five-fold cross-validation regarding the auditory stimuli (i.e. testing on data from a subject and from a part of the stimulus unused during training). To train the model, the testing data from all-but-one participants was concatenated and used to obtain the model coefficients. The unseen part of the data from the remaining subject was used to assess the performance of the model. In particular, we assessed the classifier that compared the performances of the MA and the FA model. Its classification accuracy was evaluated as described above.

4.2.14 Speaker-averaged attention decoding

We also wondered how well selective attention could be decoded from the brainstem response if the specific models of the brainstem responses to the individual voices were not available. We therefore followed a similar analysis as used for decoding auditory attention based on the speech envelope (O'Sullivan et al. 2015). For each subject, we computed a single backward model for an attended voice, irrespective if it was the male or the female one. This model was accordingly trained on the data from both the condition when the subject attended the male voice and the condition when they listened to the female speaker. The male fundamental waveform was used as the reconstruction target when the male speaker was attended, and the female fundamental waveform was the target when the female voice was attended. An equal proportion of data from each attention condition was included in each cross-validation fold. To determine the focus of attention, we then considered short testing segments as described above. For each testing segment we computed the correlation coefficient between the reconstructed fundamental waveform and the actual ones of the two speakers. If the reconstruction matched the fundamental waveform of the male speaker more closely than that of the female one, we concluded that the subject had attended the male speaker. Otherwise we determined that the focus of attention was on the female voice. The performance of the classifier was evaluated as described above.

4.3 Results

4.3.1 Response to a single speaker

We first measured neural responses to a single non-repetitive speech signal from 64-channel EEG. We employed empirical mode decomposition to obtain a fundamental waveform from the speech signal (Forte et al. 2017), and linear regression with regularization to reconstruct the fundamental waveform from the multi-channel EEG data for each individual subject (linear backward model, Methods). The performance of the reconstruction was assessed through the mean Pearson's correlation coefficients over 10-s segments of the reconstructed fundamental waveform to the actual one (Fig. 4.1-A).

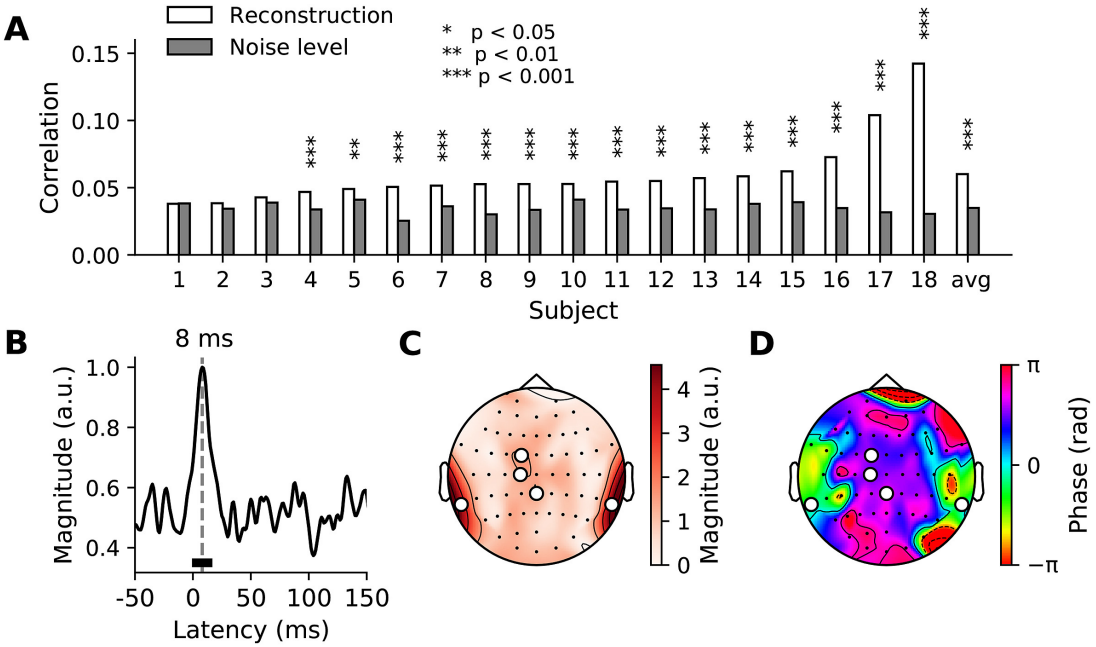

Figure 4.1: The brainstem response to natural speech detected from high-density EEG recordings using complex linear models. (**A**) The performance of the linear backward model is quantified through the Pearson's correlation coefficient of the reconstructed fundamental waveform and the actual one. For each subject the presented result is the averaged correlation coefficient obtained from 10-s long segments of the EEG and the fundamental waveform (white bars). In almost all subjects, the performance is significantly better than that of a model estimating the noise-level reconstructions. Subjects have been ordered by increasing performances. (**B**) The channel-averaged magnitude of the complex coefficients of the generic forward model obtained from the pooled data from all the participants that yielded significant reconstructions, peaks at a latency of 8 ms. Only latencies ranging from 3 to 14 ms yield a statistically-significant response (black bar, $p < 0.05$, Bonferroni correction), as compared to noise models. (**C**) At the delay of 8 ms, a significant neural response emerges from the mastoid channels as well as from the channels near the midline (white disks, $p < 0.05$, FDR correction, population average). (**D**) The phase of the complex coefficients at the delay of 8 ms shows a phase difference of around π between the temporal areas and the central one (population average).

We verified that the linear backward models did extract a significant brainstem response to

speech. To this end we also constructed models based on the fundamental waveform of unrelated speech signals from the neural data. For almost all subjects that we assessed (15 out of 18), the model that reconstructed the actual fundamental waveform significantly outperformed the one that attempted to reconstruct an unrelated fundamental waveform, showing that the former was able to extract a meaningful brainstem response (Fig. 4.1-A, two-tailed Wilcoxon signed rank test).

To investigate the spatio-temporal characteristics of the brainstem response we computed a generic linear forward model that estimated the EEG recordings from the fundamental waveform using the data from all the subjects that yielded significant reconstructions in the previous test presented in Fig. 4.1-A (Methods). The average over channels of the magnitude of the obtained complex coefficients peaked at 8 ms, and only the latencies around this peak (3–14 ms) yielded statistically-significant neural responses (Fig. 1-B). This finding demonstrated the subcortical origin of the neural activity and was in agreement with previous recordings of speech-evoked brainstem responses (Forte et al. 2017; Maddox et al. 2018; Reichenbach et al. 2016; Skoe et al. 2010). The magnitude of the coefficients at that latency showed major contributions from the mastoids as well as moderate contributions from the central scalp areas (Fig. 4.1-C). Both the mastoid channels as well as the channels near the midline of the scalp yielded significant responses. The coefficients at the central area were approximately in antiphase to those near the mastoids, reflecting the direction of the brainstem's dipole sources (Fig. 4.1-D).

We also computed linear forward models for single subjects (Fig. 4.2). We find that they yielded peak responses at similar latencies, and showed similar topographies, although these were noisier than the ones obtained from the average over all subjects.

4.3.2 Absence of stimulation artifacts

To determine if stimulus artifacts were present in the recordings, we computed a cross-correlation between the EEG data and the broadband speech signal. Broadband speech elicits neural responses from the brainstem to the cortex, at latencies ranging from 5 ms to a few hundred ms (Maddox et al. 2018). A stimulus artifact would arise, in contrast, instantaneously, at a delay of -1 ms. This delay reflects the fact that, in our analysis, we compensated for the earphone's 1 ms delay of delivering the sound to the ears. The responses that we recorded contained, however, only significant contributions between 9 and 12 ms delays, firmly in the range of subcortical neural activity (Fig. 4.3). We could accordingly not detect stimulus artifacts in our EEG recordings.

4.3.3 Attentional modulation of the response to competing speakers

We then investigated how attention modulates the brainstem response. Following a classic auditory attention paradigm we presented subjects with a male and a female voice diotically and simultaneously, instructing them to attend to either the male or the female speaker, while recording their neural activity from 64-channel EEG (Ding et al. 2012; Forte et al. 2017). For

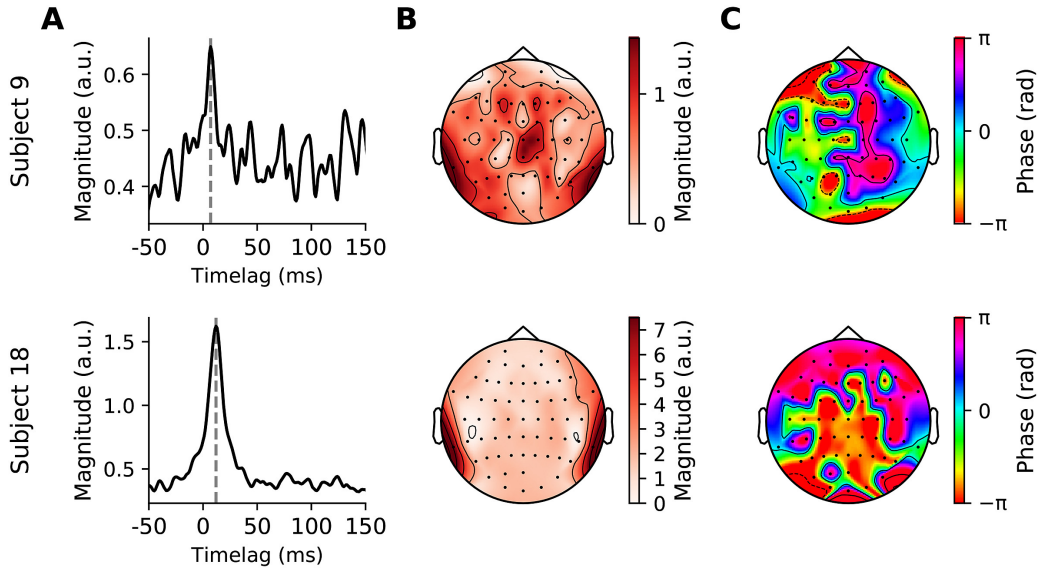

Figure 4.2: Brainstem responses to speech from two single subjects. The top row shows the brainstem response from subject 9 that yielded the median fundamental waveform reconstruction performance (Fig. 4.1). The bottom row presents the results from subject 18 that had the best reconstruction of the brainstem response to speech. (**A**) The channel-averaged magnitude of the complex coefficients of the forward model peaks at a latency of 9 ms (subject 9) and 10 ms (subject 18). (**B**) The topographic maps of the coefficient magnitudes at the peak latency are consistent with those of the generic model, although more noisy in the case of subject 9. Channels located at the mastoids show the highest magnitudes. (**C**) The phase of the complex coefficients at the peak latency. The phases differ between the two subjects since they have been taken at different latencies (9 and 10 ms, respectively). Consistent with the generic model, the topographic plots show a phase difference of around π between the temporal areas and the central area.

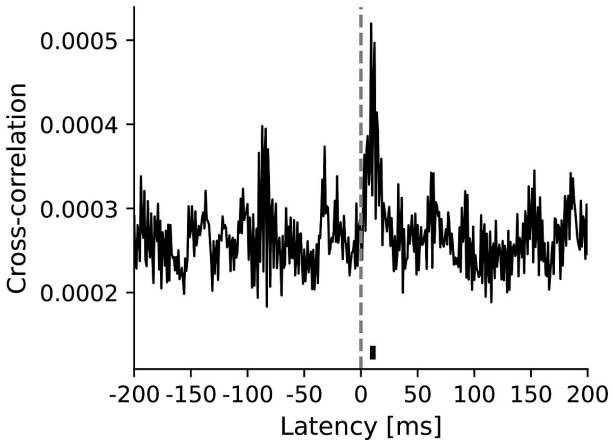

Figure 4.3: Absence of stimulus artifacts. Magnitude of the cross-correlation between the EEG data and the broadband speech stimulus averaged over channels and participants. The only time lags for which the cross correlation is significantly greater than the estimated noise floor are between 9 and 12 ms. In particular, the model shows no significant response at the delay of -1 ms, the delay of the earphones, evidencing the absence of stimulus artifacts.

each subject, we computed four linear backward models. The first model, MA, reconstructed the fundamental waveform of the male voice when the subject attended to it. The second model, MI, reconstructed the fundamental waveform of the male speaker when the subject ignored it. Analogously, a third and fourth model, FA and FI, reconstructed the fundamental waveform of the female voice when it was attended or ignored, respectively. We observed that the performance of the two models that reconstructed the fundamental waveform of a speaker when they were attended was, in most subjects, significantly better than that of the corresponding model for the ignored voice (Fig. 4.4, two-tailed Whitney-Mann rank test). The average ratio between the reconstruction performance of the model for the attended male voice to that for the ignored male voice was 1.22, significantly larger than one ($Z(17) = 7$, $p < 0.001$, two-tailed Wilcoxon signed rank test). The ratio was 1.15 in the case of the female voice, which was significantly above one as well ($Z(17) = 38$, $p = 0.039$, two-tailed Wilcoxon signed rank test). The two ratios did not differ significantly ($Z(17) = 69$, $p = 0.47$, two-tailed Wilcoxon signed rank test). The better reconstruction performance of the fundamental waveform of an attended speech signal demonstrates the attentional modulation of the brainstem response to speech that we described previously (Forte et al. 2017).

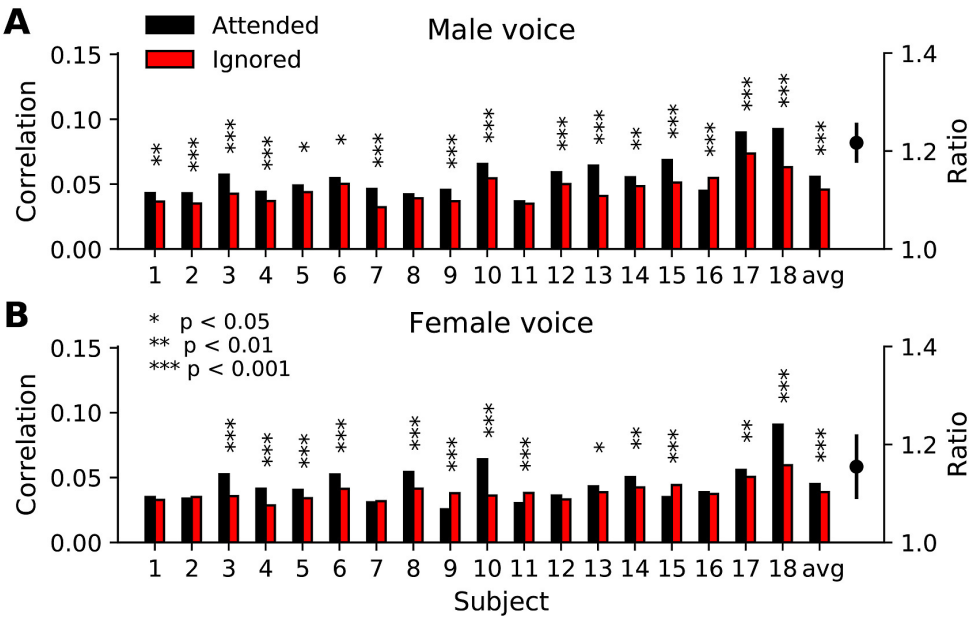

Figure 4.4: Attentional modulation of the auditory brainstem response to natural speech. The order of the subjects is as in Fig. 4.1A. (**A**) The performance of the linear backward model for the male voice is better when the male speaker is attended (black) then when he is ignored (red). The two performances differ significantly in most subjects, and so do the two average performances (avg). The average ratio between the two performances is 1.22 and is significantly larger than one ($p = 0.01$). (**B**) The performance of the linear backward model that reconstructs the fundamental waveform of the attended female voice is likewise significantly better than that for the ignored female voice in most subjects, as well as on average (avg). The average ratio of the two performances is 1.15 and is significantly larger than one ($p = 0.039$). The ratios for the male and female voices do not differ significantly ($p = 0.47$).

We wondered if the difference between the attended and the ignored brainstem response

reflected merely a difference in the strength of the response, or if there were other differences as well. To investigate the nature of these differences, we compared the coefficients of the attended forward models to those of the ignored models, at the peak delay of their average amplitude (9 ms). We found that the ratio of the magnitude of the coefficients did not differ statistically from unity, neither for the male nor for the female voice (Fig. 4.4-A,C; Wilcoxon signed rank test, FDR correction for multiple comparison over electrodes). However, we found a statistically significant clustering of phase differences between the attended and the ignored models at several electrodes near the midline as well as near the mastoids (Fig. 4.5-B,D; Rayleigh test for non-uniformity of circular data, FDR correction for multiple comparison over electrodes). For the male voice, the mean phase difference was found to be -0.51π (95% confidence interval: $[-0.56\pi; -0.47\pi]$), while it was -0.12π for the female voice (95% confidence interval: $[-0.17\pi; -0.08\pi]$). This shows that the ignored models were not merely a scaled version of the attended models, but that the brainstem response to ignored speech occurred at a different phase from that to attended speech.

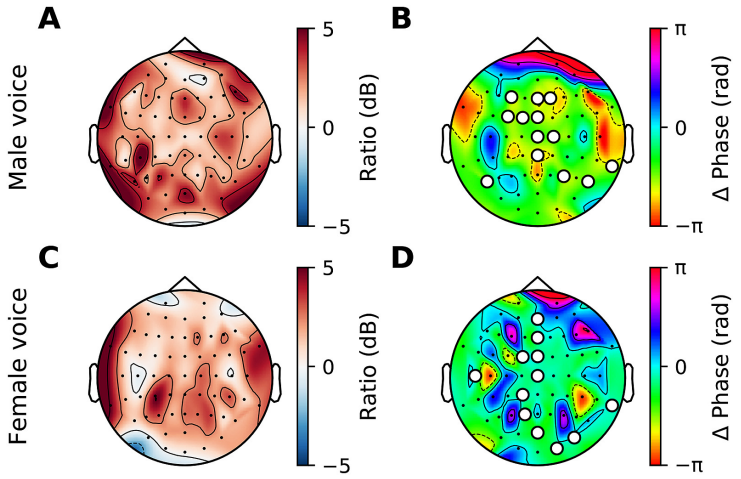

Figure 4.5: Differences in the brainstem response to attended and to ignored speech. (**A**, **C**) The subject-averaged ratio of the magnitude of the complex coefficients of the attended forward model to those of the ignored model, at the average peak latency of 9 ms. None of these ratios are statistically different from unity (FDR correction). (**B**, **D**) The subject-averaged phase difference between the coefficients of the attended and the ignored forward models, at the average peak latency of 9 ms. Channels close to the midline as well as at channels near the mastoids yielded a significant phase difference ($p < 0.05$, FDR correction). The male models exhibit a phase difference of -0.51π (95% CI: $[-0.56\pi; -0.47\pi]$), while the female model phase difference is -0.12π (95% CI: $[-0.17\pi; -0.08\pi]$).

Due to the range of frequencies that constitute the fundamental waveform, the phase shift between the attended and the ignored models did not equate to a consistent temporal shift. We did indeed not find a statistically-significant difference in the timing between the peak amplitude of the attended and the ignored models across the different subjects, for the male or female voice ($p = 0.17$ and $p = 0.69$ respectively, two-tailed Wilcoxon signed rank test).

To facilitate comparison with previous work we also computed the difference of the mastoid electrodes and the electrode at CPz, yielding a dipolar response (Forte et al. 2017). We found

that the response's ratio between the attended and ignored condition was significantly greater than unity, for both the male and female voices ($p = 0.016$, and $p = 0.003$ respectively, Wilcoxon signed rank test).

4.3.4 Decoding of auditory attention

Having verified the attentional modulation of the brainstem response to speech using high-density EEG recordings and linear backward models, we sought to investigate whether this approach could be used to decode auditory attention. We expected the focus of attention to emerge, for instance, from the difference in the performances of the models MA and FA. This difference should typically be positive when the subject attended to the male voice and be negative otherwise (Fig. 4.6-A). Similarly, attention could potentially be decoded from the difference of the reconstruction performance of the models FI and MI. A subject's attention to the male voice should mostly lead to a positive difference, and a focus on the female voice to a negative difference.

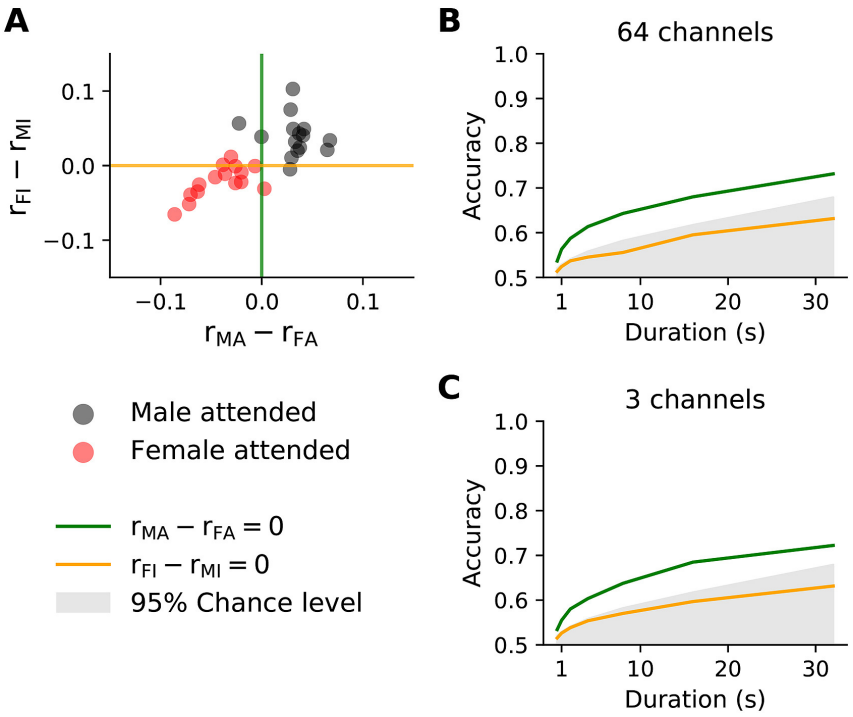

Figure 4.6: Decoding of auditory attention. (**A**) Testing data of a duration of 32 s that were obtained from a subject listening to the male speaker (black) can potentially be discriminated from those obtained when a subject listened to the female voice (red) through the performances r from four linear backward models (MA, MI, FA, FI; Methods). The classification can employ the difference in the performances between the models MA and FA (green) or the difference between the models FI and MI (orange). (**B**) The subject-averaged decoding accuracy obtained from the models MA and FA reaches 73% at a duration of 32 s and remains above chance level (grey) for very short durations of 500 ms. Decoding based on the models FI and MI remains below chance level (average over all subjects). (**C**) Employing only three recording channels to decode attention reduces the performance of the classifiers only slightly, if at all.

We tested the accuracy of the decoding on samples of a duration that varied from half a second to over 30 s (Fig. 4.6-B). The averaged decoding accuracy based on the attended models (MA, FA) remained significantly above chance even for very short samples that lasted only half a second. It was, for instance, 59% and 69% for 2-s and 16-s samples, respectively. In contrast, the models MI and FI by themselves did not allow for a decoding of the attentional focus with an accuracy that was better than chance. In the following we therefore discuss decoding obtained from the attended models only.

Practical applications of the decoding of auditory attention benefit from a small number of required recording channels. We therefore investigated how well the developed decoding works if the linear backward models use only three EEG channels, the left and right mastoid as well as the central channel Cz. Strikingly, the subject-averaged decoding accuracy was barely smaller than that of the 64-channel model; for instance, it remained at 69% for a 16-s sample when the classifier based on the attended models was used (Fig. 4.6-C).

Both for the 64-channel as well as for the 3-channel decoding we observed variation in the decoding accuracy from subject to subject (Fig. 4.7-A). For a duration of 16 s, for instance, some subjects showed decoding accuracy close to 90%, whereas other subjects exhibited significantly lower decoding accuracies that did not exceed the change level. However, even for short testing segments and for the majority of subjects, the decoding remained above chance level. We note in addition that the subjects that did not allow for significant decoding include those for whom we did not obtain significant brainstem responses to speech in quiet (Fig. 4.1-A).

Because of the complexity of empirical mode decomposition (EMD), the computation of the fundamental waveform through this method cannot typically be performed online. We therefore wondered if attention could be decoded based on a similar waveform obtained through band-pass filtering the audio signal in the range of the fundamental frequency. Band-pass filtering is indeed a comparatively simple operation that can run in real time. We found that decoding based on the band-pass filtered audio has a similar accuracy as the one based on the waveform obtained from EMD, which is encouraging for real-time applications (Fig. 4.7-B).

Real-world settings will often feature voices that have not been encountered before and for which no speaker-specific model of the brainstem response is available. In an attempt to generalise our results, we computed a speaker-averaged backward model for any attended speaker, irrespective of whether it was the male or the female one. We then decoded attention from the performance of this speaker-averaged model in reconstructing the fundamental waveform of either the male of the female speaker. The averaged decoding accuracies that we obtained were slightly lower than those from the speaker-specific models but were above chance level for durations down to 0.5 s (Fig. 4.7-C).

The decoding described above utilized linear backward models that were subject specific and hence required prior training from EEG recordings for each individual. Such subject-specific

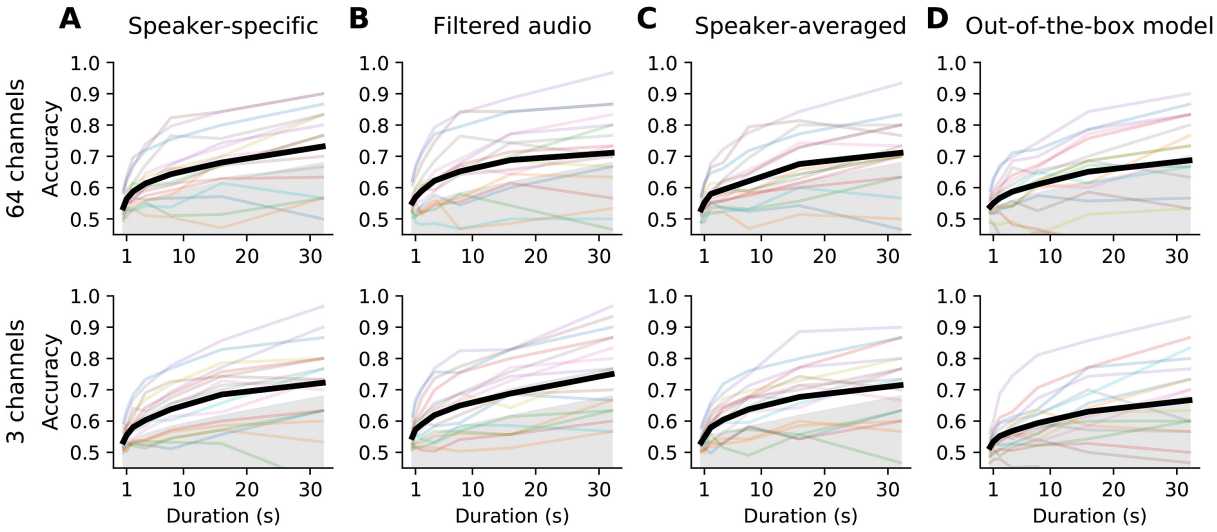

Figure 4.7: Different types of attention decoding and intra-subject variability. The two rows of panels correspond to the 64-channel and to the 3-channel decoders, respectively. (**A**) The attention decoding accuracies from the speaker-specific models achieved per individual subject (coloured lines, consistent across panels) varies by up to approximately 50% around the average (bold black line). However, for each individual subject the decoding based on 64 channels (top) is similar to that achieved from three channels (bottom). Here, the decoding is based on the difference between the attended models (same data as presented on the population level in Fig. 4.6-B,C by the green lines). (**B**) Instead of using empirical mode decomposition (EMD), a fundamental waveform can be estimated by band-pass filtering the speech signal, which can be implemented in an online fashion. Attention decoding based on the band-pass filtered audio achieves a similar performance as the one based on the EMD. (**C**) Attention can be efficiently decoded using a single attended model for both speakers as well. (**D**) The use of the out-of-the-box backward models for reconstructing the fundamental waveforms, leads to reduced, yet better than chance, decoding accuracies for most subjects.

training may, however, not always be available. We thus assessed the performance of a linear backward model that was trained on the whole population of subjects, and thus represented an average model that could be used out-of-the-box to decode attention. As expected, the decoding accuracies were then lower than those for the subject-specific models. While the decoders based on the attended models with all 64-channels remained above the chance level for all the tested durations, the 3-channel setup yielded worse performance only slightly exceeding the chance level for all but the longest duration. For duration of 16 s, for instance, the 64-channel setup yielded 65% accuracy, while the 3-channel only 63% (Fig. 4.7-D). Although the accuracy of this decoding when averaged across subjects was not very high, we note that this average was significantly reduced by a few subjects that showed particularly poor accuracies of around 50%, reflecting poor brainstem recordings from these subjects. The majority of the subjects, in contrast, yielded decoding accuracies that exceeded the chance level.

4.4 Discussion

We showed that the brainstem response to the fundamental frequency of speech can be measured reliably from high-density EEG recordings in most subjects through a statistical modelling approach. The response is most evident in the differences between the electrodes near the mastoids and those close to the vertex, in agreement with the dipolar structure of scalp-recorded auditory brainstem activity (Bidelman 2015; Grandori 1986; Norrix et al. 1996; Ono et al. 1984). Moreover, the response latency of 8 ms evidenced a subcortical origin.

The frequency-following response (FFR) to simpler acoustic signals such as long vowels has recently been found in an MEG study to contain cortical contributions (Coffey et al. 2016). However, when measured through EEG, the cortical contributions emerge earliest at a latency of 20 ms, are smaller than the subcortical ones, and mostly apparent for frequencies up to about 100 Hz (Bidelman 2018). The response to the fundamental frequency of running speech that we have measured here does not show a measurable signal at latencies longer than 14 ms and was recorded in response to a fundamental waveform high-pass filtered above 100 Hz. While contributions from cortical structures cannot be entirely ruled out, we did not observe any within our measurement accuracy.

When subjects switched attention from one to another of two competing speakers, we found that the fundamental frequency of each speaker was better encoded in the brainstem response when that speaker was attended rather than ignored. These results align with those that we obtained previously from different recording equipment and with a different analysis procedure that did not involve statistical modelling and that did not address attention decoding (Forte et al. 2017). Here we found, however, that the ratio of the attended to the ignored temporal response functions, as obtained from the forward models, did not differ significantly between the male and the female voice. Indeed, although the scalp maps that we derived largely showed a larger response to the attended than to the ignored speaker (Fig. 4.5-A, C), the modulation was not statistically significant. This presumably reflected the inclusion of all electrodes in the forward model, including many electrodes that displayed a poor signal-to-noise ratio. The backward models, in contrast, employed a weighting of the contribution from each electrode which boosted those with a large signal-to-noise ratio and thus led to a more significant result. To further investigate this issue, we also computed the response at a single channel that was obtained as the difference between the electrodes at the mastoids and at CPz, mimicking our previous bipolar recordings (Forte et al. 2017). The amplitude of this response was significantly modulated by selective attention, in agreement with our previous results.

The modelling work that we developed here allowed us to further investigate the origin of the difference in the brainstem response to attended and to ignored speech. We thereby found a significant difference between the phases of the response to attended versus ignored speech. Such a phase shift could in principle emerge from a difference in latency between the attended and ignored model. However, we found no statistically significant difference in peak latency of the attended and ignored responses. The phase shift might instead signify different relative

contributions of different parts of the brainstem to the scalp-recorded response. The different values of the phase shift that we obtained for the male and female voice may reflect the differences in the fundamental frequencies of both stimuli.

Most importantly, we developed a procedure to decode the attentional focus of a subject to speech based on her or his brainstem response as measured from as few as three recording channels. This will enable the future characterization and investigation of the subcortical mechanisms through which the brain solves the cocktail party problem. Potential practical applications include brain computer interfaces, such as neuro-steered auditory prostheses, as well as clinical assessments of supra-threshold hearing impairments that cannot be identified from pure-tone audiometry. Any of these applications will benefit from a decoding method that is fast and requires only a small number of recording channels.

We showed that the best decoding is achieved when linear models that relate the neural recording to the speech signal are computed for each subject individually. Such subject-specific models may cause difficulty in practice as sufficient training data per subject may not always be obtainable. The out-of-the-box models reflect the generalized version of the models obtained from the data pooled over many subjects and can be readily applied to other subjects for which no training data is available. We have shown that while the decoding performance of the out-of-the-box models is below those of the subject-specific models, the average decoding accuracy still exceeds the noise level for the high-density EEG setup. This suggests a consistency of the brainstem responses to speech across the participants. We also note that the out-of-the box models were fitted using the data from all subjects, including those that did not yield a significant reconstruction of the fundamental waveform in the speech-in-quiet condition.

Potential real-world applications will also often require the decoding of attention to a speaker that has not been encountered before. As an important step in this direction, we showed that speaker-averaged models that are trained on both attended speech signals, thereby computing an attended model that was averaged over the different voices, still performed well and allowed to decode attention. Future work could investigate how well these models generalise to speakers for which no training data is available.

Another important feature for real-time attention decoding is that the whole computational pipeline – from the processing of the audio signal to the computation of reconstructed waveforms and the attention decoding – can run online. Our reconstruction of the fundamental waveform through a backward model, the assessment of its performance as well as the subsequent attention decoding were all based on linear operations that can easily run in real time. However, the EMD that we employed for the computation of the fundamental waveform comes with large computational costs. We therefore explored how a computationally much simpler operation, band-pass filtering of the audio signal, performed regarding the decoding of attention. Promisingly we found that this method still allowed to decode attention from very short segments of data, evidencing the potential for real-time decoding. While two bandpass filters with different

corner frequencies were applied to the male and female voice, this approach could be extended to use filterbanks or use online pitch estimation algorithms.

The decoding procedure that we developed relies on the correlation between the reconstructed fundamental waveform from the brainstem response and the actual fundamental waveform of the speech signal. The obtained correlation coefficients are small, typically between 0.05 and 0.1 (Fig. 4.1-A, Fig. 4.2). Cortical responses allow to reconstruct the brainstem response from EEG recordings and yield somewhat higher correlation coefficients. However, the attentional decoding based on the brainstem responses that we show here is comparable to the decoding based on the reconstructed speech envelope, obtained from 64 EEG channels. A 16-s trial, for instance, yields an average decoding accuracy of about 69% when based on the fundamental waveform, which is similar to the corresponding decoding accuracy that was reported in several previous studies (Biesmans et al. 2016; Bleichner et al. 2016; O'Sullivan et al. 2015). We attribute this similarity of the attention decoding accuracies to the rapidness of the brainstem response: because the brainstem response to speech occurs at the fundamental frequency of a voice, it is ten-to hundredfold faster than the cortical response to the speech envelope. This rapidness appears to compensate for the smaller magnitude of the response.

Although brainstem responses and cortical responses allow for similarly efficient attention decoding when high-density EEG is available, the decoding based on the brainstem response to speech may have advantages when only a few channels are available. The accuracy of attention decoding based on the speech envelope drops indeed below 80% for a trial of at least 20 s when relying on subject-specific five-electrode montages (Fuglsang et al. 2017; Mirkovic et al. 2015). Similarly, the attention decoding based on the brainstem response that we have developed here achieves an averaged accuracy of 69% when based on three electrodes (TP9, TP10 and Cz) and on 16 s of data, and reaches 72% when 32 s of data are available (Fig. 4.5-B). This good decoding performance from a few EEG channels may be due to the effective capturing of the brainstem response by sparse montages, as well as due to a consistent dipole orientation across subjects (Dale et al. 1993). Importantly, we employed only band-pass filtering as a pre-processing step for the EEG data. The simplicity of this attention decoding method and its good accuracy when based on a few EEG channels may make this method attractive for practical applications.

The mixed-speaker stimuli that we employed were obtained by superimposing two speech signals, and our decoding was based on the knowledge of these separate voices. The individual components of a complex acoustic scene are, however, in general not available and need to be estimated from the acoustic mixture. The application of our method for decoding attention to steer an auditory prosthesis towards an attended voice, for instance, will thus require to first segregate the different voices that are present in the acoustic space, and to then determine the focus of the user's attention. The segregation of the different individual speakers may be achieved through multi-microphone arrays together with methods such as beamforming (Gannot et al. 2001) or non-negative blind source separation (Van Eyndhoven et al. 2016).

Certain applications may, however, not require the separation of the individual voices from an acoustic mixture but have them already available. Many locked-in patients, for instance, cannot communicate overtly, not even through eye motion (Giacino et al. 2002). Current brain-computer interfaces for them are mostly based on the P300 response, an evoked cortical potential that arises 300 ms after the occurrence of an oddball stimulus. It is typically elicited through visual or through sound stimuli and requires a few seconds to achieve a single binary response (Nijboer et al. 2008; Piccione et al. 2006; Schreuder et al. 2011). A brain-computer interface based on auditory attention, in contrast, could present a mixture of two auditory streams to the patient. The patient could then answer a question with yes or no through attending to a particular stream. Because the stimuli are merely used as a locus of attention, they would be available individually beforehand, and could be engineered to enhance decoding speed. Similarly, clinical assessments of the brainstem response to speech and its modulation through selective attention can employ predefined acoustic mixtures.

The decoding that we have described here is based on linear backward models that reconstruct the fundamental waveform of the speech signal from the EEG recordings. This method determined the brainstem response to the voiced parts of speech, and in particular to its pitch, but did not measure the brainstem response to the voiceless speech components (Maddox et al. 2018). Improved performance may be obtained through canonical correlation analysis that relates the neural recording to more speech features in an optimized space (de Cheveigné et al. 2018) or through an artificial neural network that is able to extract highly nonlinear relations between the two datasets (Yang et al. 2015).

Finally, decoding of auditory attention could leverage both cortical and sub-cortical responses as they can be obtained from the same EEG recordings. The framework for attentional decoding based on the brainstem response to running speech presented here could be readily extended to include cortical responses to the speech envelope, which could boost the overall decoding accuracy. Moreover, measuring both subcortical and cortical responses to speech from the same EEG data will be useful for fundamental auditory research and clinical assessment of hearing impairments.

Chapter 5

The neural response at the fundamental frequency of speech is modulated by word-level acoustic and linguistic information

The work presented in this chapter is currently under review at *NeuroImage*. I would like to express my gratitude to Dr Hugo Weissbart, who contributed to this work by sharing his EEG dataset collected in Weissbart et al. 2020.

5.1 Introduction

Spoken language consists of both lower-level acoustic as well as higher-level linguistic information that need to be rapidly and continuously processed in the brain (Brodbeck et al. 2020b; Giraud et al. 2012; Meyer 2018). The lower level acoustic processing is thereby typically attributed to the primary auditory cortex, and the processing of higher-level information to the secondary auditory cortex as well as other cortical areas such as the prefrontal cortex (Golumbic et al. 2012; Hickok et al. 2007; Peelle et al. 2010).

Linguistic processing encompasses both context-independent and context-dependent aspects. An important context-independent aspect is word frequency, that is, the frequency of a word in a large text corpus (Baayen 2001). This information has been found to be reflected in neural activity from the cerebral cortex (Brennan et al. 2012; Brennan et al. 2016; Brodbeck et al. 2018c). Context-dependent processing is another important linguistic aspect of speech encoding, especially in noisy auditory scenes. Behavioural studies have, for instance, shown that sentences with missing parts or added noisy intrusions can still be understood by the participants (Clarke et al. 2014; Dilley et al. 2010; Miller et al. 1963; Rubin 1976; Warren 1970).

The word expectancy resulting from context is reflected in cortical responses. Indeed, words elicit a cortical negativity at a latency of about 400 ms, the N400 response, and the N400 is modulated by word expectancy (Kutas et al. 1984). Word prediction and violations of such predictions are reflected in further aspects of cortical activity such as the beta- and gamma-band power, as has been found in studies using single sentences (Bastiaansen et al. 2006; Friederici 2002; Friederici et al. 1993; Kielar et al. 2014; Kutas et al. 2011). Moreover, we and others recently showed that cortical activity recorded from electroencephalography (EEG) acquired when

subjects listened to stories consisting of many sentences exhibited correlates of word surprisal, that is, of the violation of word predictions, as well as of the precision at which predictions were made (Donhauser et al. 2020; Gillis et al. 2021b; Weissbart et al. 2020). The word-level surprisal is thereby defined as the conditional probability of the current word, given previous words. The word-level precision is the inverse entropy of the word, given the past context. Cortical activity during natural story comprehension has also been found to reflect the semantic dissimilarity between consecutive words (Broderick et al. 2018, 2019).

Although the auditory system is often viewed as a feed-forward network of different neural processing stages, there exist corticofugal feedback connections from the cortex to the midbrain as well as to different parts of the auditory brainstem (Huffman et al. 1990; Winer 2005). A particular early neural response to speech that can potentially be under such top-down control is the neural tracking of the fundamental frequency. Voiced parts of speech are characterized by a fundamental frequency, typically between 100 Hz and 300 Hz, as well as many higher harmonics. The elicited neural activity as recorded by EEG exhibits a response primarily at the fundamental frequency, as well as, to a lesser extent, at the higher harmonics (Chandrasekaran et al. 2010; Skoe et al. 2010). The response has a short latency of around 10 ms and originates mainly in the auditory brainstem and in the midbrain, although cortical contributions have been discovered recently as well (Bidelman 2018; Chandrasekaran et al. 2010; Coffey et al. 2016, 2017, 2019).

The early neural response at the fundamental frequency of speech can reflect different aspects of speech processing. It can, in particular, be shaped by language experience as well as by musical training (Bidelman et al. 2011; Kraus et al. 2017; Krishnan et al. 2010; Wong et al. 2007). In addition, we recently showed that this response is modulated by selective attention to one of two competing speakers (Etard et al. 2019a; Forte et al. 2017). Moreover, a strong bi-directional coupling between cortical activity and subcortical contributions through corticofugal pathways was found in a speech-in-noise perception task (Price et al. 2021).

The frequency-following response (FFR) to the frequency of a pure tone can occur in a similar frequency range as the neural response at the fundamental frequency of speech, and presumably reflects related processing. This FFR may be under top-down control as well. An oddball paradigm in which many repeated tones are presented together with occasional deviant tones showed that the FFR is larger for expected than for unexpected ones, although a later study could not replicate the effect (Font-Alaminos et al. 2021; Slabu et al. 2012). Invasive recordings in animals likewise showed correlates of prediction errors at different subcortical as well as cortical stages (Parras et al. 2017).

Whether the early neural response at the fundamental frequency of speech is modulated by linguistic processing has not yet been investigated. A main difficulty is thereby the complexity of natural speech that complicates both the measurement of the neural response as well as the assessment of its modulation through linguistic information. However, recent studies have developed the methodology to measure the neural response at the fundamental frequency of

speech even for continuous, non-repetitive speech stimuli. We recently proposed an approach in which we extracted a fundamental waveform from voiced speech, that is, a waveform that, at each time instance, oscillated at the time-varying fundamental frequency of speech (Forte et al. 2017). We then related this waveform to EEG that was recorded simultaneously through linear regression with regularization (Etard et al. 2019a). As an alternative approach, the envelope of the higher harmonics of a speech signal is modulated by the fundamental frequency, and one can infer a neural response to this envelope modulation (Kulasingham et al. 2020).

Here we employed these two recently developed methodologies to measure neural responses at the fundamental frequency of individual words that occur in continuous natural speech. We also quantified key word features, including both acoustic and linguistic ones. We then investigated whether the early response to speech was shaped by these word features.

5.2 Materials and methods

5.2.1 Dataset

We analyzed EEG responses to continuous speech that were collected for an earlier study on cortical correlates of word prediction (Weissbart et al. 2020). The recording of this dataset is described in detail below.

5.2.2 Participants

13 young and healthy native English speakers (25 ± 3 years, 6 females) were recruited for the experiment. They were all right-handed and had no history of hearing or neurological impairment. All volunteers provided written informed consent. The experimental protocol was approved by the Imperial College Research Ethics Committee.

5.2.3 Experimental setup

The experiment consisted of a single session of EEG recording. During the experiment, the participants listened to continuous narratives in the form of audiobooks that were openly available at '*librivox.org*'. In particular, we used three short stories: '*Gilray's flower pot*', '*My brother Henry*' by J.M. Barrie and '*An undergraduate's aunt*' by F. Anstey Patten 1910[1]. Both audiobooks were read by a male speaker, Gilles G. Le Blanc. The total length of the audio material was 40 min. The stories were presented in 15 parts, each approximately 2.6 min long (2.6 min ± 0.43 min). The acoustic signals were presented to the participants through Etymotic ER-3C insert earphones (Etymotic, USA) at 70 dB SPL. The audiobooks' transcriptions used for computing word-level features were obtained from the project Gutenberg[2].

[1]https://librivox.org/international-short-stories-vol-2-by-william-patten/
[2]http://www.gutenberg.org/ebooks/32846

After each part, the participants answered multiple-choice comprehension questions presented on a monitor. Each participant was asked 30 questions throughout the experiment. The questions were designed to keep the volunteers engaged and to assess whether they paid attention to the stories. The participants answered the questions with an average accuracy of 96%, showing that they paid attention to the audio material and understood it.

5.2.4 EEG acquisition

The brain activity of the participants was measured using a 64-channel EEG system (active electrodes, actiCAP, and EEG amplifier actiCHamp, BrainProducts, Germany). The left ear lobe served as a reference. The impedance of all EEG electrodes was kept below 10 kΩ. The audio material presented to the participants was simultaneously recorded through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts, Germany) and used for aligning the EEG recordings to the audio signals. The EEG and the audio data were both recorded at a sampling rate of 1 kHz.

5.2.5 Auditory stimulus representations

For modelling the early neural response at the fundamental frequency of speech from the high-density EEG, we followed the methodology developed in Etard et al. 2019a. In particular, we used the fundamental waveform as well as the high-frequency envelope modulation extracted from the speech signals as the audio stimulus features (Fig. 5.1).

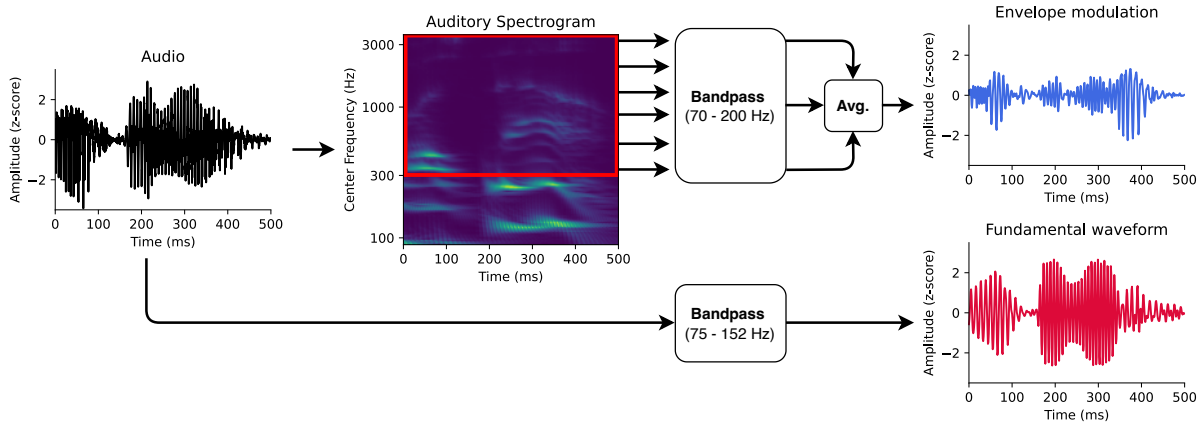

Figure 5.1: Auditory features for modelling the neural responses at the fundamental frequency. First, the fundamental waveform (red) was obtained by band-pass filtering the audio signal. Second, the high-frequency modulation of the envelope modulation was computed as well. The audio input was therefore transformed into an auditory spectrogram using a model of the auditory periphery. The frequency bins of the auditory spectrogram above 300 Hz were then filtered in the range of the fundamental frequency. The filtered bins were finally averaged to obtain the envelope modulation (blue).

The fundamental waveform is a waveform that oscillates at the fundamental frequency of

the speaker's voice. The original algorithm for computing the fundamental waveform was based on empirical mode decomposition (Forte et al. 2017; Huang et al. 2006). However, Etard et al. 2019a showed that direct band-pass filtering of the speech signal is considerably simpler, faster to compute and leads to the same result (Bachmann et al. 2021; Kulasingham et al. 2020; van Canneyt et al. 2021c, 2021d). Here, we also employed a band-pass filter to extract the fundamental waveform from the voice recordings.

We used the software Praat (Boersma 2001) and its Python interface Parselmouth (Jadoul et al. 2018) to estimate the fundamental frequency in the voice recordings presented to the participants (44,100 Hz sampling rate). The mean fundamental frequency of the speaker was 113.2 Hz with a standard deviation of 27.2 Hz. The frequencies corresponding to the 5th and 95th percentiles of the speaker's pitch distribution (75 Hz and 152 Hz, respectively) were used as corner frequencies of the bandpass filter. A FIR bandpass filter (7785th order, one-pass, zero-phase, non-causal, Hamming window, lower transition bandwidth: 18.7 Hz, upper transition bandwidth: 38.12 Hz) was then applied to filter the speech recordings. The resulting fundamental waveform was finally downsampled to 1 kHz to match the sampling rate of the EEG. The extracted signal was manually checked for artifacts to assure the validity of the automatically selected filter.

Since the neural response at the fundamental frequency might not emerge directly from the tracking of the speaker's pitch but reflect the high-frequency envelope modulation, we used the latter as an additional feature in our analysis. The high-frequency envelope modulation was extracted from the audio signal as originally introduced (Kulasingham et al. 2020). In particular, first, the audio signal was processed through a model of the auditory periphery reflecting the early stages of the auditory processing, including the cochlea, the auditory nerve and the subcortical nuclei (Chi et al. 2005)[3] to obtain the auditory spectrogram with a millisecond temporal resolution (matching the sampling rate of the EEG).

The frequency bins of the obtained auditory spectrogram corresponding to the higher harmonics above 300 Hz were then band-pass filtered in the range of the fundamental frequency, between 70 Hz and 200 Hz (177th order, one-pass, zero-phase, non-causal, Hamming window, lower transition bandwidth: 17.5 Hz, upper transition bandwidth: 50 Hz). The filtered signals were averaged to form the high-frequency envelope modulation feature. Similarly to the previous study employing the pair of features (Kulasingham et al. 2020), we found a negative correlation of $r = -0.22$ (Pearson's) between the fundamental waveform and the high-frequency envelope modulation.

[3]We used the open-source Python implementation of the model available at `https://github.com/MKegler/pyNSL`

5.2.6 EEG modelling

Firstly, the acquired EEG data (1 kHz sampling rate) was band-pass filtered between 50 Hz and 280 Hz (265th order FIR one-pass, zero-phase, non-causal filter, Hamming window, lower transition bandwidth: 12.5 Hz, upper transition bandwidth: 70 Hz) and re-referenced to the average. The pre-processed EEG data and the stimulus features obtained from the corresponding speech signal were used to fit linear models following the methodology developed in Etard et al. 2019a. EEG pre-processing and modelling pipelines were implemented through custom-written Python scripts using NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020) and MNE open-source packages (Gramfort et al. 2014).

Forward model

The forward models were designed to have complex coefficients. This approach allowed us to assess both the magnitude and the phase of the underlying neural response (Etard et al. 2019a; Forte et al. 2017). The complex forward model was designed to predict the multichannel EEG response $r(t, c)$ at channel c from the two stimulus features $f_1(t)$ and $f_2(t)$, where, $f_1(t)$ represents the fundamental waveform and $f_2(t)$ the envelope modulation (Fig. 5.1). In particular, at each time instance t, the EEG signal was estimated as a linear combination of the stimulus features $f_1(t)$ and $f_2(t)$ as well as their Hilbert transforms $f_1^{(h)}(t)$ and $f_2^{(h)}(t)$ at a time lag τ:

$$r(t, c) = \sum_{j=1}^{2} \sum_{\tau=1}^{T} [\alpha_{\tau,c,j}^{(r)} f_j(t-\tau) + \alpha_{\tau,c,j}^{(j)} f_j^{(h)}(t-\tau)] \tag{5.1}$$

where $\alpha_{\tau,c,j}^{(r)}$ and $\alpha_{\tau,c,j}^{(i)}$ are real coefficients that can be interpreted as real and imaginary parts of a complex set of coefficients $\alpha_{\tau,c,j} = \alpha_{\tau,c,j}^{(r)} + i \cdot \alpha_{\tau,c,j}^{(i)}$. These coefficients are referred to as temporal response function (TRF) since they describe the time course of the neural response r to the two stimulus features f_1 and f_2. We note that the forward models is fitted using the two stimulus features simultaneously, analogously to Kulasingham et al. 2020.

We used $T = 750$ time lags ranging from -250 ms (i.e. the stimulus is preceded by the EEG signal, thus anticausal) up to 499 ms. We chose a broad range of time lags, including a latency range typical for cortical responses, to include both early and putative late responses. The model coefficients were obtained using ridge regression (Hastie et al. 2009) with a regularization parameter $\lambda = \lambda_n \cdot e_m$, where λ_n is a normalized regularization parameter and e_m is the mean eigenvalue of the covariance matrix, to which the regularization was added (Biesmans et al. 2017). For the forward model, we used a fixed normalized regularization parameter of $\lambda_n = 1$. Prior to fitting the model, each EEG channel and the stimulus features were standardized by subtracting their mean and dividing them by their standard deviation.

A complex forward model was computed separately for each participant. The subject-specific models were then averaged to obtain a population-averaged model. The magnitudes of the complex coefficients were computed by taking their absolute values, and the phases by comput-

ing their angles. To summarize the contribution of different time lags, the magnitudes of the population-averaged model were additionally averaged across channels to obtain a single value per time lag. This value, reflecting the contribution of each time lag to the model, allowed us to estimate the latency of the predominant neural response.

To assess the significance of the forward model, we established null models using time-reversed stimulus features. Due to the mismatch between the speech features and the EEG signal, the null models were purposefully designed to reflect no meaningful brain response across the entire range of time lags. One null model was obtained for each subject. We bootstrapped the population-level null models by re-sampling null models across participants (with replacement), averaging them and computing their magnitudes across time lags in the same way as for the actual forward model. This procedure was repeated 10,000 times to form a distribution of null model magnitudes across time-lags. We therefrom computed an empirical p-value for each time lag by counting how many values from the null distribution exceeded the actual forward model for each time lag. Finally, the obtained p-values were corrected for multiple comparisons using the Benjamini-Yekutieli method (Benjamini et al. 2001).

Backward model

Backward models were designed to reconstruct the two stimulus features $f_1(t)$ and $f_2(t)$ from the time-lagged multi-channel EEG response $r(t, c)$. In particular, for each time instance t, the stimulus features were reconstructed as follows:

$$f_j(t) = \sum_{\tau=1}^{T} \sum_{c=1}^{N} \beta_{\tau,c,j} \cdot r(t-\tau, c) \tag{5.2}$$

where $\beta_{\tau,c,j}$ are real-valued model coefficients, $j \in \{1, 2\}$ denotes the stimulus feature, c represents the index of the EEG channel, and τ is a time lag between the auditory stimulus features and the EEG recording. Here, we used $T = 55$ time lags ranging from -5 ms (i.e. the EEG signal preceded the stimulus) to 49 ms. We only used real and not complex model coefficients, since the use of the former did not impact the reconstruction performance but greatly decreased the computational cost. The coefficients of the backward models were obtained in the same way as for the forward model using ridge regression. We evaluated 51 logarithmically-spaced normalized regularization parameters λ_n ranging from 10^{-10} to 10^{10}. Analogously to the forward models, the backward models were fitted using two stimulus features simultaneously.

Backward models were evaluated through five-fold cross-validation (Hastie et al. 2009). In particular, all the available data were split into five folds of the same duration of approximately eight minutes. Four folds were used to train the backward model, and the remaining one was kept aside for evaluating the model. Each time, 51 models, one for each regularization parameter, were trained and evaluated.

To investigate how the amount of available data influenced the reconstruction performance,

we split the testing data into either segments of arbitrary lengths or according to the word boundaries. The performance of each model was quantified by computing Pearson's correlation coefficient between the reconstructed stimulus features and the actual one. After evaluating the models on all segments, another fold of the data was selected as the testing set. The procedure was repeated five times until all the available data was used. This yielded reconstruction scores that reflected the strength of the neural response.

To test whether segmenting the evaluation data according to word boundaries yielded different performances in reconstructing the fundamental waveform, we compared it to the reconstruction scores obtained using segments of arbitrary duration agnostic of word onsets. For the latter, we considered six different durations of the arbitrary testing segments. Since the averaged word duration was 260 ms, we chose the fixed evaluation segment durations to be 100 ms, 260 ms (the mean word duration), 310 ms (the median word duration), 1 s, 10 s and 30 s. We evaluated the backward models as specified above for all 13 subjects. In particular, reconstruction scores from all testing segments across all the folds were averaged to summarize the model performance for each subject. For this analysis, we used the fixed normalized regularization parameter $\lambda_n = 1$.

We thereby obtained 13 averaged reconstruction scores (one per subject) for each stimulus feature (fundamental waveform and envelope modulation) and for each segment duration. For each stimulus feature, we performed the Friedman test, a non-parametric equivalent of ANOVA, to assess whether at least one of the evaluation segment lengths yielded different reconstruction scores from the others. Then, we performed a post-hoc test on the results for each pair of segment durations through the Wilcoxon signed-rank test. In addition, the reconstruction scores for the two stimulus features were compared for each segment duration. The p-values obtained from the above tests were corrected for multiple comparisons using the Benjamini-Yekutieli method (Benjamini et al. 2001).

The null models that represented the chance-level reconstruction scores were obtained in the same way as described above, but using the time-reversed stimulus features. Following the same reasoning as for the forward model, these models contained no actual brain response and estimated the chance-level reconstruction scores.

5.2.7 Word-level features

We used seven distinct word-level features to study the neural response at the fundamental frequency of the continuous narratives (Fig. 5.2A). Four linguistic features were developed in Weissbart et al. 2020 and are openly available on figshare.com[4]. In short, the transcriptions of the stories presented to the participants were processed through a language model to obtain the frequency, surprisal and precision of each word. The word frequency reflects the probability of a word out of context and was computed from Google N-grams by taking only the unigram values. As a result, this feature estimated the unconditional probability of the occurrence of

[4] 10.6084/m9.figshare.9033983.v1

a word $P(w)$. To match the remaining information-theoretic features, we computed the negative logarithm of this probability $-ln(P(w))$, and refer to this feature as the 'inverted word frequency' in the following. Importantly, less frequent words were therefore assigned a higher inverted word frequency, and more frequent words were assigned lower values.

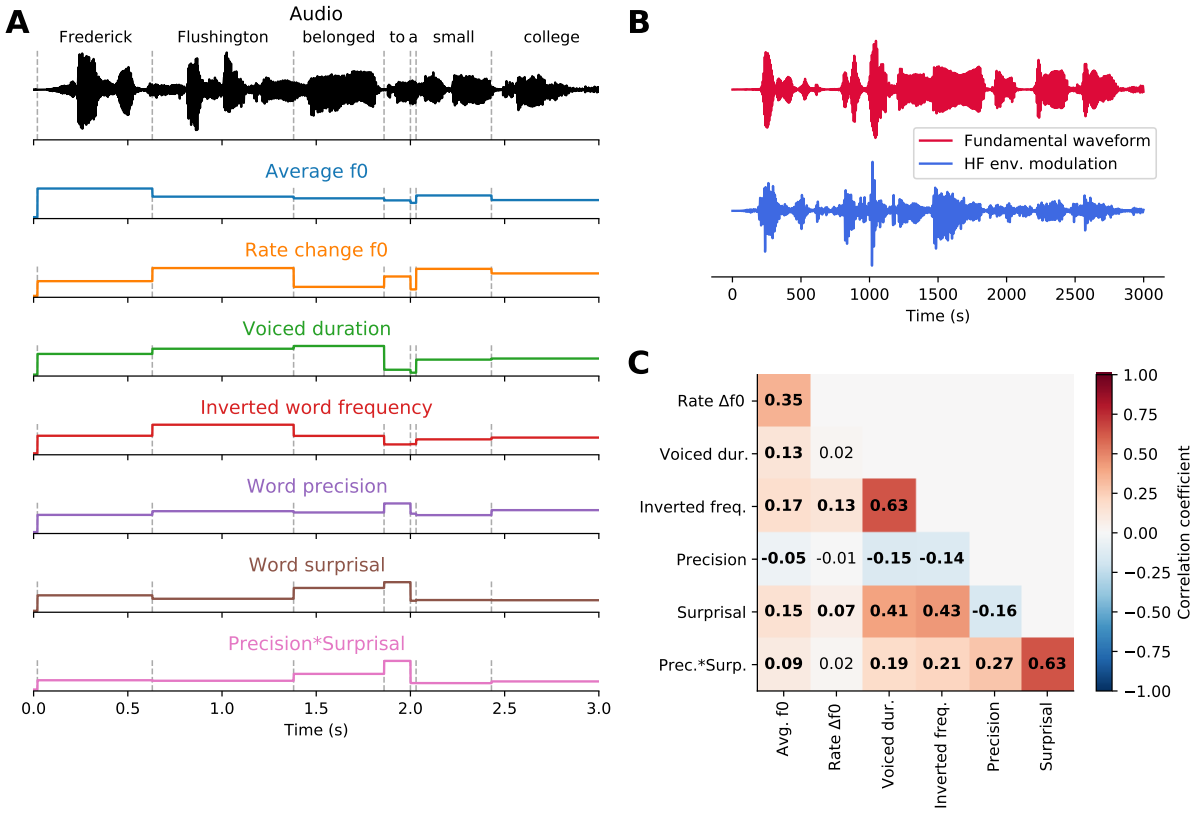

Figure 5.2: Word-level features. (**A**), An exemplary part of a speech signal (black). Dashed vertical lines represent word onsets. Seven features (coloured) were used to describe each word in the stories presented to the participants. Three of them – the averaged fundamental frequency (f0), the rate of f0 change and the duration of the voiced part – were acoustic features based on the voiced parts of each word. The remaining four features – inverted word frequency, word surprisal, word precision and the interaction of precision and surprisal – were derived from a language model and characterized linguistic properties of each word. (**B**), The two stimulus features extracted from the exemplary audio segment, the fundamental waveform (red) and the high-frequency (HF) envelope modulation (blue). (**C**), Pairwise correlations between word-level features. Significant correlations ($p < 0.05$, after correcting for multiple comparisons using the Benjamini-Yekutieli method) are denoted in bold.

In contrast to the inverted word frequency, the word precision and surprisal were derived from conditional probabilities of a particular word given the preceding words. In particular, the probability of the nth word w_n can be expressed as $P(w_n|w_1, w_2, ..., w_{n-1})$. Word surprisal quantifies the information gain that an upcoming word generates given the previous words and reflects how unexpected the word is in its context. Here, the surprisal of the word w_n was

computed as the negative logarithm of its conditional probability:

$$S(w_n) = -ln[P(w_n|w_1, w_2, ..., w_{n-1})]. \tag{5.3}$$

In contrast to the word surprisal, the word precision reflects the confidence about the prediction of the next word given the previous words. Here, the word precision was computed as the inverse word entropy, $[E(w_n)]^{-1}$. On its own, the word entropy represents the uncertainty of predicting the next word w_n from the past context $(w_1, w_2, ..., w_{n-1})$, and is formulated as:

$$E(w_n) = -\sum_{w_k} P(w_k|w_1, w_2, ..., w_{n-1}) \cdot ln(P(w_k|w_1, w_2, ..., w_{n-1}), \tag{5.4}$$

where w_k denotes the kth word from the text corpus.

Finally, to investigate a possible modulating effect that precision may have on surprisal, an interaction term was obtained by multiplying precision with surprisal. This feature can be interpreted as a confidence-weighted surprisal or a surprisal-dependent precision.

The conditional probabilities, required for computing the word surprisal and precision, were obtained from a recurrent neural network (RNN) language model introduced in Mikolov et al. 2011. The model was designed to predict the current word w_n given the previous words $w_1, w_2, ..., w_{n-1}$. Firstly, embeddings of words in the input text were obtained using pre-trained global vectors for word representation (GLOVE) trained on the Wikipedia 2014 and Gigaword 5 datasets (Pennington et al. 2014). The obtained embeddings were projected to 350 recurrent units forming the hidden layer of the model. The output layer of the model was a softmax function, from which the word probabilities were computed. Such model was trained on the *text8* dataset, consisting of 100 MB of text data from Wikipedia (Mahoney 2011), using backpropagation through time and a 0.1 learning rate. Prior to the training, the text data was cleaned to remove punctuation, HTML, capitalization and numbers. In addition, to facilitate model training, the vocabulary was limited to the 35,000 most common words in the dataset. The remaining rare words were assigned an '*unknown*' token. For more implementation details of the model itself and its training, please see Weissbart et al. 2020 where the method was originally developed.

Having obtained the above-described linguistic features, each word in the story was aligned to the acoustic signal using a forced alignment algorithm implemented in the Prosodylab-Aligner software (Gorman et al. 2011). Subsequently, we computed three additional acoustic features for each word. In particular, we used the Praat & Parselmouth Python interface (Boersma 2001; Jadoul et al. 2018) to obtain the evolution of the speaker's fundamental frequency across the story recording.

For each word, we then computed the duration of its voiced part, its mean fundamental frequency and the rate of the change in the fundamental frequency. The latter feature was

obtained by averaging the absolute value of the first derivative of the fundamental frequency's time course across the voiced duration, as described by van Canneyt et al. 2021d. Including these three features in our analysis allowed us to control for purely acoustic modulation of the neural response at the fundamental frequency.

5.2.8 Stepwise hierarchical regression

We first determined the strength of the neural response at the fundamental frequency for the ith word. To this end, the backward models for each participant were evaluated to obtain a reconstruction score for each word in the story ($N = 6,345$). 100 words did not contain a voiced part, and were therefore discarded from further analysis. For each remaining word, we obtained 51 reconstruction scores, one for each normalized regularization parameter λ_n. We picked the optimal regularization parameter λ_{model}, leading to the best reconstruction.

To control for overfitting, the same procedure was applied to the backward null models that did not contain a meaningful brain response. Possible inflation of the reconstruction score $r(i)$ from overfitting was corrected by subtracting the score obtained by the null model $r_{null}(i)$ from that of the actual decoder $r_{model}(i)$:

$$r(i) = \max_{\lambda_{model}} r_{model}(i) - \max_{\lambda_{null}} r_{null}(i). \tag{5.5}$$

The above procedure was applied independently to the reconstruction scores obtained for the two stimulus features, the fundamental waveform and the envelope modulation.

We note that the optimal word-level regularization parameter was picked independently for the actual (λ_{model}) and the null model (λ_{null}). Controlling for overfitting in this empirical manner allowed to avoid pre-selecting a fixed regularization parameter, which could either inflate or deflate reconstruction scores. However, as an additional control, we also computed the reconstruction scores with a pre-selected regularization parameter of $\lambda_{model} = 1$. The resulting reconstruction scores were not significantly different from those obtained with the procedure outlined above ($p > 0.121$, Wilcoxon signed-rank test).

Having computed the single-word reconstruction scores for each participant, we averaged them across the subjects for each word in the story to obtain population-level single-word reconstruction scores.

We then investigated whether the word-level acoustic and linguistic features modulated the early neural response at the fundamental frequency, that is, whether they modulated the single-word reconstruction scores. To this end, we first standardized both the single-word reconstruction scores r and the word-level features x by subtracting their mean and dividing them by their standard deviation. In addition, we used the isolation forest (Liu et al. 2008), an unsupervised algorithm based on the random forest, for detecting outliers and anomalies.

In this method, data points corresponding to the words in the stories and described by the word-level features and the reconstruction scores (eight descriptors) were processed through a set of 1,000 random trees established based on the dataset statistics. The algorithm measured the average path it took each data point to traverse from the roots of the trees in the random forest to their leaf nodes. Since the outliers contained extreme descriptor values, they reached leaf nodes earlier, yielding shorter paths. The threshold for the path length qualifying the data point as an outlier was determined automatically based on the dataset-wide statistics. Here, we used the implementation of the method included in the scikit-learn Python package (Pedregosa et al. 2011). Following the outlier removal, 5,732 data points corresponding to different words in the stories remained.

We then related the single-word reconstruction scores to the word-level features through stepwise hierarchical regression. The approach was inspired by the stepwise and hierarchical regression commonly used for feature selection in multiple regression models (Lewis 2007). However, neither of the standard approaches is suited for the cases in which the explanatory variables exhibit a degree of multicollinearity. Since the word-level linguistic and acoustic features were indeed correlated (Fig. 5.2C), we employed a stepwise approach based on the expected effect size for each feature, in a hierarchical manner.

The n word-level features were first ordered from that with the highest expected predictive power, $x^{(1)}$, to that with the lowest expected predictive power, $x^{(n)}$. In the first step of the procedure, the word-level feature $x^{(1)}$ from the ordered list was used to fit a linear model to predict word-level reconstruction scores r: $\hat{r}_i = ax_i^{(1)}$ with a coefficient a, assuming that r and x were standardized. In this equation, x_i denotes the world-level features of the ith word, and r_i its reconstruction score.

The feature was then projected out from the word-level reconstruction scores r by subtracting the estimated word-level reconstruction scores $\hat{r}_i$ from the actual ones: $r_i^{(1)} = r_i - \hat{r}_i$. The residual reconstruction scores $r_i^{(1)}$ were used as a response variable for fitting the next linear model using the next word-level feature $x^{(2)}$ from the ordered list. The process was repeated until all available word-level features were used.

By projecting out the predictions obtained from subsequent word-level features, we assured that a possible predictive contribution of a feature with a lower expected predictive power did not result from that of a feature with a higher expected predictive power due to shared variance. The stepwise hierarchical regression therefore constituted a conservative manner to ensure that any contributions from features with lower expected predictive power were indeed real, and that their significance was not inflated due to the shared variance with features with a higher expected predictive power.

To further reduce the influence of the extreme data points on the model coefficients, we

fitted the linear models in the stepwise hierarchical regression through robust regression, using Huber weighting of the residuals (Andrews 1974), instead of the ordinary least squares regression.

Regarding the ordering of the different features, due to the reported significant impact of the acoustic features on the neural response at the fundamental frequency (Saiz-Alía et al. 2019, 2020; van Canneyt et al. 2021d), we prioritized these above the linguistic features that likely have a weaker impact. We adopted the following ordered list of the word-level features for the stepwise hierarchical regression: (1) average fundamental frequency f0, (2) rate of the change in f0, (3) duration of the voiced part, (4) inverted word frequency, (5) word precision, (6) word surprisal, and (7) precision x surprisal.

The stepwise hierarchical regression was applied for both stimulus features, the fundamental waveform and the envelope modulation, for which the reconstruction scores were computed. Each time, the output of the procedure was a set of seven linear models corresponding to the seven word-level features. The p values reflecting the significance of each model coefficient were corrected for multiple comparisons using the Benjamini-Yekutieli method (Benjamini et al. 2001).

The methods described above were implemented via custom-written Python scripts using NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020) and statsmodels open-source packages (Seabold et al. 2010).

5.3 Results

5.3.1 Relations between the word-level acoustic and linguistic features

We computed Pearson's correlation coefficient between each pair of features (Fig. 5.2C). The obtained correlation ranged from -0.157 to 0.632. The highest correlation coefficient ($r = 0.632$) emerged between surprisal and the interaction of surprisal and precision. Another significant positive correlation ($r = 0.431$) arose between inverted word frequency and surprisal, indicating that less frequent words tended to be more surprising. Similarly, a positive correlation ($r = 0.406$) between voiced duration and surprisal showed that more surprising words had longer voiced parts and were presumably longer overall. A positive correlation ($r = 0.632$) between inverted word frequency and voiced duration indicated that less frequent words tend to be longer. The remaining correlations between features were comparatively small, between -0.157 and 0.274. The rate of fundamental frequency change was the least correlated with other features. It was only significantly correlated with the inverted word frequency ($r = 0.129$).

5.3.2 Early neural response at the fundamental frequency

We investigated the neural response at the fundamental frequency through the temporal response functions (TRFs) obtained from the forward model (Haufe et al. 2014). In particular,

we examined the model coefficients associated with the two considered stimulus features, the fundamental waveform and the high-frequency envelope modulation.

For the neural response to the fundamental waveform (Fig. 5.3A-C), the channel-averaged TRFs yielded significant responses for short delays between 9 ms - 11 ms, with a peak at 10 ms. The TRFs at the peak delay showed the highest magnitudes in the central-frontal and occipital regions, as well as at the mastoid electrodes. The phase relationship of the model coefficients at the delay of 10 ms exhibited a phase shift of approximately π between the frontal and occipital areas, and a slightly larger phase difference between the central-frontal and the mastoid electrodes.

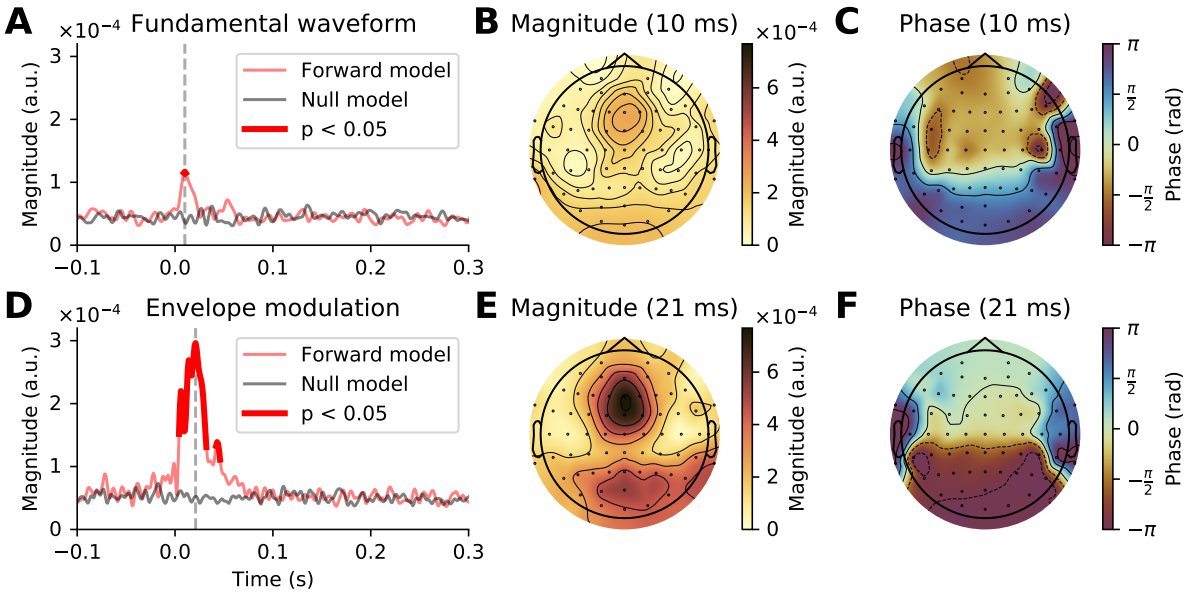

Figure 5.3: Early neural response at the fundamental frequency of continuous speech. (**A - C**), Response to the fundamental waveform. (**A**), The magnitude of the complex coefficients of the forward model (red), averaged across EEG channels and subjects exhibited a peak at an early latency of 10 ms (grey dashed line). The comparison of the complex TRF magnitudes to a null model (black solid line) showed that significant responses emerge only at latencies around the peak latency, between 9 ms and 11 ms (thicker red line, $p < 0.05$, corrected for multiple comparisons). (**B**), At the peak latency of 10 ms, the largest contribution to the TRF came from central-frontal and occipital areas, as well as from the mastoid electrodes. (**C**), The phase of the model coefficients indicated a phase shift of approximately π between the frontal area on the one hand and the occipital and mastoid electrodes on the other hand. (**D - F**), Neural response to the high-frequency envelope modulation. (**D**), The average magnitude of the complex TRF coefficients was substantially larger than that of the response to the fundamental waveform. In particular, the coefficients of the model significantly exceeded the chance level between 4 ms to 37 ms and 42 ms to 46 ms, with the peak magnitude at 21 ms. (**E, F**), At the peak latency of 21 ms the TRFs exhibited similar topographic patterns to those obtained for the response to the fundamental waveform.

For the response to the envelope modulation (Fig. 5.3D-F), the channel-averaged TRFs showed significant contributions between 4 ms - 37 ms as well as 43 ms - 46 ms, with a peak at 21 ms. Notably, the averaged magnitudes of the models coefficients for this response were

nearly three times larger than for the response to the fundamental waveform. Despite the peak magnitude occurring later, the topographical pattern of the model coefficients, both in terms of magnitudes and phases, were similar to that obtained for the response to the fundamental waveform. In particular, the largest magnitudes were obtained for frontal and occipital regions, with a phase difference of approximately π between them.

5.3.3 Reconstruction of the stimulus features from EEG

We then assessed the reconstruction of the stimulus features from the EEG recordings using backward models (Fig. 5.4). In particular, we investigated whether the reconstruction performance varied with the duration of the speech segment on which the models were tested.

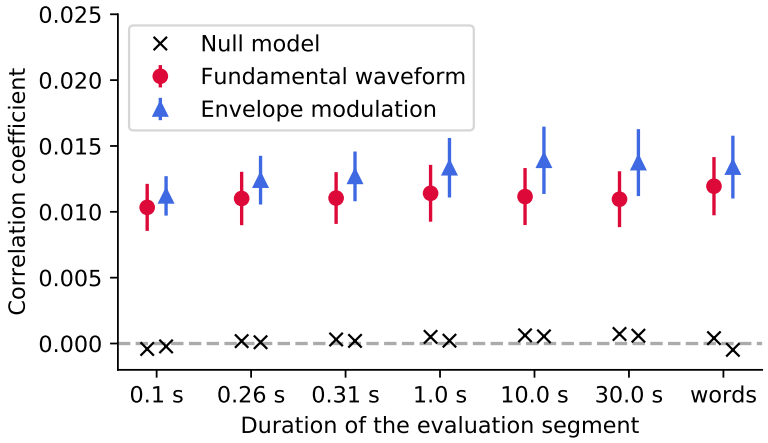

Figure 5.4: Reconstruction of the stimulus features from EEG. We evaluated the reconstruction of the stimulus features from the EEG recordings using segments of different duration, including segments aligned with the word boundaries (*words*). For each segment duration, the population-averaged reconstruction scores obtained for the fundamental waveform feature are denoted with red circles, and for the envelope modulation with blue triangles. The error bars correspond to the standard error of the mean across participants. For both features, the reconstruction scores yielded much higher correlation coefficients as compared to their respective null models (black crosses).

The Friedman test was applied to the reconstruction scores to assess whether either of the segment duration yields significantly better reconstructions than the other. For both features, the test yielded significant results at $p < 0.048$ (corrected for multiple comparisons), suggesting that at least one segment duration yielded different reconstruction scores from the rest. However, post-hoc Wilcoxon signed-rank tests performed between different segment durations yielded no significant results after correcting for multiple comparisons ($p > 0.099$).

Similarly, the reconstruction scores obtained for the two considered stimulus features were not significantly different (Wilcoxon signed-rank test, $p > 0.26$, corrected for multiple comparisons).

5.3.4 Modulation of the early neural response at the fundamental frequency through acoustic and linguistic features

We used stepwise hierarchical regression to investigate the acoustic and linguistic modulation of the early neural response at the fundamental frequency of continuous speech. Through this method, we predicted the word-level reconstruction scores of the backward model, reflecting the strength of the neural response, from the seven word-level features (Fig. 5.2A).

We first predicted the reconstruction scores related to the fundamental waveform (Table 5.1, Fig. 5.5). Of the considered acoustic word-level features, the average fundamental frequency (f0) of a word's voiced part (-0.081, $p = 1 \cdot 10^{-8}$) and the rate of change of the fundamental frequency (-0.082, $p = 3 \cdot 10^{-5}$) both yielded significant model coefficients with similar values. The negative values of the model coefficients showed that higher average fundamental frequency and higher associated variability leads to less neural tracking of the fundamental waveform. However, neither the duration of a word's voiced part nor any of the four considered linguistic features had a significant influence on the reconstruction scores ($p > 0.285$).

Table 5.1: Word-level modulation of the neural response to the fundamental waveform. The table presents the model coefficients obtained from stepwise hierarchical regression, using the reconstruction scores of the fundamental waveform. It details the model coefficient (Coeff.), the standard error (SE), the 95% confidence interval (CI), the z statistic (z) and the p-value after the FDR correction for multiple comparison using the Benjamini-Yekutieli method. Word-level features that yield a significant contribution are denoted in bold, as well as with an asterisk

Feature	Coeff.	SE	95% CI	z	p (FDR)
Average f0 (*)	-0.081	0.013	$(-0.106; -0.055)$	-6.145	$1 \cdot 10^{-8}$
Rate change f0 (*)	-0.082	0.018	$(-0.117; -0.047)$	-4.640	$3 \cdot 10^{-5}$
Voiced duration (n.s.)	-0.016	0.013	$(-0.042; 0.010)$	-1.185	0.857
Inverted word frequency (n.s.)	-0.027	0.014	$(-0.054; -0.001)$	-1.984	0.286
Word precision (n.s.)	-0.039	0.039	$(-0.115; 0.037)$	-1.008	0.948
Word surprisal (n.s.)	-0.019	0.014	$(-0.047; 0.009)$	-1.336	0.823
Precision x Surprisal (n.s.)	-0.005	0.023	$(-0.050; 0.040)$	-0.222	1.000

We then investigated which word features could predict the reconstruction scores of the high-frequency envelope modulation (Table 5.2, Fig. 5.6). As for the neural response to the fundamental waveform, both the average fundamental frequency and its rate of change significantly modulated the reconstruction scores. In particular, the average fundamental frequency was related to an even larger negative coefficient (-0.168, $p = 4 \cdot 10^{-36}$) than for the neural response to the fundamental waveform. In contrast, the rate of change of the fundamental frequency within words led to a slightly smaller negative coefficient (-0.066, $p = 0.002$). The duration of the voiced portion of each word did not significantly modulate the reconstruction scores.

Importantly, the inverted word frequency, the 4^{th} feature in the hierarchy, was a significant

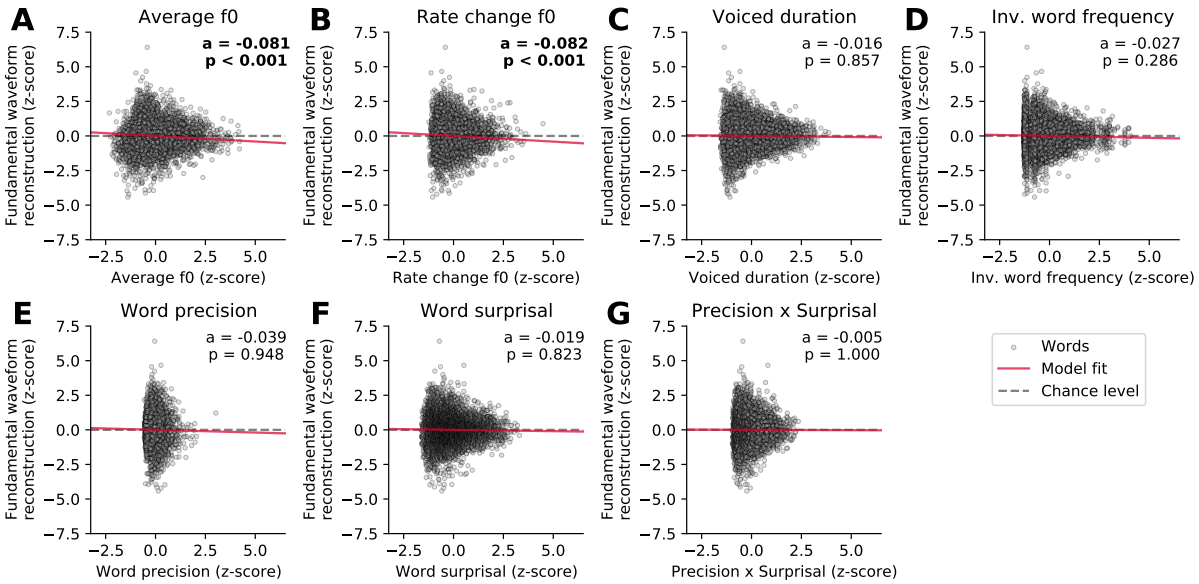

Figure 5.5: Dependency of the strength of the neural response to the fundamental waveform on the different word-level features. Panels **A** - **G** show the standardized single-word reconstruction scores averaged across 13 participants against the seven standardized word-level features. Each scatter plot shows data points corresponding to 5,732 words from the story presented to the participants during the EEG acquisition. The slopes of the red lines correspond to the coefficients obtained from the stepwise hierarchical regression. Each panel also depicts the model coefficient (a) for a given feature and the associated p-value (FDR-corrected). The gray dashed lines are horizontal and indicate no dependency.

Table 5.2: Word-level modulation of the neural response to the high-frequency envelope modulation. The table presents model coefficient obtained from stepwise hierarchical regression. It lists the model coefficient (Coeff.), the standard error (SE), the 95% confidence interval (CI), the z statistic (z) and the p-value after the FDR correction for multiple comparison using the Benjamini-Yekutieli method for the different word-level features. Word-level features that yield significant contributions ($p < 0.05$) are denoted in bold, as well as with an asterisk.

Feature	Coeff.	SE	95% CI	z	p (FDR)
Average f0 (*)	-0.168	0.013	$(-0.194; -0.142)$	-12.782	$4 \cdot 10^{-36}$
Rate change f0 (*)	-0.066	0.018	$(-0.101; -0.031)$	-3.713	0.002
Voiced duration (n.s.)	-0.015	0.013	$(-0.041; 0.011)$	-1.113	1.000
Inverted word frequency (*)	-0.039	0.014	$(-0.066; -0.011)$	-2.743	0.037
Word precision (n.s.)	-0.037	0.039	$(-0.114; 0.039)$	-0.956	1.000
Word surprisal (n.s.)	-0.003	0.014	$(-0.031; 0.025)$	-0.202	1.000
Precision x Surprisal (n.s.)	0.005	0.023	$(-0.040; 0.050)$	0.226	1.000

predictor of the reconstruction scores related to the envelope modulation. This linguistic feature was assigned a small, but significant, negative coefficient (-0.039, $p = 0.037$), indicating that less frequent words (with higher inverted frequency value) led to less neural tracking of the high-frequency envelope modulation. In contrast, none of the context-dependent word-level features (precision, surprisal and their interaction) yielded significant model coefficients ($p = 1.0$, for all three features).

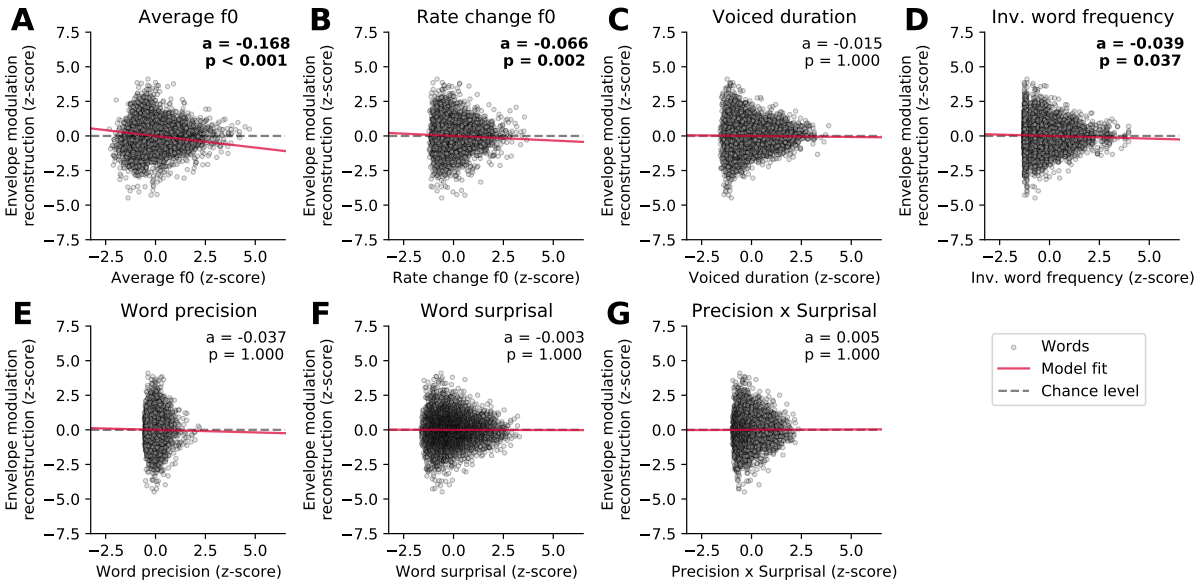

Figure 5.6: Dependency of the strength of the neural response to the high-frequency envelope modulation on the different word-level features. The world-level features and the reconstruction scores were standardized. The data points in each plot correspond to 5,732 words, and the slopes of the red lines show the coefficients of the stepwise hierarchical regression. We also detail the model coefficient (a) and the associated p-value (FDR-corrected). The gray dashed lines are horizontal and indicate no dependency.

5.4 Discussion

We showed that the word-level early neural response at the fundamental frequency of natural speech is modulated predominantly by acoustic features, but also by one of the four considered linguistic features, the inverted word frequency. Previous studies have shown significant modulation of the neural response at the fundamental frequency by acoustic differences between different speakers (Saiz-Alía et al. 2019, 2020; van Canneyt et al. 2021d). Here, we extended these findings by showing that the same effect persists for fluctuation of acoustic properties between distinct words produced by the same speaker.

The fundamental frequency of the speech that we employed varied between 75 Hz and 152 Hz. The neural response occurred accordingly at comparatively high frequencies. It could be evoked either directly by the fundamental frequency or by the high-frequency modulations of higher harmonics (Kulasingham et al. 2020). We considered both of these features and included them in our EEG modelling framework. We found that the response associated with the high-frequency envelope modulation was considerably stronger than that associated with the fundamental waveform. as observed previously in MEG recordings (Kulasingham et al. 2020). We furthermore found that the response associated with the fundamental waveform occurred earlier, around 10 ms, as compared to that associated with the high-frequency envelope modulation, at about 21 ms and at 44 ms.

The neural response at the fundamental frequency, as well as the related FFR to pure tones, is mostly attributed to the subcortical nuclei, the inferior colliculus and the medial geniculate body (Chandrasekaran et al. 2010; Skoe et al. 2010). However, recent MEG and EEG investigations have also identified a cortical contribution, in particular at frequencies below 100 Hz (Bidelman 2018; Coffey et al. 2016, 2017; Gorina-Careta et al. 2021). Regarding the measurements presented here, the earlier response associated to the fundamental waveform (at a delay of 10 ms) may have resulted predominantly from the brainstem and midbrain, as suggested by the low latency, high frequency and sensor-space topography of the response.

The later response to the high-frequency envelope modulation (at a delay of about 21 ms as well as at 44 ms) might, however, represent cortical contributions. Previous MEG recordings did indeed find cortical responses to the high-frequency envelope modulation of speech at a delay of about 40 ms (Kulasingham et al. 2020). The latencies of this response are similar to those in the auditory middle latency response that is assumed to originate in Heschl's gyrus (Borgmann et al. 2001; Liegeois-Chauvel et al. 1994; Yoshiura et al. 1995). However, due to the considerable autocorrelation of the stimulus features, our measurements did not allow us to further resolve these different neural components in the temporal domain, and the relatively low spatial resolution of our EEG measurements prevented us from more detailed spatial source localization as well. We could therefore not distinguish whether the modulation of the neural response at the fundamental frequency through the acoustic and linguistic features occurred at the subcortical level, at the cortical level, or at both.

Irrespective of the precise neural origin of the response, however, the small latency of the response implies that its modulation through the linguistic features must result from feedback from higher cortical areas at which the linguistic information in speech is processed. If the relevant contribution to the neural response originates from subcortical areas, such as the inferior colliculus, this would require corticofugal feedback to be involved in linguistic processing. If a cortical source of the neural response was modulated by the linguistic features, then the linguistic processing would involve feedback projections between different cortical areas.

To investigate the modulation of this neural response by the different acoustic and linguistic word-level features, we developed the methodology to estimate the neural response at the fundamental frequency at the word level. We tested the validity of our method by comparing the accuracy of the stimulus feature reconstruction by the backward models for different lengths of audio segments. As expected, since the models were optimized on the same training data, the segmentation of the evaluation set did not impact the feature reconstruction scores. Furthermore, we did not find a significant difference in the reconstruction performance between the two stimulus features, the fundamental waveform and the high-frequency envelope modulation.

We employed three acoustic features, the average fundamental frequency, the rate of the fundamental frequent change and duration of the voiced portion of a word. As discussed above,

the first two word-level acoustic features strongly modulated the neural response at the fundamental frequency, both that related to the fundamental frequency and that related to the high-frequency envelope modulation. The stronger modulation of the neural tracking of the high-frequency envelope modulation might be explained by the slightly stronger neural response to this feature.

The duration of the voiced portion of words in the story, however, had no significant impact on the neural tracking of either of the stimulus features. We note that we excluded entirely voiceless words from the analysis, since we could not infer a neural response for those. The neural response at the fundamental frequency is accordingly relatively similar for shorter and for longer voiced durations. Although longer voice durations will allow a better estimate of the neural response, that is, at a better signal-to-noise ratio, the response itself is indeed expected to stay constant. In other words, while longer segments of training data will lead to a more accurate backward model, the model's inference capability is independent of the duration of the data on which it is tested. This result concurs with our finding that the strength of the neural response remains unaffected by the duration of the data on which it is evaluated (Fig. 5.4).

Regarding the linguistic features, we considered four different ones: the inverted frequency of a word irrespective of its context, the surprisal of a word in its context, the associated precision, and the interaction of the surprisal and the precision. We found that the inverted word frequency had a small but significant impact on the neural response: words with a higher frequency (i.e. probability out of context) led to a larger response. Because listeners are exposed to more common words more often, this modulation may emerge due to the long-term plasticity. Similar modulation has been observed before in FFR, where the strength of the response was strongly modulated by the language experience or musical training (Bidelman et al. 2011; Krishnan et al. 2010; Krizman et al. 2019).

Importantly, this effect was present only for the neural response to the high-frequency envelope modulation, but not for that to the fundamental frequency. Because, as discussed above, the former response may contain more cortical contributions that the latter response, the modulation of the neural response by the word frequency may emerge from a cortical rather than subcortical origin.

The remaining context-sensitive word-level features did not yield a significant modulation of the neural response at the fundamental frequency. If such a modulation existed, its magnitude were accordingly too weak to be detected in the non-invasive EEG recording.

In summary, we found that the early neural response at the fundamental frequency of speech is predominantly modulated by acoustic features, but also by a linguistic feature, the frequency of a word. The latter result suggests that linguistic processing at the word level involves feedback from higher cortical areas to either very early cortical responses or even further to subcortical structures. We expect that the further investigation of the underlying neural mechanisms will

increasingly clarify the role and importance of feedback loops in spoken language processing, with potential applications in speech-recognition technology.

Chapter 6

Conclusions and future work

6.1 Summary

Understanding the neural mechanisms of speech perception is of importance for a range of applications. Firstly, it is crucial for clinical practice. Current batteries of audiological tests, such as pure tone audiometry (PTA) or click-evoked auditory brainstem response, allow the audiologist to only diagnose lacks in *audibility*. However, the results of these tests, and the associated hearing prosthetic and/or rehabilitation advice, often fail to meet patients' needs. This might be because synthetic test stimuli used in the clinic, such as pure tones or clicks, do not directly inform about the patient's *comprehension*. Despite decades of research, neither the use of naturalistic speech stimuli nor non-invasive electrophysiology is an integral part of audiological practice. Due to the still limited knowledge of mechanisms underlying speech perception, none of the recently discovered biomarkers of speech comprehension was yet to become a part of the standard clinical audiologist toolbox (Etard et al. 2019b; Iotzov et al. 2019; van Canneyt et al. 2021a; Vanheusden et al. 2020; Vanthornhout et al. 2018).

For similar reasons, auditory brain-computer interfaces (BCIs) have not yet been applied outside research laboratories. While more effective novel machine learning methods, such as unsupervised learning (Geirnaert et al. 2021a, 2022) or deep neural networks (DNNs) (Accou et al. 2021; Das et al. 2020; Vandecappelle et al. 2021), can significantly improve the performance and usability of auditory BCIs, a better understanding of neural mechanisms underlying speech perception could further improve their performance and/or reduce the computational complexity. Likewise, automatic speech processing can also benefit from a deeper understanding of how the brain processes speech. While modern systems, often based on DNNs, implement some features of the human auditory system, such as separation of sources (Luo et al. 2019; Wang et al. 2018), recovering missing or distorted parts of utterances (Kegler et al. 2020) or adapting to a range of different tasks (Beckmann et al. 2021; Elbanna et al. 2022; Niizumi et al. 2021; Scheidwasser-Clow et al. 2021), they often require extensive computational resources. By deepening our understanding of neural mechanisms of speech processing, it might be possible to translate certain algorithmic principles from the incredibly energy-efficient brain to computationally-heavy artificial systems.

The work conducted in this thesis sought to develop novel computational models charac-

terizing neural mechanisms underlying speech processing across different time scales and stages of neural processing in the auditory pathways. In particular, the developed models focused on either speech-in-noise encoding through coupled oscillations or the top-down modulation of the early neural responses to speech. The remainder of this summary is split into two sections, each addressing one of the above-outlined groups of models.

6.1.1 The role of cortical oscillations in speech-in-noise perception

While the correlation-based neuroimaging studies (i.e., correlating neural activity with experimental conditions) support the functional role of *neural entrainment* in speech processing, they do not provide direct evidence for the causal role of this mechanism in speech perception. Recently, studies utilizing non-invasive brain stimulation sought to perturb neural oscillations during the speech-in-noise listening tasks to investigate the effects on participants' comprehension. In particular, sine-wave tACS has been shown to modulate the comprehension of rhythmic speech in noise (Riecke et al. 2018; Zoefel et al. 2018). Similar results have been observed when the tACS waveform was derived from the envelope of that target talker in the experiments employing non-rhythmic speech in noise (Kadir et al. 2019; Wilsch et al. 2018). While these findings support the causal role of neural oscillations in speech processing, none has shown a significant improvement in speech comprehension, just modulation, and the optimal stimulation condition was often inconsistent across a cohort of participants. In fact, Erkens et al. 2020 recently failed to replicate the results from the previous studies. These unclear or even contradicting results suggest either a suboptimal design of the stimulation protocol or the lack of fundamental understanding of how the external current influences cortical circuits involved in speech processing.

To shed light on the inconclusive results of the previous studies and overcome their limitations, in **chapter 2** we proposed a new experimental protocol for studying the causal role of neural delta- and theta-band oscillations in speech in noise comprehension. In particular, unlike previous studies, we designed two narrowband envelope stimulation waveforms to enhance the target talker's voice masked by the background noise. By phase-shifting the stimulation waveform, we quantified the evolution of participants' speech comprehension as a function of the stimulation phase with respect to the envelope of the target speech stimulus. We found that only stimulation in the theta-band frequency range yielded significant phase-dependent modulation of speech in noise comprehension. Notably, this modulation was consistent across participants and significantly improved their speech in noise comprehension by approximately 6%. However, this improvement might be even higher for hearing-impaired listeners, whose speech comprehension tends to be facilitated via tACS more than that of young and healthy listeners (Erkens et al. 2021).

The experimental study was conducted in parallel to the development of the spiking neural network model of cortical encoding of speech in noise, described **in chapter 3**. Based on the recent theory of speech encoding through coupled oscillations (Giraud et al. 2012; Hyafil et al. 2015), the model was designed to process spoken sentences embedded in noise. We assessed

its speech encoding performance by analysing the generated spiking patterns obtained for sentences in different levels of background noise. The model's encoding performance decayed in a sigmoidal fashion with increasing background noise, which closely resembled the normal-hearing adult speech in noise comprehension. Having shown that the model may be able to estimate human speech in noise comprehension, we used it to simulate the experiments introduced in chapter 2 (Keshavarzi et al. 2020a), as well as other recent tACS studies (Kadir et al. 2019; Keshavarzi et al. 2020b; Wilsch et al. 2018). In particular, we simulated different types of tACS interventions applied to the neural network as it was encoding sentences in noise. The obtained changes in the encoding accuracy of the model matched the effects that tACS had on the participants' comprehension in the experimental studies. By studying the model dynamics, we found that theta-band tACS had a major impact on the slow oscillations segmenting the utterance into syllables. In turn, the temporal modulation of the high-frequency gamma activity via tACS led to comparatively smaller effects on the model is speech encoding performance. The investigation of the model behaviour suggests that tACS-induced modulation of speech in noise comprehension emerges from the alteration of neural dynamics of targeted cortical networks, as the model predictions agree with experimental findings. This, in turn, indicates that neural entrainment of neural oscillations in the theta frequency range plays a causal role in cortical speech-in-noise processing and actively facilitates speech comprehension.

6.1.2 Mechanisms of cognitive top-down modulation of early neural responses to speech

While late, low-frequency cortical responses have traditionally been the focus of studies investigating neural mechanisms of speech processing, in recent years, early and rapid subcortical responses have drawn much of researchers' interest (Bachmann et al. 2021; Krizman et al. 2019). Speech stimuli in the form of isolated syllables or short words have been traditionally used to study the speech processing in subcortical structures of the human auditory pathways (Skoe et al. 2010). However, hundreds of repetitions of syllables or words are far from how humans communicate on a daily basis. Following this motivation, a notable portion of the research efforts focused on developing methods for studying subcortical responses to speech in the form of long narratives, such as audiobooks. In particular, Forte et al. 2017 was the first to propose a cross-correlation based method for detecting neural responses at the fundamental frequency of the talker is voice in continuous narratives. Around the same time, Maddox et al. 2018 established another modelling framework based on the mapping of the half-wave rectified speech stimulus to the EEG recordings. The method has been further refined and validated in-depth in Polonenko et al. 2021. The two approaches yielded similar results when compared side-by-side (Bachmann et al. 2021; Bachmann et al. 2020). However, they varied in terms of attentional modulation of the detected response. Only Forte et al. 2017 found significant attention-dependent differences in the detected response.

Unlike most high-density EEG experiments focused on cortical responses, the above-outlined methods for detecting subcortical responses to speech often rely on the specialized dipolar record-

ing setup with a high sampling rate and ideally additional pre-amplifiers. In **chapter 4** we proposed a novel method for studying early neural responses at the fundamental frequency using conventional EEG acquisition setup commonly used in speech processing studies. In fact, the dataset used in this study was originally collected to study cortical correlates of speech comprehension in Etard et al. 2019b. The proposed complex statistical modelling framework ($cTRF$) relates the fundamental waveform, which oscillates according to the speaker's pitch, to the high-frequency neural response. The neural response reflected by the model's coefficients exhibited low latency below 10 ms, which was similar to the previous results reported by Forte et al. 2017 & Maddox et al. 2018. By using a multichannel EEG system, the model revealed the topography of the response, which matched that of the evoked auditory brainstem response (Bidelman 2015; Grandori 1986; Ono et al. 1984). Furthermore, the method allows studying both cortical and subcortical responses within the same experiment due to the use of a conventional high-density EEG setup.

We applied the model to study the attentional modulation of the response in a two-talker *cocktail party*. We found that the neural tracking of the attended talker's fundamental waveform was stronger than that of the ignored talker, which is in agreement with findings reported in Forte et al. 2017. Interestingly, the top-down attentional effect exhibited itself not as a gain modulation but as a phase difference between the responses obtained for the attended and ignored voices. Furthermore, we built an efficient auditory attention decoding algorithm, which allowed us to rapidly identify the listener's attentional focus from their early high-frequency neural response to speech. The proposed decoder was capable of achieving comparable results to analogous methods based on cortical responses (Mirkovic et al. 2015; O'Sullivan et al. 2015), while using only three EEG channels. These results support the functional role of top-down attentional modulation of the early neural response at the fundamental frequency in the speech stream segregation. Moreover, combining the proposed rapid attention decoding approach with existing strategies, predominantly based on cortical responses, can contribute to reducing latency and improving the responsiveness of auditory BCIs.

In **chapter 5**, we extended the model introduced in chapter 4 (Etard et al. 2019a) to investigate whether the early neural response at the fundamental frequency is modulated by the acoustic and linguistic properties of different words from a continuous spoken narrative. In particular, we extended the original methodology to measure the strength of the word-level early neural responses at fundamental and then correlate it with acoustic and linguistic descriptors of each word. We found that the detected word-level neural responses were predominantly modulated by the acoustic features derived from the speaker's pitch. These results extend previous studies reporting that acoustic characteristics of different voices can significantly impact the neural responses at the fundamental frequency (Saiz-Alía et al. 2019, 2020; van Canneyt et al. 2021d). Here, we showed that these effects are also present for different words produced by the same speaker. We also found that the early neural response at the fundamental frequency is modulated by context-independent word frequency, but to a lesser extent, as compared to the acoustic features. This language-specific effect observed while controlling for acoustic modula-

tion suggests that early neural responses at the fundamental frequency can be modulated by efferent feedback from the higher-level cortical areas involved in language processing.

6.2 Future work

The work conducted in this thesis focused on developing computational models of neural mechanisms underlying natural speech processing, ranging from early, high-frequency responses, of predominantly subcortical origin, to slower cortical activity spanning across hundreds of milliseconds. We believe that following the current trend of using naturalistic, ecologically-valid stimuli to study neural speech processing, the importance of computational models characterizing mechanisms of speech perception will only increase. Although capable of uncovering neural dynamics of speech processing, the proposed models and methods have limitations and should be refined in future work.

In particular, the proposed tACS protocol allowed us to study the causal role of neural oscillations in speech in noise comprehension. The protocol could be used to study the role of neural oscillations in other speech-related mechanisms. In fact, Keshavarzi et al. 2021 used it to study the role of theta-band oscillations in stream segregation. Moreover, the protocol could be used to further study the causal role of delta-band activity, which is commonly reported as a neural correlate of speech processing in neuroimaging studies and is believed to be involved in language processing (Meyer 2018; Molinaro et al. 2018; Weissbart et al. 2020). Future experiments should, for instance, focus on studying language comprehension rather than speech-in-noise perception. Furthermore, our experimental setup could benefit from simultaneous neuroimaging, such as fMRI (as in Zoefel et al. 2018), to investigate whether the effect emerges directly from the modulation of the auditory cortex, or indirectly through activation of other networks.

Since the behaviour of the proposed spiking neural network model matched the effects of tACS on the modulation of speech in noise comprehension in the experimental studies, it could be used to perform model-based design and optimization of stimulation protocols. In particular, this could help improve the efficacy of TCS, commonly associated with small effect sizes and considerably large inter-subject variability (Guerra et al. 2020). Furthermore, the proposed biophysically-plausible model could be integrated with structural finite element models estimating the distribution of current flow in the brain (Datta et al. 2009; Huang et al. 2019a). Such a joint structural-functional model could be used to design personalized stimulation protocols. Analogously, neuroimaging could be used to further fine-tune the functional model to a particular participant (Kasten et al. 2019). For instance, M/EEG could be used to adjust intrinsic oscillations of the model to precisely match the participant's individual frequency of the theta-band activity (Zaehle et al. 2010).

The proposed spiking neural network model reliably predicted the online effects of tACS on ongoing neural oscillations. However, weak external currents could also facilitate plastic changes in the neural circuits. While, the effects of tACS are mostly associated with short-term

entrainment of neural oscillations to the phase of the external stimulation (Herrmann et al. 2016; Johnson et al. 2020; Krause et al. 2019; Krause et al. 2021; Vieira et al. 2020), some studies reported lasting after-effects of tACS (Moliadze et al. 2019; Rufener et al. 2016; Vossen et al. 2015). The model proposed in this thesis might be a suitable framework for studying the poorly understood neural basis underlying the effects of tACS, and for generating testable hypotheses useful for designing future experimental studies (Fröhlich et al. 2015).

It is important to note that, in its current form, the proposed spiking neural network model for cortical encoding of speech-in-noise reflects only bottom-up sensory encoding in the primary auditory cortex (Hyafil et al. 2015). As shown by the numerous experimental studies, including some in this thesis, feedback interaction between different brain regions is critical for speech comprehension. The future extensions of the model should focus on incorporating modulation of its activity through higher-level cognitive mechanisms such as selective attention, predictive coding or linguistic processing. While Hovsepyan et al. 2020 recently proposed the model for the role of neural oscillations in predictive coding, it diverged from the biologically-plausible implementation via spiking neural network. While the exact neural implementation of these top-down cognitive mechanisms is unknown, the model proposed here could be used to test different theories and hypotheses to seek similarities with experimental studies. For instance, recently, Kulkarni et al. 2021 extended the model introduced here to study the role of neural oscillation in audiovisual speech processing.

Unlike the proposed computational model for cortical speech encoding, the modelling framework for detecting early neural responses at the fundamental frequency of continuous speech focuses on investigating plausible top-down modulation through corticofugal pathways. The proposed methodology allowed us to decode selective attention using only short segments of neural data and to study acoustic and linguistic modulation of this response to words in the spoken narrative. However, the method did not inform us what exact neural mechanism produces the top-down observed effect. In other words, we could detect significant top-down modulation, but we do not know *how* it emerged. This problem is related to the recently debated neural origin of FFR (Coffey et al. 2019), which cannot be clearly determined using the proposed models.

The above-outlined limitation might be addressed in several ways. Firstly, since the proposed modelling method utilizes high-density EEG, it can be employed to simultaneously detect cortical and subcortical responses to continuous speech. In particular, this property can be applied to develop a methodology to study reciprocal modulation of cortical and subcortical responses, in a similar way to how Price et al. 2021 investigated directed connectivity between the primary auditory cortex and the brainstem. The development of such a method might be challenging due to the use of continuous speech in the experiment. Thus, the recently introduced method for *Enhanced Neural Tracking of the Fundamental Frequency of the Voice* (van Canneyt et al. 2021c), proposing extensive speech stimulus feature extraction for neural response modelling, might be useful for improving the detection of weak subcortical responses to speech.

Secondly, the mechanisms underlying the observed top-down modulation of early neural responses can be studied using biologically-plausible spiking neural network models. In particular, there exist many biologically-detailed models of neural circuitry in the auditory periphery and the brainstem (Verhulst et al. 2018; Zilany et al. 2014). Recently, Saiz-Alía et al. 2020 developed a model for predicting subcortical response to continuous speech. However, it implemented only bottom-up sensory encoding. Extending the model by incorporating plausible mechanisms of top-down cognitive modulation through corticofugal pathways could allow us to investigate what might drive the effects observed in the EEG experiments. Ideally, such a detailed biophysically-plausible model of the auditory periphery and the brainstem could replace the comparatively simple one included in the spiking neural network for cortical speech encoding introduced in this thesis. Together, the two would form a biologically-detailed model of the intact human auditory pathways suited for studying neural speech processing across different nuclei and timescales.

Appendix

The introduction chapter of this thesis reused several figures from publicly available sources. Three main chapters of this thesis were published as peer-reviewed scientific papers from which I am either the first or the second author making a significant contribution. Most of the reused items were published open access under a **CC-BY 4.0** license (Creative Commons Attribution 4.0 International License). The CC-BY 4.0 license allows for reuse provided the source is cited. Permission requests were sent in the case of materials not published under the CC-BY license. Table 6.1 summarises the source and copyright license of each reused item. Copies of permissions for reusing the items not published under the CC-BY license are presented in Fig. 6.1.

Figure / Chapter	**Journal**	**Publisher**	**Copyright License**	**Proofs**
Figure 1.1	Nature Reviews Endocrinology	Springer Nature		Yes
Figure 1.2	Current Opinion in Physiology	Elsevier		Yes
Figure 1.3	Journal of Cognitive Neuroscience	MIT Press	CC-BY 4.0	N/A
Figure 1.4	Nature Neuroscience	Springer Nature		Yes
Figure 1.5	Ear and Hearing	Wolters Kluwer		Yes
Figure 1.6	Trends in Cognitive Sciences	Elsevier		Yes
Figure 1.7	NeuroImage	Elsevier	CC-BY 4.0	N/A
Chapter 2	NeuroImage	Elsevier	CC-BY 4.0	N/A
Chapter 3	NeuroImage	Elsevier	CC-BY 4.0	N/A
Chapter 4	NeuroImage	Elsevier	CC-BY 4.0	N/A

Table 6.1: A summary of the sources and copyright license of items included in the thesis.

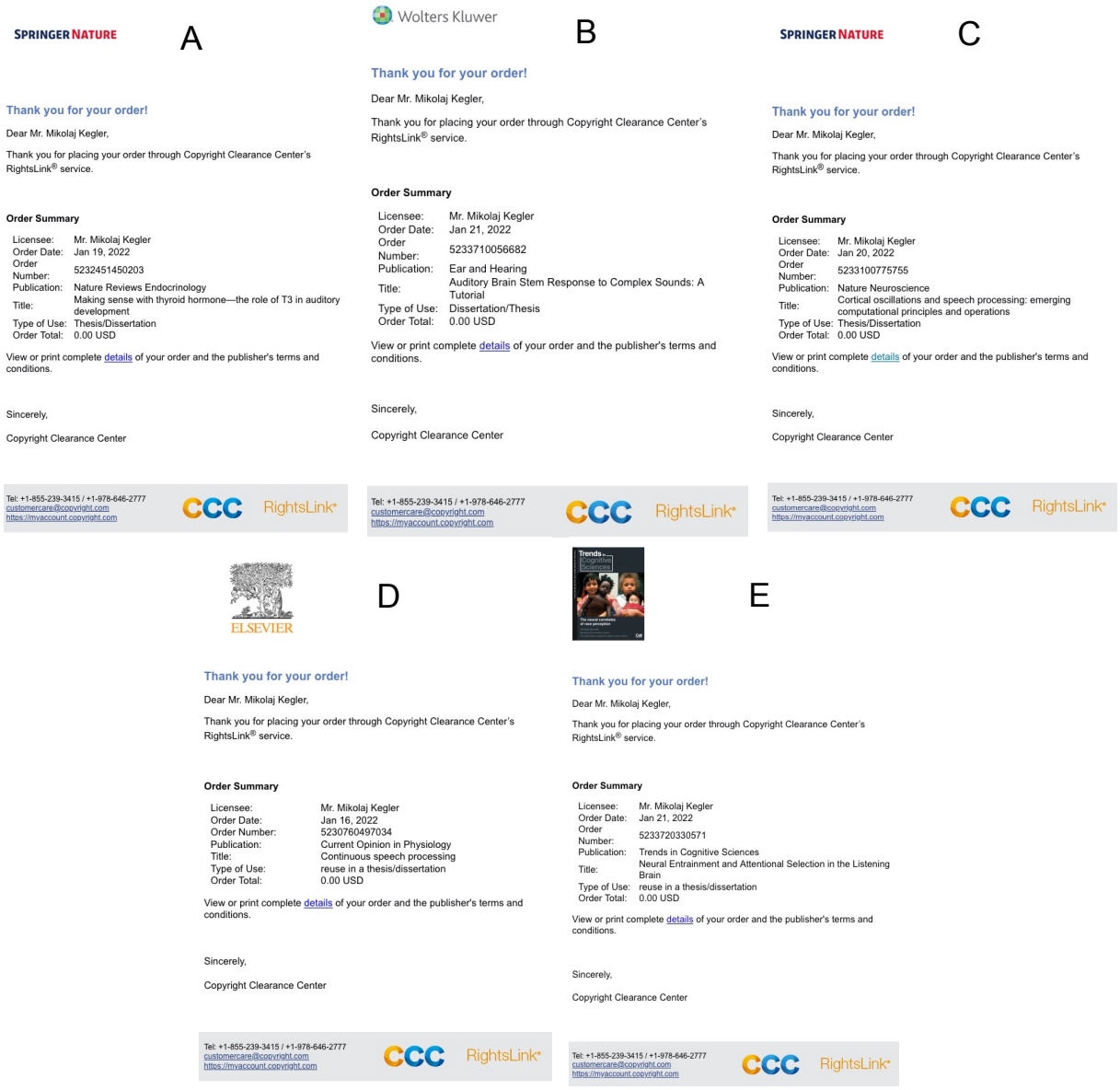

Figure 6.1: Proofs of permission. **A**: Fig. 1.1; **B**: Fig. 1.5; **C**: Fig. 1.4; **D**: Fig. 1.2; **E**: Fig. 1.6

Bibliography

Accou, B., Monesi, M. J., hamme, H. V., and Francart, T. (2021). Predicting speech intelligibility from EEG in a non-linear classification paradigm. *Journal of Neural Engineering* 18:066008.

Adam, V. and Hyafil, A. (2020). Non-linear regression models for behavioral and neural data analysis. *arXiv preprint arXiv:2002.00920*.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences* 98:13367–13372.

Ainsworth, M., Lee, S., Cunningham, M. O., Roopun, A. K., Traub, R. D., Kopell, N. J., and Whittington, M. A. (2011). Dual gamma rhythm generators control interlaminar synchrony in auditory cortex. *Journal of Neuroscience* 31:17040–17051.

Ali, M. M., Sellers, K. K., and Fröhlich, F. (2013). Transcranial alternating current stimulation modulates large-scale cortical network activity by network resonance. *Journal of Neuroscience* 33:11262–11275.

Alickovic, E., Lunner, T., Wendt, D., Fiedler, L., Hietkamp, R., Ng, E. H. N., and Graversen, C. (2020). Neural representation enhanced for speech and reduced for background noise with a hearing aid noise reduction scheme during a selective attention task. *Frontiers in neuroscience* 14:846.

Anderson, S. and Kraus, N. (2010). Sensory-cognitive interaction in the neural encoding of speech in noise: a review. *Journal of the American Academy of Audiology* 21:575–585.

Anderson, S., Parbery-Clark, A., White-Schwoch, T., and Kraus, N. (2012). Aging affects neural precision of speech encoding. *Journal of Neuroscience* 32:14156–14164.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics* 16:523–531.

Armstrong, M., Pegg, P., James, C., and Blamey, P. (1997). Speech perception in noise with implant and hearing aid. *The American journal of otology* 18:S140–1.

Asamoah, B., Khatoun, A., and Mc Laughlin, M. (2019). Analytical bias accounts for some of the reported effects of tACS on auditory perception. *Brain stimulation* 12:1001–1009.

Ashida, G. and Carr, C. E. (2011). Sound localization: Jeffress and beyond. *Current opinion in neurobiology* 21:745–751.

Baayen, R. H. (2001). *Word frequency distributions*. Vol. 18. Springer Science & Business Media.

Bachmann, F. L., MacDonald, E. N., and Hjortkjær, J. (2021). Neural Measures of Pitch Processing in EEG Responses to Running Speech. *Frontiers in Neuroscience* 15:738408.

Bachmann, F. L., MacDonald, E. N., and Hjortkjær, J. (2020). Subcortical responses to continuous speech: A comparison of two computing methods. *Advances and Perspectives in Auditory Neuroscience 2020*.

Baltus, A., Wagner, S., Wolters, C. H., and Herrmann, C. S. (2018). Optimized auditory transcranial alternating current stimulation improves individual auditory temporal resolution. *Brain stimulation* 11:118–124.

Bastiaansen, M. and Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. *Progress in Brain Research* 159:179–196.

Beckmann, P., Kegler, M., and Cernak, M. (2021). Word-Level Embeddings for Cross-Task Transfer Learning in Speech Processing. *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 446–450.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*:1165–1188.

Berens, P. (2009). CircStat: a MATLAB toolbox for circular statistics. *Journal of statistical software* 31:1–21.

Bergmann, T. O. and Hartwigsen, G. (2021). Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *Journal of cognitive neuroscience* 33:195–225.

Bestmann, S., de Berker, A. O., and Bonaiuto, J. (2015). Understanding the behavioural consequences of noninvasive brain stimulation. *Trends in cognitive sciences* 19:13–20.

Beysolow II, T. (2018). *Applied natural language processing with python*. Springer.

Bidelman, G. M. (2015). Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. *Hearing research* 323:68–80.

Bidelman, G. M. (2018). Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. *Neuroimage* 175:56–69.

Bidelman, G. M., Gandour, J. T., and Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience* 23:425–434.

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2016). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25:402–412.

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25:402–412.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bleichner, M. G., Mirkovic, B., and Debener, S. (2016). Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. *Journal of neural engineering* 13:066004.

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International* 5:341–345.

Bonaiuto, J. J. and Bestmann, S. (2015). Understanding the nonlinear physiological and behavioral effects of tDCS through computational neurostimulation. *Progress in Brain Research* 222:75–103.

Borgmann, C., Roß, B., Draganova, R., and Pantev, C. (2001). Human auditory middle latency responses: influence of stimulus type and intensity. *Hearing research* 158:57–64.

Boto, E., Holmes, N., Leggett, J., Roberts, G., Shah, V., Meyer, S. S., Muñoz, L. D., Mullinger, K. J., Tierney, T. M., Bestmann, S., et al. (2018). Moving magnetoencephalography towards real-world applications with a wearable system. *Nature* 555:657–661.

Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in statistics*. Elsevier, 201–236.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., and Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language* 120:163–173.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., and Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language* 157:81–94.

Brix, R (1984). The influence of attention on the auditory brain stem evoked responses preliminary report. *Acta oto-laryngologica* 98:89–92.

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A., Resnik, P., Simon, J. Z., and Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife* 11:e72056.

Brodbeck, C., Hong, L. E., and Simon, J. Z. (2018a). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology* 28:3976–3983.

Brodbeck, C., Jiao, A., Hong, L. E., and Simon, J. Z. (2020a). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLoS biology* 18:e3000883.

Brodbeck, C., Presacco, A., Anderson, S., and Simon, J. Z. (2018b). Over-representation of speech in older adults originates from early response in higher order auditory cortex. *Acta Acustica united with Acustica* 104:774–777.

Brodbeck, C., Presacco, A., and Simon, J. Z. (2018c). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *Neuroimage* 172:162–174.

Brodbeck, C. and Simon, J. Z. (2020b). Continuous speech processing. *Current Opinion in Physiology* 18:25–31.

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

Broderick, M. P., Anderson, A. J., and Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience* 39:7564–7575.

Brosch, M., Budinger, E., and Scheich, H. (2002). Stimulus-related gamma oscillations in primate auditory cortex. *Journal of neurophysiology* 87:2715–2725.

Buxton, R. B. (2013). The physics of functional magnetic resonance imaging (fMRI). *Reports on Progress in Physics* 76:096601.

Cakan, C. and Obermayer, K. (2020). Biophysically grounded mean-field models of neural populations under electrical stimulation. *PLoS computational biology* 16:e1007822.

Cardin, J. A., Carlén, M., Meletis, K., Knoblich, U., Zhang, F., Deisseroth, K., Tsai, L.-H., and Moore, C. I. (2009). Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* 459:663–667.

Chandrasekaran, B. and Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology* 47:236–246.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25:975–979.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America* 118:887–906.

Chung, K. (2004). Challenges and recent developments in hearing aids: Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends in Amplification* 8:83–124.

Clarke, J., Gaudrain, E., Chatterjee, M., and Başkent, D. (2014). T'ain't the way you say it, it's what you say–Perceptual continuity of voice and top–down restoration of speech. *Hearing Research* 315:80–87.

Coffey, E. B., Herholz, S. C., Chepesiuk, A. M., Baillet, S., and Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature communications* 7:1–11.

Coffey, E. B., Musacchia, G., and Zatorre, R. J. (2017). Cortical correlates of the auditory frequency-following and onset responses: EEG and fMRI evidence. *Journal of Neuroscience* 37:830–838.

Coffey, E. B., Nicol, T., White-Schwoch, T., Chandrasekaran, B., Krizman, J., Skoe, E., Zatorre, R. J., and Kraus, N. (2019). Evolving perspectives on the sources of the frequency-following response. *Nature Communications* 10:1–10.

Combrisson, E. and Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods* 250:126–136.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience* 10:604.

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., and Lalor, E. C. (2021). Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research. *Frontiers in Neuroscience* 15.

Dale, A. M. and Sereno, M. I. (1993). Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of cognitive neuroscience* 5:162–176.

Das, N., Zegers, J., hamme, H. V., Francart, T., and Bertrand, A. (2020). Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding. *Journal of Neural Engineering* 17:046039.

Datta, A., Bansal, V., Diaz, J., Patel, J., Reato, D., and Bikson, M. (2009). Gyri-precise head model of transcranial direct current stimulation: improved spatial focality using a ring electrode versus conventional rectangular pad. *Brain stimulation* 2:201–207.

Daube, C., Ince, R. A., and Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology* 29:1924–1937.

David, S. V., Mesgarani, N., and Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in neural systems* 18:191–212.

Davidesco, I., Matuk, C., Bevilacqua, D., Poeppel, D., and Dikker, S. (2021). Neuroscience Research in the Classroom: Portable Brain Technologies in Education Research. *Educational Researcher* 50:649–656.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. 1. Cambridge university press.

Dayan, E., Censor, N., Buch, E. R., Sandrini, M., and Cohen, L. G. (2013). Noninvasive brain stimulation: from physiology to network dynamics and back. *Nature neuroscience* 16:838–844.

De Cheveigné, A., Wong, D. D., Di Liberto, G. M., Hjortkjaer, J., Slaney, M., and Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage* 172:206–216.

De Cheveigne, A. (2005). Pitch perception models. *Pitch*. Springer, 169–233.

Debener, S., Minow, F., Emkes, R., Gandras, K., and De Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology* 49:1617–1621.

Decruy, L., Vanthornhout, J., and Francart, T. (2019). Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of neurophysiology* 122:601–615.

Decruy, L., Vanthornhout, J., and Francart, T. (2020). Hearing impairment is associated with enhanced neural tracking of the speech envelope. *Hearing Research* 393:107961.

Devlin, J. T. and Watkins, K. E. (2007). Stimulating language: insights from TMS. *Brain* 130:610–622.

Di Liberto, G. M., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S. A., and Mesgarani, N. (2021). Neural representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage* 227:117586.

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology* 25:2457–2465.

Dilley, L. C. and Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21:1664–1670.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience* 19:158–164.

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

Ding, N. and Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience* 33:5728–5735.

Ding, N. and Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience* 8:311.

Doelling, K. B. and Assaneo, M. F. (2021). Neural oscillations are a start toward understanding brain activity rather than the end. *PLoS Biology* 19:e3001234.

Donhauser, P. W. and Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron* 105:385–393.

Driscoll, W. C. (1996). Robustness of the ANOVA and Tukey-Kramer statistical tests. *Computers & Industrial Engineering* 31:265–268.

Drullman, R. (1995). Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level. *The Journal of the Acoustical Society of America* 98:1796–1798.

Elbanna, G., Biryukov, A., Scheidwasser-Clow, N., Orlandic, L., Mainar, P., Kegler, M., Beckmann, P., and Cernak, M. (2022). Hybrid Handcrafted and Learnable Audio Representation for Analysis of Speech Under Cognitive and Physical Load. *arXiv preprint arXiv:2203.16637*.

Elhilali, M. and Shamma, S. A. (2008). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America* 124:3751–3771.

Engel, A. K., Moll, C. K., Fried, I., and Ojemann, G. A. (2005). Invasive recordings from the human brain: clinical insights and beyond. *Nature Reviews Neuroscience* 6:35–47.

Erkens, J., Schulte, M., Vormann, M., and Herrmann, C. S. (2020). Lacking effects of envelope transcranial alternating current stimulation indicate the need to revise envelope transcranial alternating current stimulation methods. *Neuroscience Insights* 15:2633105520936623.

Erkens, J., Schulte, M., Vormann, M., Wilsch, A., and Herrmann, C. S. (2021). Hearing Impaired Participants Improve More Under Envelope-Transcranial Alternating Current Stimulation When Signal to Noise Ratio Is High. *Neuroscience insights* 16:2633105520988854.

Etard, O., Kegler, M., Braiman, C., Forte, A. E., and Reichenbach, T. (2019a). Decoding of selective attention to continuous speech from the human auditory brainstem response. *NeuroImage* 200:1–11.

Etard, O., Messaoud, R. B., Gaugain, G., and Reichenbach, T. (2021). No Evidence of Attentional Modulation of the Neural Response to the Temporal Fine Structure of Continuous Musical Pieces. *Journal of Cognitive Neuroscience*:1–14.

Etard, O. and Reichenbach, T. (2019b). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience* 39:5750–5759.

Fant, G. (1970). *Acoustic theory of speech production*. 2. Walter de Gruyter.

Farahani, F., Kronberg, G., FallahRad, M., Oviedo, H. V., and Parra, L. C. (2021). Effects of direct current stimulation on synaptic plasticity in a single neuron. *Brain Stimulation* 14:588–597.

Fawcett, J., Risko, E., and Kingstone, A. (2015). *The handbook of attention*. MIT Press.

Ferrari, M. and Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63:921–935.

Fiedler, L., Ala, T. S., Graversen, C., Alickovic, E., Lunner, T., and Wendt, D. (2021). Hearing Aid Noise Reduction Lowers the Sustained Listening Effort During Continuous Speech in Noise—A Combined Pupillometry and EEG Study. *Ear and hearing* 42:1590–1601.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., and Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of neural engineering* 14:036020.

Fiedler, L., Wöstmann, M., Herbst, S. K., and Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* 186:33–42.

Fitch, R. H., Miller, S., and Tallal, P. (1997). Neurobiology of speech perception. *Annual review of neuroscience* 20:331–353.

Font-Alaminos, M., Ribas-Prats, T., Gorina-Careta, N., and Escera, C. (2021). Emergence of prediction error along the human auditory hierarchy. *Hearing Research* 399:107954.

Forte, A. E., Etard, O., and Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *elife* 6:e27203.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences* 6:78–84.

Friederici, A. D., Pfeifer, E., and Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research* 1:183–192.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Frisina, D. R. and Frisina, R. D. (1997). Speech recognition in noise and presbycusis: relations to possible neural mechanisms. *Hearing research* 106:95–104.

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007). Auditory attention—focusing the searchlight on sound. *Current opinion in neurobiology* 17:437–455.

Fröhlich, F. (2015). Experiments and models of cortical oscillations as a target for noninvasive brain stimulation. *Progress in brain research* 222:41–73.

Fröhlich, F. and McCormick, D. A. (2010). Endogenous electric fields may guide neocortical network activity. *Neuron* 67:129–143.

Frohlich, F. and Schmidt, S. L. (2013). Rational design of transcranial current stimulation (TCS) through mechanistic insights into cortical network dynamics. *Frontiers in human neuroscience* 7:804.

Fröhlich, F., Sellers, K. K., and Cordle, A. L. (2015). Targeting the neurophysiology of cognitive systems with transcranial alternating current stimulation. *Expert review of neurotherapeutics* 15:145–167.

Fuglsang, S. A., Märcher-Rørsted, J., Dau, T., and Hjortkjær, J. (2020). Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *Journal of Neuroscience* 40:2562–2572.

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

Galbraith, G. C., Olfman, D. M., and Huffman, T. M. (2003). Selective attention affects human brain stem frequency-following response. *Neuroreport* 14:735–738.

Gannot, S., Burshtein, D., and Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing* 49:1614–1626.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n* 93:27403.

Geirnaert, S., Francart, T., and Bertrand, A. (2021a). Unsupervised Self-Adaptive Auditory Attention Decoding. *IEEE Journal of Biomedical and Health Informatics*.

Geirnaert, S., Francart, T., and Bertrand, A. (2022). Time-adaptive Unsupervised Auditory Attention Decoding Using EEG-based Stimulus Reconstruction. *bioRxiv*.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigne, A., Lalor, E., Meyer, B. T., Miran, S., Francart, T., and Bertrand, A. (2021b). Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine* 38:89–102.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in psychology* 2:130.

Giacino, J., Ashwal, S., Childs, N., Cranford, R., Jennett, B., Katz, D., Kelly, J., Rosenberg, J., Whyte, J., Zafonte, R., and Zasler, N. (2002). The minimally conscious state: definition and diagnostic criteria. *Neurology* 58:349–353.

Gillis, M., Decruy, L., Vanthornhout, J., and Francart, T. (2021a). Hearing loss is associated with delayed neural responses to continuous speech. *bioRxiv*.

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., and Brodbeck, C. (2021b). Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics. 41:10316–10329.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., and Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* 56:1127–1134.

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience* 15:511–517.

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

Golumbic, E. M. Z., Poeppel, D., and Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain and language* 122:151–161.

Gorina-Careta, N., Kurkela, J. L., Hämäläinen, J., Astikainen, P., and Escera, C. (2021). Neural generators of the frequency-following response elicited to stimuli of low and high frequency: A magnetoencephalographic (MEG) study. *NeuroImage* 231:117866.

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

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience* 7:267.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86:446–460.

Grandori, F. (1986). Field analysis of auditory evoked brainstem potentials. *Hearing research* 21:51–58.

Grant, K. W., Ardell, L. H., Kuhl, P. K., and Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *The Journal of the Acoustical society of America* 77:671–677.

Griffiths, T. D. and Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience* 5:887–892.

Groppe, D. M., Urbach, T. P., and Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology* 48:1711–1725.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS biology* 11:e1001752.

Guerra, A., López-Alonso, V., Cheeran, B., and Suppa, A. (2020). Variability in non-invasive brain stimulation studies: reasons and results. *Neuroscience letters* 719:133330.

Hallett, M. (2007). Transcranial magnetic stimulation: a primer. *Neuron* 55:187–199.

Hamilton, L. S., Edwards, E., and Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology* 28:1860–1871.

Han, C., O'Sullivan, J., Luo, Y., Herrero, J., Mehta, A. D., and Mesgarani, N. (2019). Speaker-independent auditory attention decoding without access to clean speech sources. *Science advances* 5:eaav6134.

Hansen, J. C., Dickstein, P. W., Berka, C., and Hillyard, S. A. (1983). Event-related potentials during selective attention to speech sounds. *Biological Psychology* 16:211–224.

Hari, R. and Puce, A. (2017). *MEG-EEG Primer*. Oxford University Press.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature* 585:357–362.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87:96–110.

Hausfeld, L., Riecke, L., Valente, G., and Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage* 181:617–626.

Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. (2019). A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation. *The Journal of the Acoustical Society of America* 145:1378–1388.

Heimrath, K., Fiene, M., Rufener, K. S., and Zaehle, T. (2016). Modulating human auditory processing by transcranial electrical stimulation. *Frontiers in cellular neuroscience* 10:53.

Helfrich, R. F., Schneider, T. R., Rach, S., Trautmann-Lengsfeld, S. A., Engel, A. K., and Herrmann, C. S. (2014). Entrainment of brain oscillations by transcranial alternating current stimulation. *Current biology* 24:333–339.

Herrmann, C. S., Murray, M. M., Ionta, S., Hutt, A., and Lefebvre, J. (2016). Shaping intrinsic neural oscillations with periodic stimulation. *Journal of Neuroscience* 36:5328–5337.

Herrmann, C. S., Rach, S., Neuling, T., and Strüber, D. (2013). Transcranial alternating current stimulation: a review of the underlying mechanisms and modulation of cognitive processes. *Frontiers in human neuroscience* 7:279.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience* 8:393–402.

Hölle, D., Meekes, J., and Bleichner, M. G. (2021). Mobile ear-EEG to study auditory attention in everyday life. *Behavior Research Methods*:1–12.

Horton, C., D'Zmura, M., and Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of neurophysiology* 109:3082–3093.

Hovsepyan, S., Olasagasti, I., and Giraud, A.-L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature communications* 11:1–12.

Huang, H. and Pan, J. (2006). Speech pitch determination based on Hilbert-Huang transform. *Signal Processing* 86:792–803.

Huang, Y., Datta, A., Bikson, M., and Parra, L. C. (2019a). Realistic volumetric-approach to simulate transcranial electric stimulation—ROAST—a fully automated open-source pipeline. *Journal of neural engineering* 16:056006.

Huang, Y. and Parra, L. C. (2019b). Can transcranial electric stimulation with multiple electrodes reach deep targets? *Brain stimulation* 12:30–40.

Huffman, R. F. and Henson Jr, O. (1990). The descending auditory pathway and acousticomotor systems: connections with the inferior colliculus. *Brain Research Reviews* 15:295–323.

Hutcherson, R. W., Dirks, D. D., and Morgan, D. E. (1979). Evaluation of the speech perception in noise (SPIN) test. *Otolaryngology–Head and Neck Surgery* 87:239–245.

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., and Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *Elife* 4:e06213.

İnce, R., Adanır, S. S., and Sevmez, F. (2020). The inventor of electroencephalography (EEG): Hans Berger (1873–1941). *Child's Nervous System*:1–2.

Iotzov, I. and Parra, L. C. (2019). EEG can predict speech intelligibility. *Journal of neural engineering* 16:036008.

Irvine, D. R. (2012). The auditory brainstem: a review of the structure and function of auditory brainstem processing mechanisms.

Jadi, M. P. and Sejnowski, T. J. (2014). Cortical oscillations arise from contextual interactions that regulate sparse coding. *Proceedings of the National Academy of Sciences* 111:6780–6785.

Jadoul, Y., Thompson, B., and De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71:1–15.

Jensen, O. and Colgin, L. L. (2007). Cross-frequency coupling between neuronal oscillations. *Trends in cognitive sciences* 11:267–269.

Johnson, L., Alekseichuk, I., Krieg, J., Doyle, A., Yu, Y., Vitek, J., Johnson, M., and Opitz, A. (2020). Dose-dependent effects of transcranial alternating current stimulation on spike timing in awake nonhuman primates. *Science advances* 6:eaaz2747.

Joris, P. X. and Verschooten, E. (2013). On the limit of neural phase locking to fine structure in humans. *Basic aspects of hearing*:101–108.

Junqua, J.-C., Fincke, S., and Field, K. (1999). The Lombard effect: A reflex to better communicate with others in noise. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. Vol. 4. IEEE, 2083–2086.

Kadir, S., Kaza, C., Weissbart, H., and Reichenbach, T. (2019). Modulation of speech-in-noise comprehension through transcranial current stimulation with the phase-shifted speech envelope. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28:23–31.

Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics* 63:1377–1388.

Kasten, F. H., Duecker, K., Maack, M. C., Meiser, A., and Herrmann, C. S. (2019). Integrating electric field modeling and neuroimaging to explain inter-individual variability of tACS effects. *Nature communications* 10:1–11.

Kayser, C., Ince, R. A. A., and Panzeri, S. (Oct. 2012). Analysis of Slow (Theta) Oscillations as a Potential Temporal Reference Frame for Information Coding in Sensory Cortices. *PLoS Computational Biology* 8. Ed. by T. Behrens:e1002717.

Kegler, M., Beckmann, P., and Cernak, M. (2020). Deep Speech Inpainting of Time-Frequency Masks. *Proc. Interspeech 2020*, 3276–3280.

Kegler, M. and Reichenbach, T. (2021). Modelling the effects of transcranial alternating current stimulation on the neural encoding of speech in noise. *NeuroImage* 224:117427.

Kerlin, J. R., Shahin, A. J., and Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *Journal of Neuroscience* 30:620–628.

Keshavarzi, M., Kegler, M., Kadir, S., and Reichenbach, T. (2020a). Transcranial alternating current stimulation in the theta band but not in the delta band modulates the comprehension of naturalistic speech in noise. *NeuroImage* 210:116557.

Keshavarzi, M. and Reichenbach, T. (2020b). Transcranial alternating current stimulation with the theta-band portion of the temporally-aligned speech envelope improves speech-in-noise comprehension. *Frontiers in human neuroscience* 14:187.

Keshavarzi, M., Varano, E., and Reichenbach, T. (2021). Cortical tracking of a background speaker modulates the comprehension of a foreground speech signal. *Journal of Neuroscience* 41:5093–5101.

Keshishian, M., Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. (2020). Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *Elife* 9:e53445.

Kidd Jr, G., Mason, C. R., Best, V., and Swaminathan, J. (2015). Benefits of acoustic beamforming for solving the cocktail party problem. *Trends in Hearing* 19:2331216515593385.

Kielar, A., Meltzer, J. A., Moreno, S., Alain, C., and Bialystok, E. (2014). Oscillatory responses to semantic and syntactic violations. *Journal of Cognitive Neuroscience* 26:2840–2862.

Klobassa, D. S., Vaughan, T. M., Brunner, P., Schwartz, N., Wolpaw, J. R., Neuper, C., and Sellers, E. (2009). Toward a high-throughput auditory P300-based brain–computer interface. *Clinical Neurophysiology* 120:1252–1261.

Klonowski, W. (2009). Everything you wanted to ask about EEG but were afraid to get the right answer. *Nonlinear biomedical physics* 3:1–5.

Kluender, K. R., Lotto, A. J., and Holt, L. L. (2012). Contributions of nonhuman animal models to understanding human speech perception. *Listening to Speech*. Psychology Press, 203–220.

Kollmeier, B., Gilkey, R. H., and Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *The Journal of the Acoustical Society of America* 83:1852–1862.

Kösem, A. and Van Wassenhove, V. (2017). Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience* 32:536–544.

Koskinen, M., Kurimo, M., Gross, J., Hyvärinen, A., and Hari, R. (2020). Brain activity reflects the predictability of word sequences in listened continuous speech. *NeuroImage* 219:116936.

Kraus, N., Anderson, S., and White-Schwoch, T. (2017). *The Frequency-Following Response: A Window Into Human Communication*. Springer.

Krause, M. R., Vieira, P. G., Csorba, B. A., Pilly, P. K., and Pack, C. C. (2019). Transcranial alternating current stimulation entrains single-neuron activity in the primate brain. *Proceedings of the National Academy of Sciences* 116:5747–5755.

Krause, M. R., Vieira, P. G., Thivierge, J.-P., and Pack, C. C. (2021). tACS competes with ongoing oscillations for control of spike-timing in the primate brain. *bioRxiv*.

Krishnan, A., Gandour, J. T., and Bidelman, G. M. (2010). The effects of tone language experience on pitch processing in the brainstem. *Journal of Neurolinguistics* 23:81–95.

Krizman, J. and Kraus, N. (2019). Analyzing the FFR: A tutorial for decoding the richness of auditory function. *Hearing Research* 382:107779.

Kronberg, G., Rahman, A., Sharma, M., Bikson, M., and Parra, L. C. (2020). Direct current stimulation boosts hebbian plasticity in vitro. *Brain stimulation* 13:287–301.

Kubanek, J. (2018). Neuromodulation with transcranial focused ultrasound. *Neurosurgical focus* 44:E14.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., and Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PloS one* 8:e53398.

Kübler, A., Furdea, A., Halder, S., Hammer, E. M., Nijboer, F., and Kotchoubey, B. (2009). A brain–computer interface controlled auditory event-related potential (P300) spelling system for locked-in patients. *Annals of the New York Academy of Sciences* 1157:90–100.

Kulasingham, J. P., Brodbeck, C., Presacco, A., Kuchinsky, S. E., Anderson, S., and Simon, J. Z. (2020). High gamma cortical processing of continuous speech in younger and older listeners. *Neuroimage* 222:117291.

Kulasingham, J. P. and Simon, J. Z. (2022). Algorithms for Estimating Time-Locked Neural Response Components in Cortical Processing of Continuous Speech. *bioRxiv*.

Kulkarni, A., Kegler, M., and Reichenbach, T. (2021). Effect of visual input on syllable parsing in a computational model of a neural microcircuit for speech processing. *Journal of Neural Engineering* 18:056055.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics* 22:79–86.

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* 62:621–647.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307:161–163.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., and Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron* 77:750–761.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., and Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of neurophysiology* 94:1904–1911.

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

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human perception and performance* 21:451.

Lehmann, A. and Schönwiesner, M. (2014). Selective attention modulates human auditory brainstem responses: relative contributions of frequency and spatial cues. *PloS one* 9:e85442.

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., and Francart, T. (2019a). Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations. *Hearing research* 380:1–9.

Lesenfants, D., Vanthornhout, J., Verschueren, E., and Francart, T. (2019b). Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech. *Journal of neural engineering* 16:066017.

Lewis, A. G. and Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex* 68:155–168.

Lewis, M. (2007). *Stepwise versus Hierarchical Regression: Pros and Cons.* `https://eric.ed.gov/?id=ED534385`.

Liegeois-Chauvel, C, Musolino, A, Badier, J., Marquis, P, and Chauvel, P (1994). Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* 92:204–214.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. *2008 eighth ieee international conference on data mining*. IEEE, 413–422.

Luo, H. and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54:1001–1010.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27:1256–1266.

Maddox, R. K. and Lee, A. K. (2018). Auditory brainstem responses to continuous natural speech in human listeners. *Eneuro* 5.

Mahoney, M. (2011). *Large text compression benchmark*. `www.mattmahoney.net/dc/text.html`.

Maier, W. and Ruf, I. (2016). Evolution of the mammalian middle ear: a historical review. *Journal of anatomy* 228:270–283.

Maison, S. F. and Liberman, M. C. (2000). Predicting vulnerability to acoustic injury with a noninvasive assay of olivocochlear reflex strength. *Journal of Neuroscience* 20:4701–4707.

Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004). Mining event-related brain dynamics. *Trends in cognitive sciences* 8:204–210.

Maoiléidigh, D. Ó. and Ricci, A. J. (2019). A bundle of mechanisms: Inner-ear hair-cell mechanotransduction. *Trends in neurosciences* 42:221–236.

Marković, D., Mizrahi, A., Querlioz, D., and Grollier, J. (2020). Physics for neuromorphic computing. *Nature Reviews Physics* 2:499–510.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* 11:431–441.

Mazzoni, A., Panzeri, S., Logothetis, N. K., and Brunel, N. (2008). Encoding of naturalistic stimuli by local field potential spectra in networks of excitatory and inhibitory neurons. *PLoS computational biology* 4:e1000239.

Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences* 111:6792–6797.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience* 48:2609–2621.

Middlebrooks, J. C., Simon, J. Z., Popper, A. N., and Fay, R. R. (2017). *The auditory system at the cocktail party*. Vol. 60. Springer.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5528–5531.

Miller, G. A. and Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior* 2:217–228.

Miniussi, C., Harris, J. A., and Ruzzoli, M. (2013). Modelling non-invasive brain stimulation in cognitive neuroscience. *Neuroscience & Biobehavioral Reviews* 37:1702–1712.

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., and Babadi, B. (2018). Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach. *Frontiers in neuroscience* 12:262.

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

Moliadze, V., Sierau, L., Lyzhko, E., Stenner, T., Werchowski, M., Siniatchkin, M., and Hartwigsen, G. (2019). After-effects of 10 Hz tACS over the prefrontal cortex on phonological word decisions. *Brain stimulation* 12:1464–1474.

Molinaro, N. and Lizarazu, M. (2018). Delta (but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience* 48:2642–2650.

Morillon, B., Liégeois-Chauvel, C., Arnal, L. H., Bénar, C. G., and Giraud, A.-L. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. *Frontiers in psychology* 3:248.

Negahbani, E., Kasten, F. H., Herrmann, C. S., and Fröhlich, F. (2018). Targeting alpha-band oscillations in a cortical model with amplitude-modulated high-frequency transcranial electric stimulation. *Neuroimage* 173:3–12.

Neupane, A. K., Gururaj, K., Mehta, G., and Sinha, S. K. (2014). Effect of repetition rate on speech evoked auditory brainstem response in younger and middle aged individuals. *Audiology research* 4:21–27.

Ng, L., Kelley, M. W., and Forrest, D. (2013). Making sense with thyroid hormone—The role of T 3 in auditory development. *Nature Reviews Endocrinology* 9:296–307.

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., and Kashino, K. (2021). BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation. *arXiv preprint arXiv:2103.06695*.

Nijboer, F., Furdea, A., Gunst, I., Mellinger, J., McFarland, D. J., Birbaumer, N., and Kübler, A. (2008). An auditory brain–computer interface (BCI). *Journal of neuroscience methods* 167:43–50.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America* 95:1085–1099.

Norrix, L. W. and Glattke, T. J. (1996). Multichannel waveforms and topographic mapping of the auditory brainstem response under common stimulus and recording conditions. *Journal of communication disorders* 29:157–182.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., and Brugge, J. F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *Journal of Neuroscience* 29:15564–15574.

Noury, N., Hipp, J. F., and Siegel, M. (2016). Physiological processes non-linearly affect electrophysiological recordings during transcranial electric stimulation. *Neuroimage* 140:99–109.

Obleser, J. and Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences* 23:913–926.

Ono, I., Ichikawa, G., Yosikawa, H., Kato, E., and Fukuda, M. (1984). The Scalp Topography of ABR Stability· Reappearance. *AUDIOLOGY JAPAN* 27:292–299.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral cortex* 25:1697–1706.

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

Parkin, B. L., Ekhtiari, H., and Walsh, V. F. (2015). Non-invasive human brain stimulation in cognitive neuroscience: a primer. *Neuron* 87:932–945.

Parras, G. G., Nieto-Diego, J., Carbajal, G. V., Valdés-Baizabal, C., Escera, C., and Malmierca, M. S. (2017). Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nature Communications* 8:1–17.

Patten, W. (1910). *International Short Stories (Vol. 2)*. P.F. Collier & Son.

Paulus, W. (2011). Transcranial electrical stimulation (tES–tDCS; tRNS, tACS) methods. *Neuropsychological rehabilitation* 21:602–617.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12:2825–2830.

Peelle, J. E., Johnsrude, I., and Davis, M. H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in Human Neuroscience* 4:51.

Peelle, J. E. and Wingfield, A. (2016). The neural consequences of age-related hearing loss. *Trends in neurosciences* 39:486–497.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*:539–558.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Piastra, M. C., Nüßing, A., Vorwerk, J., Clerc, M., Engwer, C., and Wolters, C. H. (2021). A comprehensive study on electroencephalography and magnetoencephalography sensitivity to cortical and subcortical sources. *Human Brain Mapping* 42:978–992.

Piccione, F., Giorgi, F., Tonin, P, Priftis, K, Giove, S, Silvoni, S, Palmas, G, and Beverina, F (2006). P300-based brain computer interface: reliability and performance in healthy and paralysed participants. *Clinical neurophysiology* 117:531–537.

Pickles, J. (1998). *An introduction to the physiology of hearing*. Brill.

Picton, T. W. (1992). The P300 wave of the human event-related potential. *Journal of clinical neurophysiology* 9:456–479.

Picton, T. W. and Hillyard, S. A. (1974). Human auditory evoked potentials. II: Effects of attention. *Electroencephalography and clinical neurophysiology* 36:191–200.

Pikovsky, A., Rosenblum, M., Kurths, J., and Synchronization, A (2001). A universal concept in nonlinear sciences. *Self* 2:3.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454:995–999.

Plomp, R. and Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 18:43–52.

Polonenko, M. J. and Maddox, R. K. (2021). Exposing distinct subcortical components of the auditory brainstem response evoked by continuous naturalistic speech. *eLife* 10:e62329.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience* 35:1497–1503.

Presacco, A., Simon, J. Z., and Anderson, S. (2016). Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. *Journal of neurophysiology* 116:2346–2355.

Price, C. N. and Bidelman, G. M. (2021). Attention reinforces human corticofugal system to aid speech perception in noise. *NeuroImage* 235:118014.

Radman, T., Ramos, R. L., Brumberg, J. C., and Bikson, M. (2009). Role of cortical cell type and morphology in subthreshold and suprathreshold uniform electric field stimulation in vitro. *Brain stimulation* 2:215–228.

Ray, S. and Maunsell, J. H. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS biology* 9:e1000610.

Reato, D., Rahman, A., Bikson, M., and Parra, L. C. (2010). Low-intensity electrical stimulation affects network dynamics by modulating population rate and spike timing. *Journal of Neuroscience* 30:15067–15079.

Reato, D., Rahman, A., Bikson, M., and Parra, L. C. (2013). Effects of weak transcranial alternating current stimulation on brain activity—a review of known mechanisms from animal studies. *Frontiers in human neuroscience* 7:687.

Reichenbach, C. S., Braiman, C., Schiff, N. D., Hudspeth, A., and Reichenbach, T. (2016). The auditory-brainstem response to continuous, non-repetitive speech is modulated by the speech envelope and reflects speech processing. *Frontiers in computational neuroscience* 10:47.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994). On the perceptual organization of speech. *Psychological review* 101:129.

Riecke, L., Formisano, E., Sorger, B., Başkent, D., and Gaudrain, E. (2018). Neural entrainment to speech modulates speech intelligibility. *Current Biology* 28:161–169.

Ritter, P. and Villringer, A. (2006). simultaneous EEG–fMRI. *Neuroscience & Biobehavioral Reviews* 30:823–838.

Rubin, D. C. (1976). The effectiveness of context before, after, and around a missing word. *Perception & Psychophysics* 19:214–216.

Rufener, K. S., Zaehle, T., Oechslin, M. S., and Meyer, M. (2016). 40 Hz-Transcranial alternating current stimulation (tACS) selectively modulates speech perception. *International Journal of Psychophysiology* 101:18–24.

Ruhnau, P., Neuling, T., Fuscá, M., Herrmann, C. S., Demarchi, G., and Weisz, N. (2016). Eyes wide shut: transcranial alternating current stimulation drives alpha rhythm in a state dependent manner. *Scientific reports* 6:1–6.

Saenz, M. and Langers, D. R. (2014). Tonotopic mapping of human auditory cortex. *Hearing research* 307:42–52.

Saiz-Alía, M., Forte, A. E., and Reichenbach, T. (2019). Individual differences in the attentional modulation of the human auditory brainstem response to speech inform on speech-in-noise deficits. *Scientific Reports* 9:1–10.

Saiz-Alía, M., Miller, P., and Reichenbach, T. (2021). Otoacoustic emissions evoked by the time-varying harmonic structure of speech. *Eneuro* 8.

Saiz-Alía, M. and Reichenbach, T. (2020). Computational modeling of the auditory brainstem response to continuous speech. *Journal of Neural Engineering* 17:036035.

Sakata, S. and Harris, K. D. (2009). Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. *Neuron* 64:404–418.

Scharf, B., Quigley, S, Aoki, C, Peachey, N, and Reeves, A (1987). Focused auditory attention and frequency selectivity. *Perception & psychophysics* 42:215–223.

Scheidwasser-Clow, N., Kegler, M., Beckmann, P., and Cernak, M. (2021). SERAB: A multilingual benchmark for speech emotion recognition. *arXiv preprint arXiv:2110.03414*.

Schmidt, S, Redecker, C, Bruehl, C, and Witte, O. (2010). Age-related decline of functional inhibition in rat cortex. *Neurobiology of aging* 31:504–511.

Schreuder, M., Rost, T., and Tangermann, M. (2011). Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. *Frontiers in neuroscience* 5:112.

Schroeder, C. E. and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in neurosciences* 32:9–18.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Sevy, A. B., Bortfeld, H., Huppert, T. J., Beauchamp, M. S., Tonini, R. E., and Oghalai, J. S. (2010). Neuroimaging with near-infrared spectroscopy demonstrates speech-evoked activity in the auditory cortex of deaf children following cochlear implantation. *Hearing research* 270:39–47.

Shamir, M., Ghitza, O., Epstein, S., and Kopell, N. (2009). Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. *PLoS computational biology* 5:e1000370.

Shamma, S. (1989). Spatial and temporal processing in central auditory networks. *Methods in neuronal modeling: From synapses to networks*, 247–289.

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences* 34:114–123.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270:303–304.

Sharpee, T. O., Atencio, C. A., and Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current opinion in neurobiology* 21:761–767.

Shi, J. V., Xu, Y., and Baraniuk, R. G. (2014). Sparse bilinear logistic regression. *arXiv preprint arXiv:1404.4104*.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences* 12:182–186.

Skoe, E. and Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and hearing* 31:302.

Slabu, L., Grimm, S., and Escera, C. (2012). Novelty detection in the human auditory brainstem. *Journal of Neuroscience* 32:1447–1452.

Sohal, V. S., Zhang, F., Yizhar, O., and Deisseroth, K. (2009). Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature* 459:698–702.

Sohmer, H, Pratt, H., and Kinarti, R (1977). Sources of frequency following responses (FFR) in man. *Electroencephalography and clinical neurophysiology* 42:656–664.

Soli, S. D. and Wong, L. L. (2008). Assessment of speech intelligibility in noise with the Hearing in Noise Test. *International Journal of Audiology* 47:356–361.

Spille, C., Kollmeier, B., and Meyer, B. T. (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language* 52:123–140.

Spyridakou, C., Rosen, S., Dritsakis, G., and Bamiou, D.-E. (2020). Adult normative data for the speech in babble (SiB) test. *International journal of audiology* 59:33–38.

Tang, C., Hamilton, L., and Chang, E. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science* 357:797–801.

Teoh, E. S., Cappelloni, M. S., and Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience* 50:3831–3842.

Terreros, G. and Delano, P. H. (2015). Corticofugal modulation of peripheral auditory responses. *Frontiers in systems neuroscience* 9:134.

Thielscher, A., Antunes, A., and Saturnino, G. B. (2015). Field modeling for transcranial magnetic stimulation: a useful tool to understand the physiological effects of TMS? *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 222–225.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. *Doklady Akademii Nauk*. Vol. 151. 3. Russian Academy of Sciences, 501–504.

Tort, A. B., Komorowski, R., Eichenbaum, H., and Kopell, N. (2010). Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *Journal of neurophysiology* 104:1195–1210.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*:99–114.

Van Canneyt, J., Gillis, M., Vanthornhout, J., and Francart, T. (2021a). Neural tracking as an objective measure of auditory perception and speech intelligibility. *bioRxiv*.

Van Canneyt, J., Wouters, J., and Francart, T. (2021b). Cortical compensation for hearing loss, but not age, in neural tracking of the fundamental frequency of the voice. *Journal of Neurophysiology* 126:791–802.

Van Canneyt, J., Wouters, J., and Francart, T. (2021c). Enhanced neural tracking of the fundamental frequency of the voice. *IEEE Transactions on Biomedical Engineering* 68:3612–3619.

Van Canneyt, J., Wouters, J., and Francart, T. (2021d). Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics. *European Journal of Neuroscience* 53:3640–3653.

Van Eyndhoven, S., Francart, T., and Bertrand, A. (2016). EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Transactions on Biomedical Engineering* 64:1045–1056.

Vandecappelle, S., Deckers, L., Das, N., Ansari, A. H., Bertrand, A., and Francart, T. (2021). EEG-based detection of the locus of auditory attention with convolutional neural networks. *Elife* 10:e56481.

Vanheusden, F. J., Kegler, M., Ireland, K., Georga, C., Simpson, D. M., Reichenbach, T., and Bell, S. L. (2020). Hearing aids do not alter cortical entrainment to speech at audible levels in mild-to-moderately hearing-impaired subjects. *Frontiers in human neuroscience* 14:109.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology* 19:181–191.

Varghese, L., Bharadwaj, H. M., and Shinn-Cunningham, B. G. (2015). Evidence against attentional state modulating scalp-recorded auditory brainstem steady-state responses. *Brain Research* 1626:146–164.

Vaughan Jr, H. G. and Ritter, W. (1970). The sources of auditory evoked responses recorded from the human scalp. *Electroencephalography and clinical neurophysiology* 28:360–367.

Verhulst, S., Altoe, A., and Vasilkov, V. (2018). Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. *Hearing research* 360:55–75.

Verschueren, E., Vanthornhout, J., and Francart, T. (2021). The effect of stimulus intensity on neural envelope tracking. *Hearing Research* 403:108175.

Victor, J. D. (2005). Spike train metrics. *Current opinion in neurobiology* 15:585–592.

Vieira, P. G., Krause, M. R., and Pack, C. C. (2020). tACS entrains neural activity while somatosensory input is blocked. *PLoS biology* 18:e3000834.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17:261–272.

Vossen, A., Gross, J., and Thut, G. (2015). Alpha power increase after transcranial alternating current stimulation at alpha frequency (α-tACS) reflects plastic changes rather than entrainment. *Brain stimulation* 8:499–508.

Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26:1702–1726.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science* 167:392–393.

Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of cognitive neuroscience* 32:155–166.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. *Breakthroughs in statistics*. Springer, 196–202.

Wilsch, A., Neuling, T., Obleser, J., and Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *NeuroImage* 172:766–774.

Winer, J. A. (2005). Decoding the auditory corticofugal systems. *Hearing research* 207:1–9.

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience* 10:420–422.

Wood, C. C., Goff, W. R., and Day, R. S. (1971). Auditory evoked potentials during speech perception. *Science* 173:1248–1251.

Yang, M., Sheth, S. A., Schevon, C. A., Ii, G. M. M., and Mesgarani, N. (2015). Speech reconstruction from human auditory cortex with deep neural networks. *Sixteenth Annual Conference of the International Speech Communication Association*.

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

Yoshiura, T., Ueno, S., Iramina, K., and Masuda, K. (1995). Source localization of middle latency auditory evoked magnetic fields. *Brain research* 703:139–144.

Yu, D. and Deng, L. (2016). *Automatic Speech Recognition.* Springer.

Zaehle, T., Rach, S., and Herrmann, C. S. (2010). Transcranial alternating current stimulation enhances individual alpha activity in human EEG. *PloS one* 5:e13766.

Zilany, M. S., Bruce, I. C., and Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America* 135:283–286.

Zoefel, B., Archer-Boyd, A., and Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology* 28:401–408.