Neural Speech Tracking in EEG: Integrating Acoustics and Linguistics for Hearing Aid Users

Klara Almgren

Annie Mentzer

LUND
UNIVERSITY

Department of Automatic Control

MSc Thesis
TFRT-6243
ISSN 0280-5316

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

Printed in Sweden by Tryckeriet i E-huset
Lund 2024

Abstract

This master thesis explores the neural encoding of speech features for hearing aid users. The study utilizes electroencephalography (EEG) and audio data from an experiment that stimulates a Cocktail Party scenario. This is a complex auditory scene, especially difficult for individuals with hearing impairments. The primary objective of the study is to investigate how different acoustic and linguistic speech features are represented in the brain response and how these representations are influenced by hearing aid settings. The speech features analyzed are the **acoustic envelope, phonetic features, word onset**, and **word surprisal**. The word surprisal values were derived from GPT-2. Temporal Response Functions (TRFs) and multivariate TRFs (mTRFs) were employed to examine the correlation between these features and EEG signals during speech processing in both attended (target) and unattended (masker) speech scenarios.

The TRFs were estimated by training a forward model using a boosting algorithm. In this process, the speech features serve as the predictor variables (X-data), and the EEG signals serve as the response variables (Y-data). The boosting algorithm iteratively improves the model by combining multiple weak learners to better predict the EEG responses based on the given speech features.

The study found that target and masker speech are significantly distinguishable using TRF models trained on these features. It also revealed that hearing aid conditions impact their encoding. Among the features analyzed, the acoustic envelope had the highest correlation with neural responses. Adding other predictor variables to the Envelope model did not improve the correlation. Further, all speech features were found to have unique neural encoding. The acoustic envelope and phonetic features could be correlated to early processing, while word onset and word surprisal are reflected in later neural responses.

Our findings suggest that speech features are important in understanding how hearing aid users process speech, which could lead to future development of hearing aids that are not only fitted to the user's needs but also tailored to their unique neural responses.

Acknowledgements

We would like to express our gratitude to everyone who has supported us throughout the course of our master's thesis project. This work would not have been possible without the invaluable guidance, support, and collaboration of everyone involved.

First and foremost, we extend our sincere thanks to Eriksholm Research Centre, part of Oticon, for providing us with the opportunity to conduct our master's thesis. We are especially grateful to our industrial supervisors, Emina Alickovic and Martin Skoglund, whose expertise, insightful feedback, and continuous support were instrumental in the success of this project. We would also like to thank Johanna Wilroth and Sara Cartas for always being helpful with questions about the dataset and code.

We also wish to express our appreciation to Bo Bernhadsson, our academic supervisor at Lund University (LTH), for his guidance and feedback throughout the thesis process. We are also grateful to him for providing us with the necessary resources and facilitating our connection with Eriksholm, Emina, and Martin. Additionally, We would like to thank Pex Tufvesson for his insightful discussions and valuable programming tips.

A special thanks go to our fellow students at LTH and Eriksholm: Alessandro Celoria, Valentín López Jiménez, Sofia Boselli, and Gautam Sridhar. The collaborative environment and the mutual exchange of ideas have enriched our time as master's thesis students and contributed significantly to our work's progress.

Finally, we acknowledge the support from the NAISS project for providing computational resources under project NAISS 2024/22-283.

Thank you all for your support and encouragement.

Contents

1

Introduction

1.1 Introduction to Auditory Processing

When we perceive a sound, whether it’s speech, music, or environmental noise, the auditory information is initially captured by the ears and then conveyed as electrical signals to the auditory cortex located in the brain. The brain then processes and decodes spoken language in different regions in a hierarchical manner, through a complex network of neural pathways and structures [Hickok and Poeppel, 2007]. The neurophysiological mechanisms in speech recognition are affected by the many adverse speech conditions we experience in everyday life where speech intelligibility is challenged. Difficulties in perceiving and processing speech can impact speech recognition due to degradation of the speech signal, for example at the source or during transmission, which is often the case in noisy environments. These difficulties can also stem from receiver limitations due to individual hearing impairments and cognitive load [Mattys et al., 2012].

1.2 Problem Statement

The understanding of how the human brain processes speech, particularly in challenging listening environments, is a multifaceted problem that has significant implications for various fields, including neuroscience, linguistics, and audiology. In such environments, individuals must decipher speech among competing auditory inputs, which imposes additional cognitive demands, and presents a difficult challenge, especially for hearing impaired (HI) individuals. Difficulties with selective hearing attention in social settings and noisy environments often lead to feelings of frustration and exclusion for the affected, potentially resulting in social isolation, unemployment, and mental health issues [Podury et al., 2023].

1.3 Objectives

The primary objective of this study is to explore how acoustic and linguistic speech features are encoded and represented in the brain. The acoustic feature investigated is the acoustic envelope, while the linguistic features include word onsets, phonetic features, and word surprisals, the latter derived from a Large Language Model (LLM). This study examines their correlation with electroencephalography (EEG) signals during speech processing in a competing speech scenario. Specifically, it aims to determine whether **word surprisals**, as a linguistic feature, can be used to predict semantic processing of hearing aid users. Additionally, it explores whether integrating word surprisal with acoustic features enhances this correlation.

An additional objective is to investigate the potential impact of encoding acoustic and linguistic features across various hearing aid (HA) settings. Given that the data used in this study are obtained from hearing aid users, analyzing the correlation across different features will provide valuable insights into the brain's mechanisms under impaired hearing conditions. Ultimately, this research aims to advance our understanding of how different speech features are encoded in the EEG of hearing-impaired individuals in competing speech scenarios, and the impact of hearing aids and Noise Reduction (NR) on this process. Such understanding could have significant implications for the development of future hearing aids, potentially improving their effectiveness and user experience.

1.4 Methods

To address these objectives, several key questions have guided the methodology. First, the EEG data were preprocessed, and the speech features were extracted. Notably, the word surprisal values were derived and computed using the LLM GPT-2. To investigate the correlation between acoustic and linguistic speech features with EEG signals during speech processing, Temporal Response Function (TRF) models for individual features and multivariate TRF (mTRF) models combining different features were estimated using the Eelbrain toolbox in python.

The TRF models were then used to examine how acoustic and linguistic features are encoded in the EEG signals in both attended (target) and unattended (masker) speech across subjects, as well as under different hearing aid conditions. This analysis aims to evaluate if word surprisal consistently influences neural responses to target speech differently than to competing masker speech.

The TRF curves provide a graphic representation of these models, illustrating the temporal dynamics of the neural response to each specific speech feature, or combination of features, that the TRF or mTRF model was trained on.

Since different brain regions have been shown to correlate more strongly with various levels of speech processing, such as the temporal lobe for auditory processing

[Hickok and Poeppel, 2007] and the frontal lobe for semantic processing [Friederici, 2011], this study also explores the brain regions involved in the encoding of different features by analyzing topographic maps for each TRF model. This approach aims to gain a deeper understanding of the connection between speech features and neural activity during speech processing.

2

Background: Auditory Processing

This chapter covers the foundational aspects of auditory processing, particularly focusing on hearing loss and its implications, the cocktail party problem, and the stages of auditory processing. These are key elements for understanding the complexities involved in auditory processing and the challenges faced by individuals with hearing impairments. The following sections will provide an overview of hearing loss, the cognitive demands of listening in noisy environments, and the intricate processes involved in auditory signal processing, including the differentiation between masker and target speech streams.

2.1 Hearing loss

Definition and Prevalence

Hearing loss is defined as not being able to hear as well as a person with normal hearing, and it ranges from mild to profound. If the hearing loss is greater than 35 dB in the better-hearing ear, it can be labeled as a disabling hearing loss. Mild to severe hearing loss can be rehabilitated with assistive devices such as hearing aids and cochlear implants. More than 5% of the global population, equivalent to 430 million people, need rehabilitation to manage their hearing impairment. Projections suggest that by 2050, this number could surge to over 700 million people, affecting approximately 1 in every 10 people worldwide [World Health Oragnization, 2024].

Implications of Hearing Loss

Hearing loss can have great implications on the lives of those affected, both physically and socially. Psycho-social development can be disrupted in different ways during the different stages of life and can ultimately lead to increased morbidity in older adults [Podury et al., 2023].

Hearing is an integral part of language, communication, and social interaction. In early childhood, auditory input is important for the maturation of white matter in the brain's speech center, and the auditory cortex may be important for multimodal social processing. Speech contains acoustic features that carry socially rich information and convey emotion that may be difficult to interpret using the other sensory modalities. Conversations in noisy environments can be very challenging for those with impaired auditory processing. This can result in frustration, cognitive fatigue, and social avoidance. The degree of hearing loss in adults has been associated with work-related implications such as unemployment and increased need for sick leave [Podury et al., 2023].

The implications of hearing loss could lead to difficulties with social development in early childhood, diminished social well-being, social isolation, and increased risk of depression and anxiety in adulthood. In older age implications may also include cognitive decline and dementia, and although not yet fully understood, an increased risk of falls and frailty [Podury et al., 2023].

2.2 The Cocktail Party Problem

The cocktail party problem refers to the brain's ability to focus its attention on one speaker or auditory stimulus among many in a noisy setting [Cherry, 1953]. The phenomenon showcases an important human socio-cognitive capacity of sound segregation and selective attention. This ability is often more challenging for those with hearing loss and for those who use hearing aids [Reiss and Molis, 2021]. A remarkable phenomenon that can occur during a cocktail party problem is when a word of significance, such as the listener's name, is mentioned in one of the unattended speech streams and immediately catches the listeners attention.

Cognitive load

Listening effort can be described as a product of the processing demands in a certain situation and the cognitive resources of the individual, and is indirectly measured in behavior or performance. A highly demanding listening effort can lead to lower perception and comprehension. The cocktail party problem is an example of a listening scenario that can significantly increase the processing load. Other scenarios are when the linguistic properties of the stimulus are more complex or demand more memory, or when the listener is dividing their attention e.g. during multitasking [Johnsrude and Rodd, 2015].

2.3 Auditory Processing

Low-Level Processing

Low-level processing refers to the initial stages of auditory processing, where the brain analyzes basic acoustic features of the speech signal, such as spectral cues and temporal patterns. Adverse conditions, such as background noise or signal degradation, can impact low-level processing by reducing the clarity and quality of the incoming speech signal, making it harder for the brain to extract relevant acoustic information [Mattys et al., 2012].

Low-level processing is essential for extracting acoustic cues from the speech signal. At the acoustic level, the auditory cortex is primarily involved in decoding the physical properties of sound and speech, such as frequency, amplitude, and duration. The phonetic level involves the classification of acoustic patterns into phonemes, bridging the gap between acoustic information and linguistic units. This allows the brain to recognize and differentiate between different speech sounds [Brodbeck et al., 2018b].

High-Level Processing

While low-level processing is essential for extracting acoustic cues from the speech signal, high-level processing is crucial for integrating these cues into meaningful linguistic representations and facilitating comprehension. High-level processing involves more complex cognitive processes, such as lexical access, semantic interpretation, and syntactic analysis. Adverse conditions can also affect high-level processing by increasing cognitive load, reducing attentional resources, or introducing interference from competing speech signals or background noise [Mattys et al., 2012]. The lexical level allows the brain to identify and understand words in continuous speech by integrating phonetic information for word identification. Lexical processing occurs in the left temporal lobe, and is crucial for understanding the meaning of spoken language [Brodbeck et al., 2018b].

Compensatory Mechanisms of Speech Recognition

In response to adverse conditions that impact both low and high levels of processing, the brain may engage compensatory mechanisms to enhance speech recognition. These mechanisms may involve reallocating attentional resources, adjusting cognitive strategies, or relying more on contextual cues to aid in speech perception and understanding.

For example, in the process of word segmentation, the brain tends to rely more on acoustic and phonetic cues in non-noisy environments. In noisy environments, it becomes harder for the brain to rely solely on the acoustic characteristics of speech, i.e. the sound, to understand what's being said. In such situations, contextual information becomes particularly valuable for understanding speech. Even if certain speech sounds are obscured by noise, the context provided by surrounding words can help fill in the gaps and aid in comprehension [Mattys et al., 2012]. Additionally, phonetic features are more strongly encoded in the neural response when the stimulus is a comprehended language, i.e. there is high-level processing present. It has been suggested that the brain strategically relies more on low-level features when the context clues are constricted and more information is needed, and on high-level features when the context is reliable [Tezcan et al., 2023].

Stages of Auditory processing

Early Auditory Processing. In primary-like brain areas associated with early auditory processing (within approximately 85 milliseconds), previous research has demonstrated that it is not possible to distinguish between multiple talkers in individuals with normal hearing. In these early stages, the brain represents all speech with similar fidelity, regardless of the focus of attention [Alickovic et al., 2021].

Higher-Order Processing. In higher-order non-primary areas, later responses (beyond 85 milliseconds) reveal a different pattern. In these regions, there is a discernible increase in fidelity for the attended speech compared to the unattended speech. This phenomenon has been observed in both normal hearing (NH) and hearing-impaired (HI) individuals, where the neural responses related to target speech streams are enhanced while those related to masker speech streams are suppressed [Alickovic et al., 2021].

Brain Regions Involved in Target and Masker Speech Processing

Activation of Superior Temporal Gyri. Research indicates that when speech is masked by another speech signal, both the left and right superior temporal gyri (STG) are activated. However, when the speech signal is masked with spectrally rotated (unintelligible) speech, activation is observed only in the right STG. This suggests distinct roles for the left and right temporal lobes in processing different types of masked speech signals [Scott et al., 2009].

Impact of Different Maskers. It has been studied how different maskers, both energetic (e.g., white noise) and informational (e.g., speech babble), impact cortical auditory evoked potentials (CAEPs) during speech processing [Vander Werff et al., 2021]. It is known that informational masking, such as competing speech, is more challenging for normal hearing individuals compared to energetic masking [Brungart, 2001]. The study found that speech babble resulted in more significant latency delays in CAEPs compared to other noise types, indicating that the type of masking influences how speech is neurally encoded even without the participant actively attending to it [Vander Werff et al., 2021].

SNR and Age-Related Differences in Cortical Processing

How the brain responds to target speech in noisy environments is affected by age. Research indicates that elderly people need a higher Signal-to-Noise Ratio (SNR) compared to younger people to segregate words to the same extent in noisy environments, both in informational and in energetic noise [Schneider et al., 2022]. SNR compensates for age-related differences in speech processing, particularly in competing speech scenarios, which are more difficult for older adults because of the similar acoustics between the target and masker speech streams. Younger individuals can more effectively use the onset differences between the target and masker speech streams to focus on and understand the target speaker. In contrast, older adults struggle to benefit from these temporal cues, making it harder for them to process speech in noisy environments [Schneider et al., 2022].

Phase Locking and Linguistic Information

Phase Locking Mechanisms. Phase locking refers to the synchronization of neural responses to the timing of auditory stimuli. It has been investigated whether phase locking of cortical responses benefits from linguistic information or is solely a response to acoustic information in speech [Peelle et al., 2012]. The result of the study showed that phase locking between the envelopes and neural activity is the strongest between 4 Hz and 7 Hz, and that this is where the acoustic envelope of sentences contains most power. It was also demonstrated that the listener's ability to extract linguistic information is related to the brain's entrainment to connected speech. Phase locking was proven to be significantly more cerebro-acoustically coherent in the left auditory cortex than in the right for high intelligibility speech. The impact of speech intelligibility on phase-locked responses is observed in lower-level auditory regions of the temporal cortex. This sensitivity in areas that are anatomically early in the speech-processing hierarchy suggested that these regions are attuned to linguistic information [Peelle et al., 2012].

Semantic Encoding. The encoding of semantic information in EEG signals is associated with specific frequency ranges, particularly in the theta (4 Hz to 8 Hz) and high beta (13 Hz to 30 Hz) bands. Studies have shown that theta activity is linked to associative memory formation and semantic processing, whereas beta activity is often related to maintaining a defined level of cognitive processing and predicting upcoming stimuli. This suggests that these frequency ranges play a crucial role in how the brain encodes and processes semantic information, implying their importance in surprisal analysis [Zion Golumbic et al., 2013].

3

Background: EEG

This chapter explores electroencephalography (EEG) as a tool to understand the neurophysiological responses to auditory stimuli. It will cover the basics of EEG, electrode placement, frequency bands, and the phenomena of neural tracking. Additionally, the chapter will outline how EEG aids in understanding speech intelligibility for hearing aid users, the significance of event-related potentials (ERPs), and the application of temporal response functions (TRFs).

3.1 Introduction to EEG

EEG is a technique commonly used by scientists to study human brain functions and by physicians to diagnose neurological disorders such as epilepsy, sleep disorders and brain tumors. EEG involves using a cap with superficial (or in some cases intracranial) electrodes placed on the scalp to measure potentials produced by the brain cells, reflecting the electrical activity of the brain. The electrodes are connected to amplifiers and an EEG machine that records the activity. The electrodes are recorded in parallel, with each channel typically consisting of two electrodes. Each channel produces a signal, and together these signals generate a graph [Siuly et al., 2016]. The placement of the electrodes, or montage, is important as different areas of the cerebral cortex process different stimuli and tasks [Siuly et al., 2016].

3.2 Frequency Bands in EEG

The most important characteristic of an EEG recording is its frequency content. Specific frequency bands are correlated with different mental states, such as active attention, wakefulness, and various stages of the sleep cycle. The main frequency band where auditory information is processed in the brain is between 1 Hz to 7 Hz [Zion Golumbic et al., 2013].

3.3 Neural Tracking

Neural tracking is a phenomenon where neural responses, commonly measured with EEG, temporally align with specific acoustic features of a continuous sound input, e.g. speech [Tezcan et al., 2023]. This can be used to investigate the hierarchical properties of speech processing in the brain.

Methods of Neural Tracking

Frequently used methods of neural tracking are **linear decoding** and **encoding models** [Alickovic et al., 2019; Geirnaert et al., 2021]. Decoding, also called **backward modelling**, refers to the process of reconstructing the stimulus. Encoding, or **forward modelling**, involves predicting the neural response, such as the EEG signal. By modelling a linear relationship between the stimulus feature and the neural activity, it is possible to gain valuable insights into how the brain processes and responds to different auditory inputs. Frequently tracked features include the fundamental frequency f_0 [Van Canneyt et al., 2021; Kegler et al., 2022], the acoustic envelope [O'sullivan et al., 2015; Mirkovic et al., 2016; Alickovic et al., 2019; Di Liberto et al., 2020; Biesmans et al., 2016], and linguistic features such as phoneme and word surprisal [Di Liberto et al., 2015; Broderick et al., 2018; Brodbeck et al., 2018a; Gillis et al., 2021; Anderson et al., 2023; Chalehchaleh et al., 2024; Karunathilake et al., 2024]. By incorporating different tracking features, it is possible to gain insights into auditory processing and speech intelligibility at different levels [Gillis et al., 2022].

Auditory Attention Decoding

Auditory attention decoding (AAD) is a technique used to identify which auditory stimulus a listener is focusing on by analyzing neural signals, such as EEG. This method holds significant importance in applications such as hearing aids and brain-computer interfaces, where understanding auditory attention can profoundly enhance user experience. Common attention decoding methods rely on correlation and estimation techniques that link electrophysiological responses to auditory stimuli. The primary techniques include Canonical Correlation Analysis (CCA), dense estimation encoding/decoding models, and sparse estimation encoding/decoding models [Alickovic et al., 2019].

Incorporating speech features into AAD serves to enhance the accuracy and robustness of these techniques. Speech-to-language transformations, modeled by systems like Whisper [Radford et al., 2023], can predict EEG responses by using spectral speech features contextualized over time [Anderson et al., 2023]. This approach demonstrates that higher layers of speech-to-language models, which integrate longer temporal contexts, can be used to better predict EEG signals related to auditory attention, thus improving the AAD performance [Anderson et al., 2023].

3.4 Speech Comprehension for Hearing Aid Users

Hearing Aid Strategies

Historically, hearing aids have been focused on amplifying sound and reducing noise, but are not working cooperatively with the thinking brain. There can be great benefit in looking at top-down processing for complex stimuli such as speech and music. The thinking brain uses mechanisms to extract meaning from complex soundscapes and can filter out specific signals from the background. The strategies used for auditory problem-solving vary among individuals, and there is much to learn from the superior strategies to tailor hearing aids more effectively on an individually basis [Hafter, 2010].

Speech Comprehension

Research has shown that semantic information might be more important for the intelligibility of speech than the SNR. A study investigated how speech intelligibility and perceived cognitive effort were affected by two different experimental setups with reverberant background noise for hearing aid users. The results showed that maintaining the same topic for sentences presented to subjects, i.e. the same semantic context, enhanced intelligibility and reduced self-reported cognitive effort. The same topic increased intelligibility for 95% of subjects, while the reverberant room hearing aid setting (NR setting) increased intelligibility for 75% of users compared to the universal hearing aid setting. The listening effort was also lower in the same context and reverberant room conditions. There was no significant interaction between context and the Reverberant Room setting for either of the results [Holmes et al., 2018].

It has also been found that intelligibility remains largely consistent despite background noise until the noise reaches a level where it overpowers the energy in the speech signal, typically around -3 dB [Brungart, 2001]. Another study [Hafter, 2010] suggests that although NR hasn't been proven to improve speech intelligibility in noise, it is still positively commented on by the users. The appreciation could be explained by a reduction of cognitive effort that the NR offers. In shared attention scenarios, hearing aid users might enjoy the NR not because it improves speech identification, but because it makes it easier for them to focus on the task [Hafter, 2010].

However, a more recent study has shown that an NR scheme for HAs had an enhancing effect on the cortical representation of both target and maker speech in both low and high SNRs. In low SNRs, it also suppressed the representation of background babble noise [Alickovic et al., 2020; Alickovic et al., 2021].

3.5 Event Related Potentials (ERPs)

In conventional event-related analysis, it is presumed that every stimulus triggers a uniform and distinct response in the EEG. ERPs are employed to study discrete events, like words, which are integral to human speech perception. ERPs are calculated as averages of the neural signal before and after the onset of the word, which may include neural signals from other responses to other stimuli, such as the acoustic signal, as well as responses to the previous and next word [Brodbeck et al., 2021]. An example of what an ERP can look like is depicted in Figure 3.1. The use of ERPs is therefore more suitable for studying discrete sound rather than continuous. ERPs are localized in different brain regions depending on the task and type of stimulus.

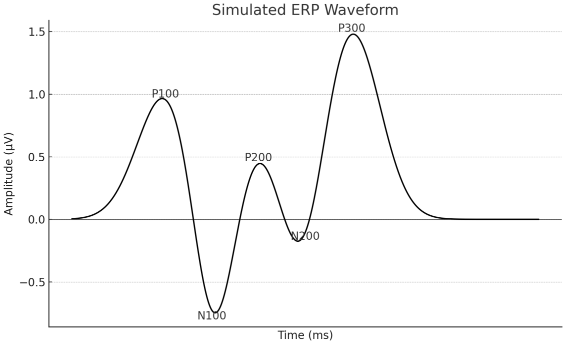


Figure 3.1 A simulated ERP waveform showing the peaks P100, N100, P200, N200, and P300. The x-axis represents time in milliseconds, and the y-axis represents amplitude in microvolts (µV).

Early Peaks

N100: A negative peak between 100-160 ms is indicative of early auditory processing and is correlated to the onset and discrimination of a stimulus [Näätänen and Picton, 1987]. The N100 peak is related to perception, and the latency of the N100 peak has been shown to increase with age [Tomé et al., 2014]. The enhancement of the N100 component when attending to a stimulus can be useful for evaluating the effectiveness of electrical stimulation prior to the implementation of cochlear implants for hearing-impaired individuals [Paulraj et al., 2015]. N100 is often followed by the P200-peak, forming the so-called N1-P2 complex, distributed over the fronto-central scalp region [Tomé et al., 2014].

P200: The P200 peak (150-200 ms latency), in ERPs typically reflects the early stage of sensory processing and is associated with the encoding and initial analysis of sensory information. The P200-peak is related to attention and perceptual processing [Bourisly and Shuaib, 2018]. The P200 amplitude has been shown to increase with age (20-60 years) and then decrease again. The latency also increases with age in the anterior brain regions [Mueller et al., 2008].

Mid-Latency Peaks

N200: The N200 peak (200-350 ms latency) is correlated with selective attention and separation of auditory stimuli [Sur and Sinha, 2009] and processing of phonological information, which is related to the sounds of the language [Tomé et al., 2014]. These peaks are related to higher cognitive processes such as attention, discrimination, and memory. They indicate how the brain allocates cognitive resources to process and respond to significant or novel stimuli. The N200 is generally observed in the frontal-central areas of the brain, although the specific cites are task dependent. [Folstein and Van Petten, 2007].

P300: The P300 peak is elicited when the subject actively listens to the target stimuli [Picton, 1992]. It is related to higher cognitive processing such as decision-making, memory and context updating. The peak latency of the P300 can indicate how fast the individual is processing the stimulus [Mueller et al., 2008]. The P300 peak is found in the parietal and frontal cortex and originates from stimulus-driven frontal attention mechanisms during task processing, and from temporal-parietal activity associated with higher cognitive processing [Polich, 2007].

Late Peaks

N400: Around 400 ms latency, speech understanding occurs and words are encoded. Late negative peaks around 400 milliseconds reflect semantic processing and integration, indicating higher-level cognitive functions related to understanding and integrating complex stimuli. The N400 component is typically associated with activity in the frontal, temporal and parietal regions of the brain[Kutas and Federmeier, 2011]. Due to the peak's relation to integrating semantics, it could potentially encode the word surprisal.

3.6 Temporal Response Function (TRF)

TRF measures how the brain responds to continuous sounds, like speech, over time. TRFs describe the linear relationship between a stimulus and the brain's reaction. By using regularized linear regression, TRFs help understand how different features of sounds are processed by the brain. TRFs act like filters, showing how ongoing stimuli are transformed into neural responses, assuming a linear convolution between input and output [Crosse et al., 2016].

A TRF is a mathematical model that describes how a system responds over time to a given input. TRFs thereby provide a framework for analyzing neural responses to continuous stimuli, capturing both the continuous nature of the stimulus and the discrete events or features within it. This is achieved through the utilization of time series predictor variables, such as multiple impulses (ERPs), which undergoes convolution with a kernel that outlines the general pattern of responses to this event type [Brodbeck et al., 2021]. The TRF, $w(\tau,n)$, is estimated by minimizing the mean-squared error (MSE) between the actual neural response at channel n, $r(t,n)$, and the predicted response $\hat{r}(t,n)$ obtained from the convolution:

$$\min \varepsilon(t,n) = \sum_{t} [r(t,n) - \hat{r}(t,n)]^2 . \tag{3.1}$$

Although TRF analysis assumes a linear relationship between the EEG response and stimulus feature, the approach is based on the understanding that brain processes are continuous, non-linear, and hierarchical. This suggests that brain signals can be decomposed into separate responses linked to various predictor variables that are processed simultaneously at different levels. This can be achieved by estimating a **multivariate Temporal Response Function** (mTRF). This approach can provide valuable insights into the unfolding dynamics of perception and cognitive processes over time [Brodbeck et al., 2021].

Non-Linear Transformations

Convolution is limited by its tendency to model linear responses to input stimuli. However, the brain exhibits a range of transformations in response to stimuli, rather than simply mirroring them. In this context, a non-linear response is more appropriate. To address this, the stimulus can be subjected to non-linear transformations. Subsequently, brain responses can be predicted based on linear responses to these newly non-linear stimulus features [Brodbeck et al., 2021].

TRF Peaks

It is possible to interpret TRF peaks similarly to ERP peaks, as both TRFs and ERPs are derived from EEG data and reflect the brain's response to stimuli. The main difference is that TRFs provide a mapping of how the brain responds to continuous stimuli over time, while ERPs are typically associated with specific, discrete events [Crosse et al., 2016]. Nevertheless, TRF and ERP peaks can be interpreted in terms of their amplitude and latency in a similar way, reflecting the strength and timing of various neural responses.

3.7 Predictor Variables

Predictor variables can be interpreted as hypotheses about the representation of stimuli in the brain. The stimulus is represented by a chosen predictor variable, such as the acoustic envelope, audio spectrogram, word and phoneme onsets, etc., and a convolution model specifies its relation to the response. TRFs model the continuous relationship between the stimulus (e.g., speech audio signal) and the neural response over time, allowing for a detailed analysis of the temporal dynamics of brain activity in response to continuous stimuli. By assuming that the responses are additive, multiple predictor variables can be chosen to characterize the stimuli, resulting in an mTRF model consistent with macroscopical measurements of electric brain signals and more natural stimulus conditions [Brodbeck et al., 2021].

The Acoustic Envelope

The acoustic envelope describes how the amplitude of an oscillating signal, e.g. sound wave, changes over time and can be extracted in digital signal processing by for example using the Hilbert transform or the moving root-mean-square amplitude [The MathWorks Inc, n.d.]

Role of Acoustic Envelope in Neural Processing. The temporal envelope of speech is a key feature strongly represented in the cortex and closely related to neural activity. Studies have shown that there is a strong correlation between cortical activity due to auditory activity and acoustic envelopes of speech [Horton et al., 2014; O'sullivan et al., 2015]. The brain's early auditory responses are closely aligned with the temporal patterns of the speech envelope, indicating that the acoustic envelope is a primary driver of these neural activities [Oganian et al., 2023]. This encoding helps the perception of speech and other complex sounds, enabling the brain to process auditory information effectively.

Attention Prediction. The speech envelope has been shown to play a significant role in decoding attention in cocktail party situations, where the EEG response phase-locks to the acoustic envelope of the attended speaker, while remaining out of phase with the unattended speaker [O'sullivan et al., 2015]. A study demonstrated that a classifier trained on individuals' envelope responses, obtained by cross-correlating each stimulus's envelope with EEG channels, could reliably predict the attended speaker with significant accuracy. Moreover, this classifier successfully predicted the subject's attention using data from other subjects, implying promising prospects for developing auditory Brain-Computer Interfaces that don't necessitate individual user data for training [Horton et al., 2014].

A promising alternative is to use cepstrum analysis for AAD, which has been shown to improve classification accuracy [Alickovic et al., 2023a; Alickovic et al., 2023b]. This approach involves performing the Fourier transform on a signal, taking the logarithm of its magnitude, and then applying the Fourier transform again. The resulting cepstrum can highlight periodic structures in the frequency domain, such as echoes or harmonics, that are not easily visible in the original signal or its frequency representation. This advantage makes cepstrum analysis a valuable tool for identifying hidden features in complex signals.

Word Onset

Word onsets refer to the beginning sounds of words and are important for speech perception and processing. The human auditory system is sensitive to these initial sounds, which help in segmenting continuous speech into discrete units, facilitating word recognition and comprehension. The brain's response to word onsets has been shown to be robust, as these initial sounds often carry important phonetic and prosodic information that aids in predicting the rest of the word. Research using electrophysiological methods, such as ERPs, has demonstrated distinct neural responses to word onsets, typically observed as peaks in the P200 and N200 components, reflecting early auditory processing and cognitive evaluation of speech sounds [Marslen-Wilson and Zwitserlood, 1989].

Phonetic Features

Phonetic features are a combination of the onset of phonemes and how they are articulated. The phonemes of a language can be categorized by distinct characteristics such as voicing, manner of articulation, and positioning of the tongue and lips. These features are essential for distinguishing between different phonemes, the smallest units of sound that can change meaning in a language. The processing of phonetic features is a foundational aspect of speech perception, allowing listeners to decode the acoustic signal into meaningful linguistic units. Early sensory processing of phonetic features is reflected in the P100 component of ERPs, which typically occurs within the first 100 milliseconds after stimulus onset [Näätänen and Winkler, 1999].

Word Surprisal

Word surprisal is a measure of the unexpectedness of a word given its previous words in a sentence. It is represented as probability values and computed as the negative log likelihood of the word (3.2) [Heilbron et al., 2022]. Since the brain predicts upcoming content based on prior knowledge, word surprisal enables investigations into how word predictability affects the brain response in natural sentences. Studies have linked higher surprisal values to longer reading times as well as to increased activation in language-related brain areas as measured by functional MRI [Lowder et al., 2018]. Recent research has successfully used the LLM GPT-2 to generate estimates of word surprisal to capture semantic and lexical processing in normal hearing subjects [Heilbron et al., 2022; Anderson et al., 2023]. Word surprisal is calculated using the negative log likelihood of the probability of the next word given its previous context. For a word w_t at position t in a sequence, given the context of previous words $w_1, w_2, \ldots, w_{t-1}$, the surprisal is defined as:

$$\text{Surprisal}(w_t) = -\log P(w_t \mid w_1, w_2, \ldots, w_{t-1}) \tag{3.2}$$

Implications for Speech Processing Models. By investigating whether word surprisal from GPT-2 improves prediction accuracy/correlation in speech processing models, we can gain a deeper understanding of how linguistic context influences neural responses during speech comprehension. This knowledge could lead to the development of more robust speech processing algorithms and models, with applications in fields such as automatic speech recognition and natural language understanding.

4

Background: Neural and Computational Models in Speech Processing

This chapter investigates the intersection of neural and computational models in speech processing, focusing on tools and techniques that facilitate the understanding and analysis of how the brain processes language. It will introduce the Eelbrain toolkit for modeling neural responses and examine the architecture and capabilities of the GPT-2 language model in natural language processing tasks.

4.1 Eelbrain Toolkit

Eelbrain is a Python toolkit that makes time-lagged regression, using TRFs, easy and accessible to model neural responses to speech and language processes and evaluate them against electrophysiological brain responses, e.g. EEG recordings. The approach is based on the assumption that the brain processes are continuous, non-linear and hierarchical. The brain signals can be broken down into separate responses linked to various predictor variables. This can be done by estimating a mTRF and the approach has provided valuable insights into the unfolding dynamics of perception and cognitive processes over time [Brodbeck et al., 2021].

4.2 Language Models for NLP-tasks (GPT-2)

Model Training and Architecture

During training, the language model learns to predict the next word, or token, in a sentence based on preceding input, enabling it to grasp complex linguistic structures and meanings. A token is the basic unit of text used in natural language processing, typically representing words or punctuation. The model consists of neural network layers that compute hidden states for each token, capturing contextual information. To predict the next token, GPT-2 compares these hidden states with token embeddings, estimating the probability distribution over potential tokens. GPT-2 operates in an autoregressive manner, generating one token at a time based on preceding tokens, thus crafting cohesive sentences [OpenAI, 2019].

Hierarchical Prediction

A study has been investigating the hierarchy of linguistic prediction during natural speech comprehension by using the LLM GPT-2 to quantify contextual predictions [Heilbron et al., 2022]. The study found that the brain uses continuous probabilistic prediction of upcoming speech to support interpretation, similar to how GPT-2 predicts upcoming input. The results support hierarchical predictive processing, indicating that word prediction informs phoneme prediction, and that the prediction occurs spontaneously at multiple levels of abstraction. This suggests that the brain and LLMs share similar mechanisms for processing language [Heilbron et al., 2022].

Zero-Shot Task Transfer

Using language models for natural language processing tasks has led to several key findings and implications, such as Zero-shot Task Transfer which demonstrates the versatility and adaptability of language models to handle various tasks beyond their original training objectives. Additionally, language models can engage in unsupervised multitask learning, learning and performing various tasks demonstrated in natural language sequences without interactive communication. This showcases their ability to generalize to various tasks as well as their versatility and adaptability in handling different natural language processing tasks effectively [Radford et al., 2019].

Importance of Diverse Datasets

The success of zero-shot task transfer is influenced by the model’s capacity, with larger models exhibiting improved performance across tasks. It’s essential to build large and diverse datasets for training language models, gathering natural language demonstrations across different domains and contexts (sources like Common Crawl provide valuable data for this purpose) [Radford et al., 2019].

Model Sizes and Parameters

GPT-2 comes in different sizes with varying numbers of layers and dimensions. The model sizes and corresponding parameters are as follows

- 117M: 12 layers, 768 dimensions
- 345M: 24 layers, 1024 dimensions
- 762M: 36 layers, 1280 dimensions

- 1542M: 48 layers, 1600 dimensions

Byte Pair Encoding (BPE)

GPT-2 employs a Transformer-based architecture, enabling it to estimate the probability of any Unicode string, which the model treats as sequences of UTF-8 bytes.

GPT-2 uses Byte Pair Encoding (BPE) as part of its input representation strategy. Directly applying BPE to byte sequences can lead to suboptimal token merges, particularly for common words. This can result in inefficient token allocation and fragmentation of words across multiple vocabulary tokens. To address this issue, Radford et al. [Radford et al., 2019] suggest a modification where BPE does not merge symbols from different character categories (e.g., alphabetic characters, punctuation symbols) for byte sequences, with an exception for spaces. This is done with the aim to improve the efficiency of token allocation and reduce fragmentation of words across multiple vocabulary tokens when using BPE with byte sequences.

5

The Data

The data utilized in this thesis were collected in a study by Eriksholm Research Centre. For comprehensive details regarding the dataset, readers are referred to [Alickovic et al., 2021]. The study was approved by the ethics committee, and all subjects signed a written consent. The experiment initially involved 34 hearing aid users, aged 21 and 84 years. However, due to missing or poor-quality data, this thesis uses data from 28 of these subjects. Throughout the thesis, all subjects are identified by an ID number, such as ID1031.

5.1 Experimental Setup

The experimental setup is illustrated in Figure 5.1 and aimed to simulate a cocktail party effect. Subjects were placed in a sound-proofed room surrounded by six loudspeakers. Two loudspeakers were positioned in front of the subject at slight offset angles symmetrically from the center, while the remaining four were symmetrically distributed behind the subject. Participants were exposed to 4-talker babble noise (3 dB SNR) from each of four background loudspeakers and presented with target (attended) and masker (unattended) speech streams from the two foreground loudspeakers. The selection and order of the presented audio were different for each subject, session, and trial. The target speaker was either male or female, with the masker being of the opposite gender to the target.

During each trial, EEG data were collected from the subjects using a high-density BioSemi ActiveTwo 64-channel montage, from which the data in this thesis were derived. The sensor layout is illustrated in Figure 6.3

The experiment was divided into four sessions, with each session comprising 20 trials. In each session, one of four hearing aid conditions was randomly selected and utilized. The order of conditions was randomized independently for each subject across sessions. Before each trial, a visual prompt appeared on a screen, indicating the gender of the speaker and the designated side for attention. The gender and speaker positions were randomized, but the same male and female voices were used throughout all trials. The audio stimuli were neutral-toned Danish speech streams in WAV format with a sampling frequency of 44.1 kHz. The speakers maintained a consistent rhythm and pauses longer than 200 ms had been removed. Each trial began 5 seconds after the start of the babble noise and lasted for 33 seconds.

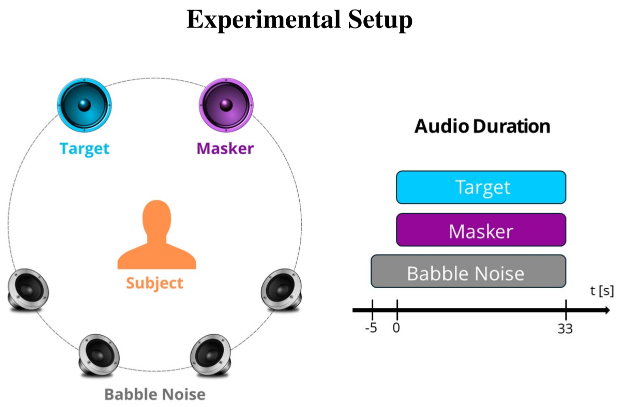


Figure 5.1 Schematic of the experimental setup and duration of the speech streams played for one trial.

6

Method

6.1 EEG

The preprocessing of the EEG data was done prior to this thesis project by Johanna Wilroth and included e.g. referencing, resampling, removing artifacts, and manual Independent Component Analysis (ICA). Further details about the dataset can be found in [Alickovic et al., 2021]. To facilitate its utilization with the Eelbrain toolbox, the data underwent conversion into an NDVar (N-dimensional variable), a widely employed data container in neuroscience. This EEG-NDVar contained temporal information alongside EEG signals from each channel in the montage and trials within a session, with four sessions per subject.

The EEG electrode montage was modeled after the Biosemi64 configuration, omitting the two reference channels, also known as mastoid channels, labeled 'EXG1' and 'EXG2'. Following this setup, EEG signals underwent bandpass filtering within the 1 Hz to 8 Hz range, employing a finite impulse response (FIR) filter, as this frequency band aligns with auditory processing in the brain [Zion Golumbic et al., 2013].The FIR-filter used zero-phase filtering, ensuring that the temporal characteristics of the signals are preserved, which is crucial for TRF analysis. Subsequently, the EEG data was downsampled from its original 256 Hz to 64 Hz. Given that the EEG recording extended beyond the duration of the target and masker speech streams playback, it was padded to synchronize with the start and stop times of the speech streams. Finally, the time axis was re-centered to zero for coherence with subsequent analyses. Examples of the resulting EEG signals for a trial can be viewed in Figure 6.1 (64 channels) and Figure 6.2 (one channel).

64 Channel EEG Example

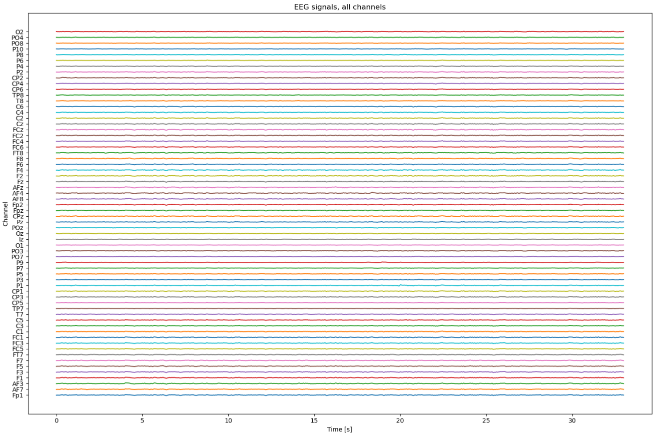


Figure 6.1 Example of the EEG data for one subject, and trial, representing all the 64 channels.

Single Channel EEG Example

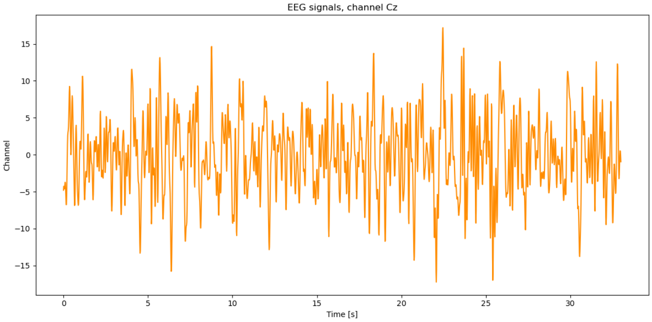


Figure 6.2 Example of the EEG data for one trial measured by the 'Cz' channel.

Sensors

For most of the analysis, 17 electrodes from the fronto-central region were used to reduce the influence of channels that might capture neural activity unrelated to the auditory stimuli. An illustration of the EEG montage and the selected sensors, marked in blue, can be seen in Figure 6.3. The chosen channels cover areas of the brain involved in auditory processing, including the auditory cortex and regions associated with attention and working memory related to auditory stimuli, see Section 3.5 for reference.

Annotation. The channels are named after the brain region they capture:

Brain regions: F - Frontal, T - Temporal, C - Central, P - Parietal, O - Occipital, I - Inion.

Positions: A - anterior, z - midline

Odd and Even Numbers: Odd numbers (1, 3, 5, 7, 9) indicate electrodes on the left hemisphere, while even numbers (2, 4, 6, 8, 10) indicate electrodes on the right hemisphere.

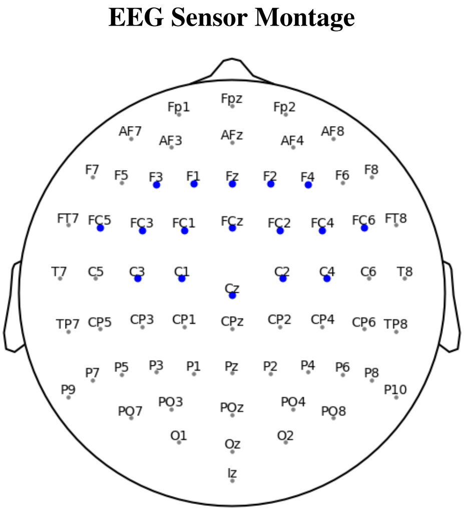


Figure 6.3 Depiction of the EEG sensor montage. The 17 channels chosen for analysis are marked with blue circles.

6.2 Predictor Variables

The features extracted from the audio files were the acoustic envelopes, word onset times, and phonetic features. Word surprisal values were computed from transcriptions using a GPT-2 model and combined with the word onsets to create the surprisal. Further details are given in the text below.

These features were used as predictor variables in the training of the (m)TRF models. For each subject, session, and trial, these features were saved as NDVars, matching the format of the EEG data, with the additional dimension indicating whether the specified audio file was used as the target or masker speech.

Envelopes

The dataset included all the speech tracks used in the experiment as target and masker speakers, as well as babble noise. However, only the speech tracks were used in the analysis. Acoustic envelopes were extracted from WAV files using Eelbrain's function for computing the Hilbert envelope. These envelopes were then bandpass filtered, using a zero-phase FIR filter, within the 1 Hz to 8 Hz range and were resampled to 64 Hz to align with the frequency content and sampling rate of the EEG. Figure 6.4 shows an example plot of one of the resulting envelopes.

Envelope Example

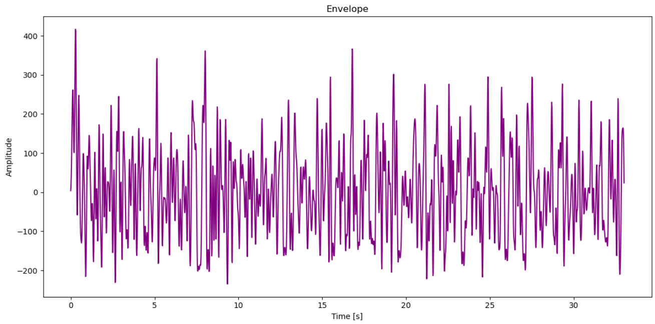


Figure 6.4 Example of the filtered acoustic envelope from one trial. The duration is 33 s and the sampling frequency is 64 Hz.

Word and Phoneme Onsets

The features 'word onsets' and 'phoneme onsets' both encode the starting time, or onset, of said feature during speech. Complementary information about the audio files was provided in Textgrid files [Team, 2024] and processed using Matlab. The word onsets and phoneme onsets were extracted along with a word list for every speech audio file.[1] The onset of a word or phoneme is represented by an impulse with an amplitude of one, while all other samples, including the onsets of silent pauses, are set to zero. An example of word onsets can be seen in Figure 6.5, where every vertical line represents the onset of a word. The phoneme onsets were used to create the phonetic features.

[1] A script authored by Yen-Liang Shue was used to read the Textgrid files, and a function developed by Sara Cartas was utilized to extract the phoneme feature map, word onset times, and list of words from each audio track.

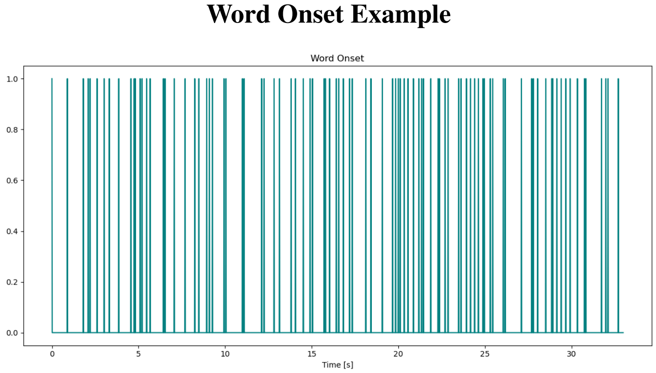
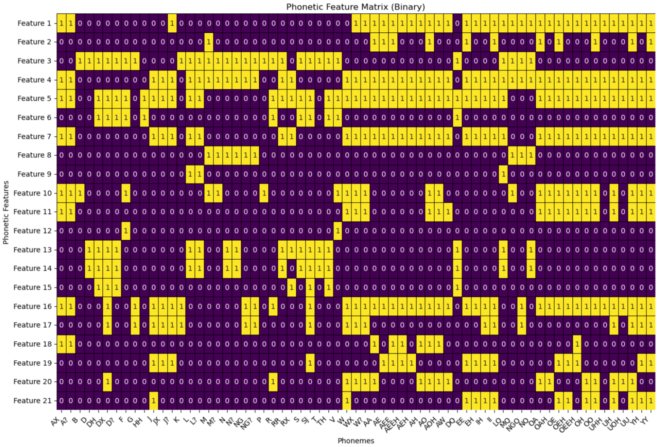


Figure 6.5 Example of the word onsets from one trial.

Phonetic Features

The phonetic features were constructed using a phonetic feature matrix and a phonetic occurrence matrix, containing the onsets of each phoneme in the speech track. The phonetic feature matrix is a binary representation of Danish phonemes and their corresponding phonetic features. The matrix consists of 65 rows and 21 columns. Each row corresponds to a specific Danish phoneme, and each column represents a distinct phonetic feature. The presence of a feature in a phoneme is indicated by a value of one, while its absence is indicated by a value of zero, see Figure 6.6 for reference. The phonetic features include properties such as voicing, place of articulation, manner of articulation, and other articulatory and acoustic characteristics. When this matrix is multiplied by a time-dependent phonetic occurrence matrix, it provides a temporal representation of phonetic features present in an audio track. An example of a resulting phonetic feature map is shown in Figure 6.7, where row is a feature and every vertical line is the onset of that feature.

Phonetic Feature Matrix



	AX	A?	B	D	DH	DX	D?	F	G	HH	J	JX	J?	K	L	L?	M	M?	N	N?	NG	NG?	P	R	RR	RX	S	SJ	T	TH	V	W	WX	W?	AA	AE	AEE	AEEH	AEH	AH	AO	AOH	AW	DQ	EE	EH	IH	II	LQ	MQ	NGQ	NQ	OA	OAH	OE	OEH	OEEH	OH	OO	UHH	UH	UOH	UU	YH	YY
Feature 1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Feature 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	1
Feature 3	0	0	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 4	1	1	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Feature 5	1	1	0	0	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Feature 6	0	0	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 7	1	1	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Feature 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 10	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	1	0	1	1	1
Feature 11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	1	0	1	1	1
Feature 12	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 13	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 14	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 15	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Feature 16	1	1	0	0	0	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Feature 17	0	0	0	0	0	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1
Feature 18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Feature 19	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1
Feature 20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	1	1	0	0
Feature 21	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0	1	1	0	1	0	1	1	1

Figure 6.6 The phonetic feature matrix used to construct the phonetic feature map. Each row corresponds to a phonetic feature and the numbers in the columns refer to absence (0 - purple) or presence (1 - yellow) of the phoneme specific to that column.

Phonetic Features Example

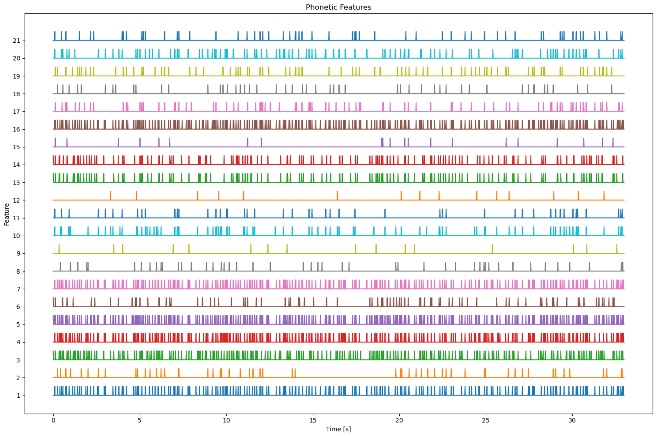


Figure 6.7 Example of a phonetic feature map for one trial. Each row represents a phonetic feature and the onsets of it.

Surprisal

To calculate word surprisal, a Python function was created which takes a string of text as input and computes surprisal values for each word in the text. The tokenizer and model used to calculate the surprisal of the speech was the pre-trained GPT-2 model ‘KennethTM/gpt2-medium-danish’ that has been fine-tuned to the Danish language [KennethTM, 2021]. The training data used to fine-tune the model is part of the ‘oscar’ dataset, with a context length of 21,024 tokens. These were loaded using the Auto classes from the Transformers library in Hugging Face [Wolf et al., 2020].

The model processed the input word by word, and therefore the context when calculating the probability of a word in that sequence is the previously processed words. Surprisal values quantify the unexpectedness of a word given past context. The wordlists from each audio file were concatenated as text string objects and were fed to the Python function separately.

First, the input text was tokenized using the tokenizer, which returned tokenized representations in the form of tensors. A tensor is a multi-dimensional array that generalizes scalars, vectors, and matrices to higher dimensions, enabling the complex computations required for tasks such as language modeling with GPT-2. Then, the text was encoded again using the same tokenizer to obtain tokenized text tensors. Next, predictions were generated by passing the tokenized text tensors through the model. The predictions consist of logits, which represent the model’s estimated probabilities for each token. The function iterates through each token in the tokenized text and calculates the softmax probabilities for each token prediction.

When text is tokenized, it is broken down into smaller units such as words or subwords, each of which is assigned a unique identifier called a token ID. However, in the tokenization scheme, words or subwords got broken down into subword units in the byte pair encoding (BPE) [Radford et al., 2019]. An additional function was implemented to bypass this and its basic concept is illustrated in Figure 6.8. It identifies the boundaries of each word in the original text and returns a list containing the word IDs corresponding to the original words in the input text. These word IDs allow for mapping between individual words in the text and their respective positions in the tokenized sequence. For each word, the function calculates its surprisal value by summing the negative logarithm of the predicted probability of each token in the word. Finally, the function returns a tuple containing the word list and corresponding surprisal values.

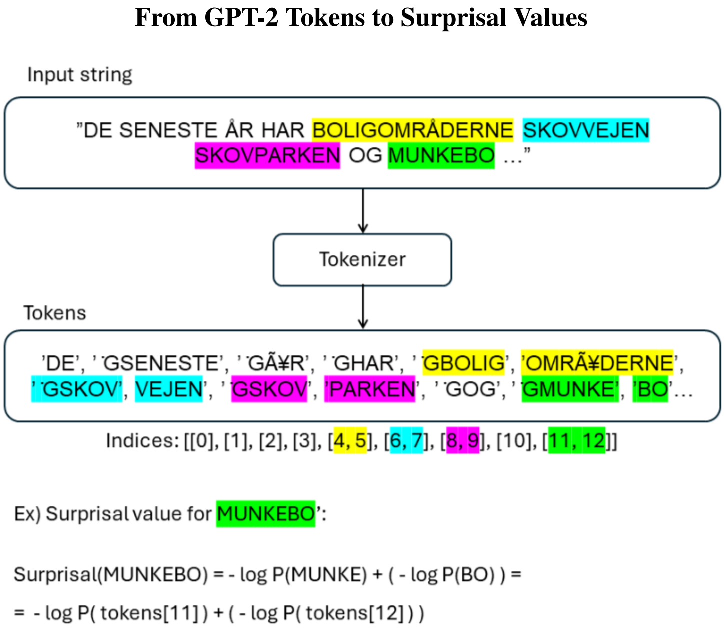

From GPT-2 Tokens to Surprisal Values
Input string
"DE SENESTE ÅR HAR BOLIGOMRÅDERNE SKOVVEJEN SKOVPARKEN OG MUNKEBO …"
Tokenizer
Tokens
'DE', ' ˙GSENESTE', ' ˙GÃ¥R', ' ˙GHAR', ' ˙GBOLIG', 'OMRÃ¥DERNE', ' ˙GSKOV', 'VEJEN', ' ˙GSKOV', 'PARKEN', ' ˙GOG', ' ˙GMUNKE', 'BO'…
Indices: [[0], [1], [2], [3], [4, 5], [6, 7], [8, 9], [10], [11, 12]]
Ex) Surprisal value for MUNKEBO':
Surprisal(MUNKEBO) = - log P(MUNKE) + (- log P(BO)) =
= - log P(tokens[11]) + (- log P(tokens[12]))

Figure 6.8 Example of surprisal calculation for words that had been split during the tokenization.

To align the surprisal values with the EEG signal, they were synchronized with the word onset times. In this alignment, each surprisal value corresponds to an instance precisely at the onset time of the word. In Figure 6.9, an example of surprisal is plotted, where the vertical lines represent the onset of a word and the amplitude is the level of surprise. At all other time points, the values are set to zero. The same was done for the word list. The surprisal data was then saved along with the acoustic features to be used as predictor variables for the linear models.

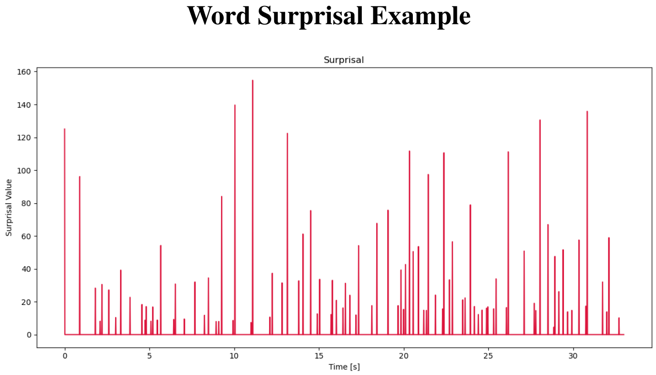


Figure 6.9 Example of word surprisals for one trial.

Frequency Content. Research has revealed that semantic information, such as word surprisal, is encoded in EEG activity at higher frequency ranges compared to acoustic information, notably up to 30 Hz [Zion Golumbic et al., 2013]. To determine the best-suited preprocessing methods for EEG and surprisal signals, an analysis was conducted involving various surprisal TRF models and their correlation with EEG data. This analysis utilized four sets of EEG signals, each processed differently: one underwent bandpass filtering between 1 Hz to 30 Hz, another between 1 Hz to 8 Hz, a third between 0 Hz to 8 Hz, and the fourth underwent highpass filtering above 1 Hz followed by resampling to 32 Hz. Similarly, four sets of the surprisal signal were prepared: one remained unfiltered, another was bandpass filtered between 1 Hz to 8 Hz, a third between 0 Hz to 8 Hz, and the fourth was highpass filtered above 1 Hz before resampling to 32 Hz. The signals that weren't resampled for the analysis had a sampling frequency of 64 Hz. The resulting combinations, as outlined in Table 6.1, were used to train TRF models, and their mean correlation values were subsequently compared.

Table 6.1 Different combinations of EEG signal and word surprisal as input when training surprisal TRF models. The EEG signal and word surprisal stimulus were preprocessed in different ways through filtration and resampling.

model nr.	Word Surprisals	EEG signal
1	Unprocessed	Bandpass filtered 1-30 Hz
2	Unprocessed	Bandpass filtered 1-8 Hz
3	Bandpass filtered 1-8 Hz	Bandpass filtered 1-8 Hz
4	Bandpass filtered 0-8 Hz	Bandpass filtered 0-8 Hz
5	Highpass filtered >1 Hz, resampled to 30 Hz	Highpass filtered >1 Hz, resampled to 30 Hz

6.3 Linear Model

The Boosting Algorithm in Eelbrain was used to train TRF- and mTRF models of different predictor variable combinations and the EEG signal. It is resilient to overfitting and favors sparsity in the TRFs as well as implements an early stopping strategy [Brodbeck et al., 2021].

The data was divided into 10 partitions and reserving 10% of the data for validation and 10% for training, see Figure 6.10 for reference. When using multiple predictors, the selective stopping option was used to stop the training when a single predictor would eventually start to overfit.

The time window was set from -100 ms to 400 ms for all models, except those that include the surprisal predictor which instead used a time window from -100 ms to 700 ms. The larger time window was used for surprisal as it has been proven that responses related to semantics are processed later in the brain than acoustic stimuli. The basis function is a meta-parameter that sets the length of the basis of windows for the kernel. A longer basis makes the TRF more smooth and realistic in shape to neural signals. This parameter was set to 100 ms in the algorithm.

Partitions in Cross-validation

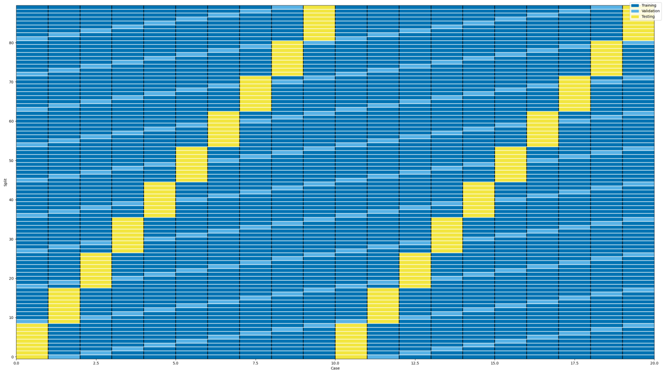


Figure 6.10 Partitioning of the data for cross-validation using data from one session (20 trials). The data have been divided into 10 partitions. The test folds are depicted in yellow, the validation folds in light blue, and the training folds in dark blue. Each fold serves as the test set once (two folds per cycle) while the remaining data is used for training, reserving one segment from each training fold for validation. Case (x-axis) refers to trial in this context.

Eleven different models were trained by combining the different predictor variables. Four TRF models were calculated, one for each predictor variable: envelopes, phonetic features, word onsets, and surprisals. Three mTRF models had a combination of two predictor variables, with one being the envelopes. Another three mTRFs included a combination of three different predictors (all including the envelopes), and finally, one mTRF model that combined all four predictors. Additionally, four surprisal TRFs were trained using surprisal and EEG data filtered and resampled in different frequency ranges to find the best combination. The data used for training each model included data from all trials, excluding the testing set, within one session for one subject, using either the features from the target or the masker speaker.

6.4 Statistics

Correlation

The performance of the models was defined by the Pearson correlation (r) between the measured EEG response (y) and the predicted response ($\hat{y}$). The Pearson correlation coefficient measures the linear relationship between two variables, ranging from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

$$\text{Pearson correlation: r} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \quad (6.1)$$

Here, y_i represents the individual measured EEG responses, $\hat{y}_i$ represents the corresponding predicted responses, $\bar{y}$ is the mean of the measured responses, $\bar{\hat{y}}$ is the mean of the predicted responses, and n is the number of data points. The equation calculates the Pearson correlation coefficient by comparing the deviations of each data point from their respective means, and normalizing by the standard deviations.

Hearing Aid Conditions

The four hearing aid conditions represent combinations of two different hearing aids with two different NR settings:

- Condition 1: Hearing Aid 1 with Noise Reduction OFF (HA 1, NR OFF)
- Condition 2: Hearing Aid 1 with Noise Reduction ON (HA 1, NR ON)
- Condition 3: Hearing Aid 2 with Noise Reduction OFF (HA 2, NR OFF)
- Condition 4: Hearing Aid 2 with Noise Reduction ON (HA 2, NR ON)

Data Analysis

In this thesis, models trained using audio stimuli from the target speech streams as the predictor are referred to as ’Target models’ (e.g., Target Envelope model). Similarly, models trained using data from the masker speech streams are referred to as ’Masker models’ (e.g., Masker Envelope model). Both target and masker models for a given session are trained on the same EEG data.

Model Hierarchy

Each subject (28 total) has eight different models per predictor variable (TRF models) and per predictor variable combination (mTRF models). The eight models are then divided into two speaker categories: target and masker, depending on which speaker data they were trained on. One model was created for each session. There are four different sessions, and each session uses a randomized condition. Each model was trained on data from the 20 trials recorded in the respective session. Figure 6.11 illustrates the hierarchy of the model structure, for one subject and predictor variable.

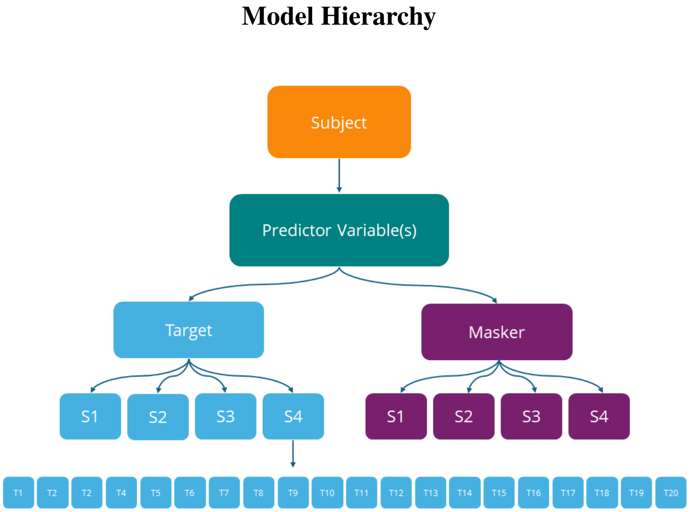

Model Hierarchy

Figure 6.11 Schematic of how the data and models are structured. Every model is trained on 80% of the trials from one session (S1-4), using either the target (blue) or the masker (purple) speech features as predictors of the EEG. This is done for every predictor variable (or combination thereof) and for every subject individually.

ANOVA and Tukey's Tests

Statistical analyses were performed using the Python toolbox statsmodels [Seabold and Perktold, 2010], specifically employing anova_lm for one-way and two-way ANOVAs. Analysis of Variance (ANOVA) is a statistical method used to analyze the differences among group means in a sample. It tests the null hypothesis that the means of several groups are equal against the alternative hypothesis that at least one group mean is different.

In this study, ANOVA was used to investigate potential effects of the different factors condition, speaker, and predtictor variable(s) on the dependent variable, which was the correlation data (r). By examining the variance between and within groups, ANOVA helps determine whether there are statistically significant differences in means among these groups.

All EEG measures were assumed to be independent, and the dependent variable approximated a normal distribution. Furthermore, the correlation data for each condition exhibited homogeneous variances. Despite multiple measurements from the same individual, each observation was considered independent due to variations in stimuli across measurements.

Initially, the mean correlation for all models across all conditions, stratified by masker and target, was computed. Subsequently, to explore potential effects on condition, speaker, and model, one- and two-way ANOVAs were conducted, with different factors fixed. In cases where significance was observed, post-hoc Tukey's Honest Significant Difference (HSD) tests were employed to explore pairwise differences between conditions or models, depending on the dependent variable in the ANOVA test. Tukey's test is a method used for comparing all possible pairs of means to determine which means are significantly different from each other after finding a significant result in ANOVA. It adjusts for multiple comparisons to control the family-wise error rate.

6.5 Structure of Analysis

Performance of all TRF and mTRF Models

The analysis comprised multiple tests to examine the relationships between different models, attended speaker, conditions and subjects. Firstly, the results for all models were computed and visualized using bar plots or box plots. The models were compared based on their correlation values, averaged across the selected sensors illustrated in Figure 6.3. The conditions were grouped in different ways: all conditions together, one by one, or as ON and OFF, where ON/OFF indicate whether an NR scheme was employed. After this initial analysis, a fourth group was formed using conditions 2, 3, and 4, while condition 1 was omitted, see Section 6.4 *Hearing Aid Conditions*.

Secondly, the models were tested for significant differences using ANOVA and post-hoc Tukey's HSD tests. The single predictor models, i.e., TRFs, were compared. The mTRF models, all including the envelope as a predictor, were then compared against the envelope TRF to determine if the added predictor variables contributed to the model's correlation. The models were averaged across all subjects and conditions (excluding condition 1) to view the average impulse response in butterfly plots and the sensor and connectivity in topographic maps.

Word Surprisal and Semantic Processing

The next part of the analysis focused on the Surprisal models to evaluate the performance of word surprisal as a predictor of EEG for hearing-impaired individuals, and whether they behaved differently based on attended speaker and hearing aid settings. Initially, different preprocessing methods for both the stimulus and response signals were analyzed to find the best-suited combination for model training. The Surprisal models were then compared by their correlation values by speaker and condition. An additional analysis of condition 1 and 4 for the Target models was done by analyzing the average TRF responses and topographic maps.

Subject-Specific Neural Responses

Given that EEG data and speech processing can be highly individual, especially for HI subjects, significant information can be lost when averaging across all subjects. Therefore, the correlations for the four TRF models — including Envelope, WordOnset, PhoneticFeatures, and Surprisal models — were plotted per subject. Butterfly plots of the TRF (i.e., the impulse response) for the best-performing subject were generated for the four TRF models, along with topographic maps of the EEG sensor responses. From these results, the six models with the highest and lowest correlations for the envelope model were selected for further analysis of the TRF response.

7

Results

7.1 Interpretation of TRF plots and Topographic maps

The interpretation of TRF plots and topographic maps provides crucial insights into neural responses to acoustic stimuli. Below are detailed explanations for both types of visualizations used in this study.

TRF plot
In the TRF plot:

- **Y-Axis**: Represents the amplitude of the neural response, which is time-locked to the predictor variable. The responses are normalized for easier comparison.
- **X-Axis**: Represents time in milliseconds (ms), with the zero point indicating the onset of the predictor variable.
- **Colored Graphs**: Show TRFs for different EEG sensors (channels). Each line illustrates how the neural response at a specific electrode evolves over time relative to the predictor variable.

Topographic Maps
The topographic map complements the TRF plot by showing the spatial distribution of the neural response at a specific time point (indicated by the black vertical line in the plot):

- **Red Areas:** Indicate regions with a strong positive response.
- **Blue Areas:** Indicate regions with strong negative responses.
- **White to Lighter Areas:** Indicate weaker responses.
- **Blue Dots:** Represent EEG sensor positions.

7.2 Performance of All TRF and mTRF Models

This section provides an overview of the performance of the TRF and mTRF models in capturing neural responses to speech features. The analysis examines mean correlation values across all subjects and conditions, highlighting the differences in model performance between Target and Masker models for each predictor variable. Additionally, the impact of hearing aid conditions on these correlations is explored, with comparisons made between different hearing aid conditions. The results are visualized using bar plots and box plots to illustrate significant differences and trends across the models and conditions. Furthermore, averaged TRFs and topographic maps for Target and Masker models across are presented to provide insights into the temporal dynamics and spatial distribution of neural responses.

Mean Correlation

The mean correlation values for all Target and Masker models were computed and plotted separately, covering all subjects and conditions. The result is presented in Figure 7.1, for more details, see Tables 10.9(target) and 10.10(masker) in Appendix 10.1. The speaker, determined by whether the model was trained on target or masker speech data, significantly influences mean correlation values across different models, with consistently higher correlations observed for the target speaker ($p < 0.0001$ for each model). On average, correlations increase by approximately 527% from the masker to the target across all models.

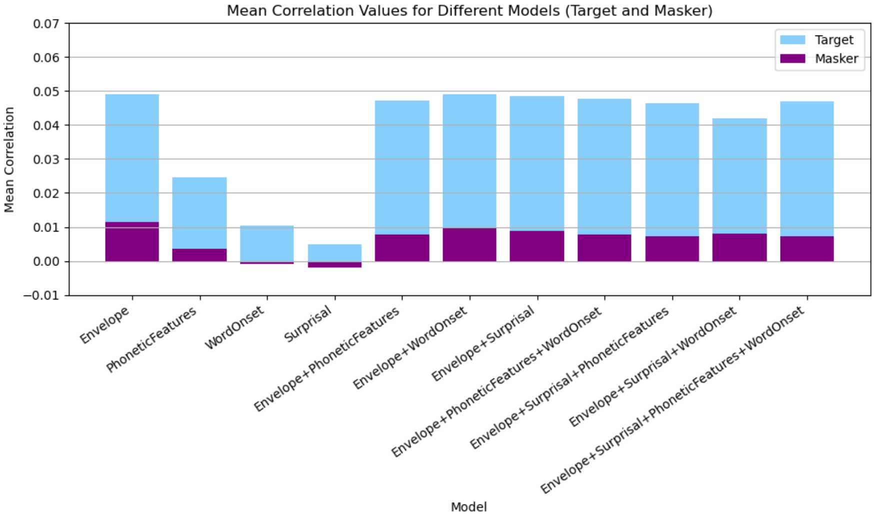
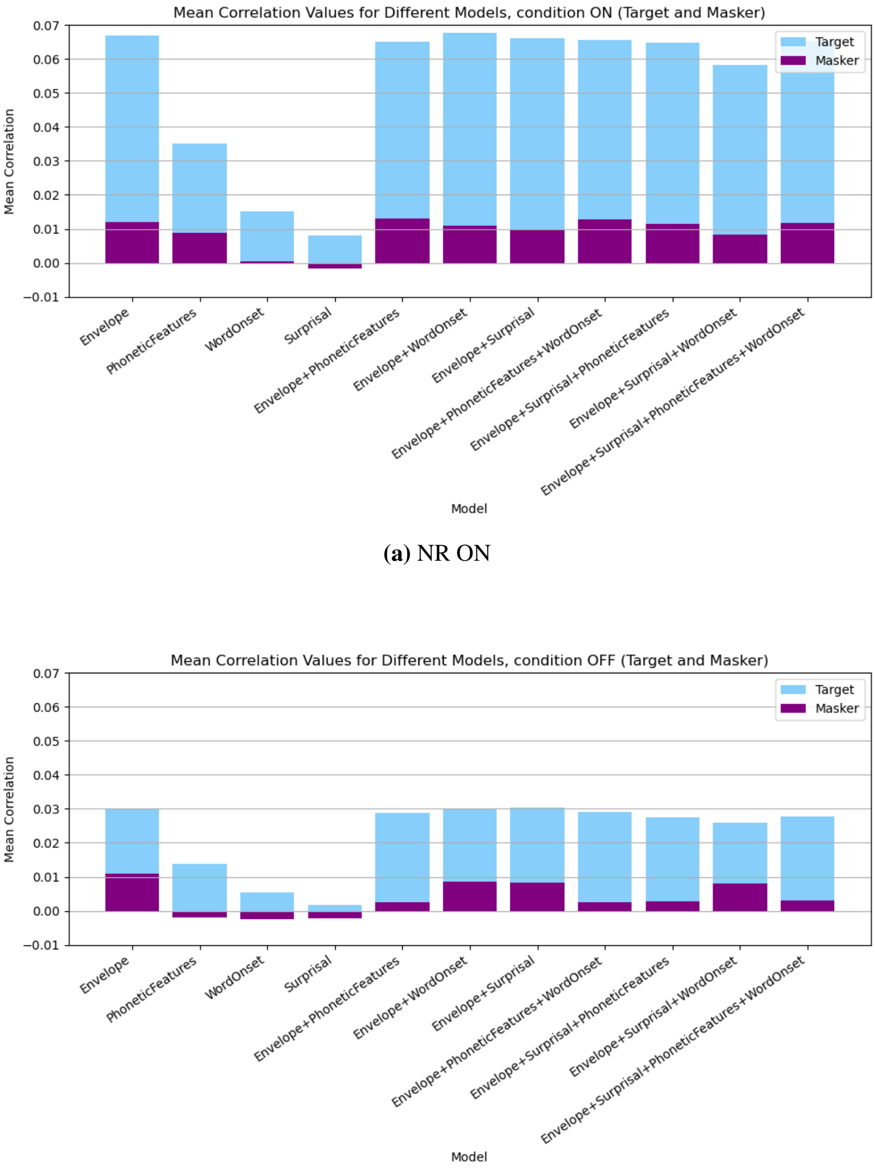


Figure 7.1 Bar plot illustrating the mean correlation values for all models, encompassing all subjects and hearing aid conditions. Mean correlation values for Target models are represented in blue, while those for Masker models are shown in purple.

Comparing Conditions: ON/OFF

To explore whether hearing aid conditions had a significant impact on mean correlation values, the first analysis involved grouping the conditions into ON (including both hearing aids with NR ON) and OFF (including both hearing aids with NR OFF). Bar plots of these grouped conditions are shown below in Figures 7.2a and 7.2b, and a box plot is shown in Figure 7.3.

Mean Correlation - NR OFF/ON



(a) NR ON

(b) NR OFF

Figure 7.2 Bar plots illustrating the mean correlation values for all models with NR ON and OFF. Mean correlation values for Target models are represented in blue, while those for Masker models are shown in purple.

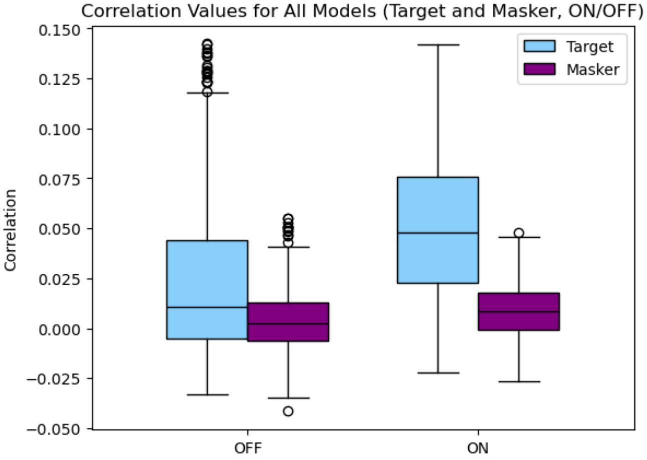


Figure 7.3 Box plot illustrating the significant differences in correlation values between Target and Masker models across the grouped NR conditions OFF and ON. Mean correlation values for all subjects and models have been used. Target correlations are shown in blue, Masker correlations in purple. Circles indicate subjects identified as outliers. The noise reduction algorithm (NR ON) provides a significant improvement in Target correlation, without any similar, unwanted, increase in Masker correlation.

One-way ANOVAs, with the speaker as the independent factor and the condition (ON/OFF) held constant, revealed a significant speaker effect on correlation values across all models in the ON condition ($p < 0.001$), see Figures 7.2a and 7.3. In the OFF condition (see Figure 7.2b) the speaker effect was insignificant for the Surprisal model, significant for the WordOnset model ($p < 0.01$), and significant for the rest of the models ($p < 0.0001$). P-values of all models can be seen in Table 10.5 in Appendix 10.1

Comparing Conditions: 1-4

To investigate each hearing aid condition separately, one-way ANOVAs were conducted with the speaker as the independent factor, followed by pairwise post-hoc Tukey tests. A box plot of the four conditions, for both Target and Masker models, is shown in Figure 7.4.

The analysis revealed a significant speaker effect (i.e. target vs masker) for all models in all conditions ($p < 0.001$ for all models, except for the Surprisal model in conditions 2 and 3 where $p < 0.05$), see Table 10.5 in Appendix 10.1. Interestingly, in condition 1, the speaker did not demonstrate a significant effect for any model except for the Envelope model, indicating a unique pattern compared to the other conditions.

The pairwise Post- hoc Tukey tests revealed significant differences between conditions 1-2, 1-3 and 1-4 for all Target and Masker models ($p < 0.05$) with condition 1 consistently yielding lower mean correlation values. For the Target models, there were no significant differences between the rest of the condition combinations, see Table 10.7 in Appendix 10.1. For the Masker models, there were additional significant differences between conditions 2-3 and 2-4 ($p < 0.05$), where both conditions 3 and 4 had higher correlation than condition 2, see Table 10.8 in Appendix 10.1.

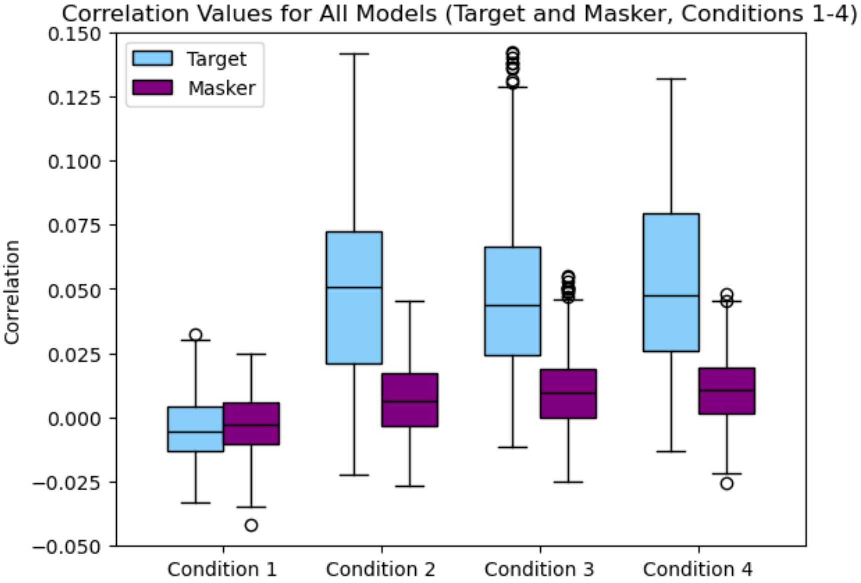


Figure 7.4 Box plot illustrating the significant differences in correlation values between Target and Masker models for conditions 1,2,3 and 4. Mean correlation values for all subjects and models have been used. Condition 1 (Hearing Aid 1 with Noise Reduction OFF) performs significantly worse than the other conditions.

Removing condition 1

Following the findings from previous results (Figure 7.4), where the mean correlation for condition 1 was significantly lower compared to other conditions, condition 1 was subsequently removed. New mean correlation values for each model were then recalculated and visualized in Figure 7.5. After removing condition 1 from the dataset, there was an average increase in mean correlation for all models of approximately 39%. Given this improvement, condition 1 was excluded for further analysis.

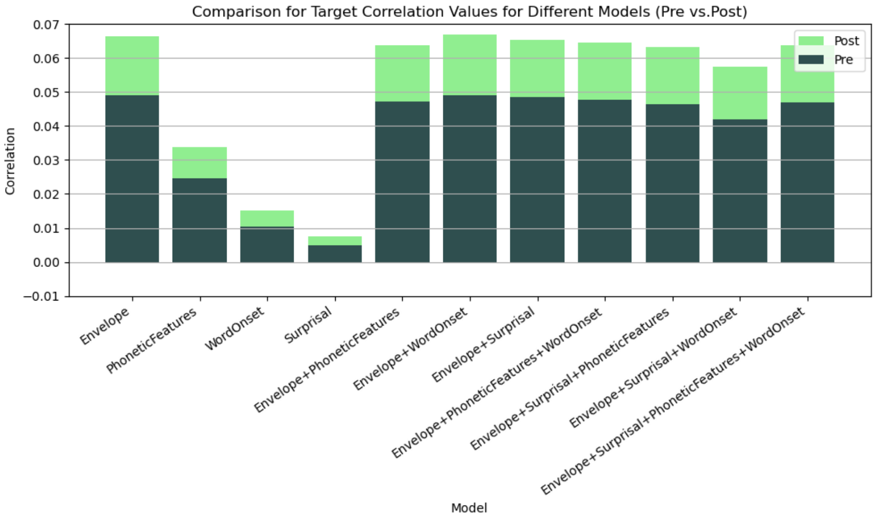


Figure 7.5 Bar plot illustrating the difference in mean correlation values for all models, including and excluding condition 1. Mean correlation for all models, encompassing conditions 1,2,3 and 4, are depicted in dark green and denoted as “Pre” (note that this is the same result as in Figure 7.1). Mean correlations for all models after removing condition 1 are depicted in light green, denoted as ‘Post’.

Model Performance Comparison

To evaluate the performance of the TRF and mTRF models, their mean correlations were compared using one-way ANOVA tests to determine if there were significant differences among the models. The models used for the analysis were the Target models trained on data from conditions 2, 3, and 4 lumped together. In the One-way ANOVA, the correlation were the dependent variable, while the model type was the independent variable.

First, the single predictor variable TRF models, trained on the acoustic envelopes, word surprisals, phonetic features and word onsets, were compared in pairs; for example evaluating the Envelope model vs. the Surprisal model. The same methodology was used for analyzing the mTRF models. To determine if incorporating multiple predictors enhanced the model’s mean correlation, each mTRF model was compared to the Envelope model.

The TRF response and topographic maps, averaged across all subjects and conditions 2-4, were plotted, see Figure 7.6, and analyzed to provide an overview of the model performance.

One-way ANOVA tests proved that all single predictor models were significantly different ($p < 10^{-6}$), see Tables 10.1, 10.2 and 10.3 in Appendix 10.1. The best-performing TRF models, based on mean correlation values across subjects, were Envelope, PhoneticFeatures, WordOnset, and Surprisal, in descending order.

None of the mTRF models were found to be significantly different from the Envelope model ($p > 0.05$, see Table 10.4 in Appendix 10.1. Incorporating other predictor variables therefore did not improve the performance of the Envelope model. This result motivates the decision to exclude the mTRF models for further analysis and instead focus on the TRF models.

Average Response - TRF models

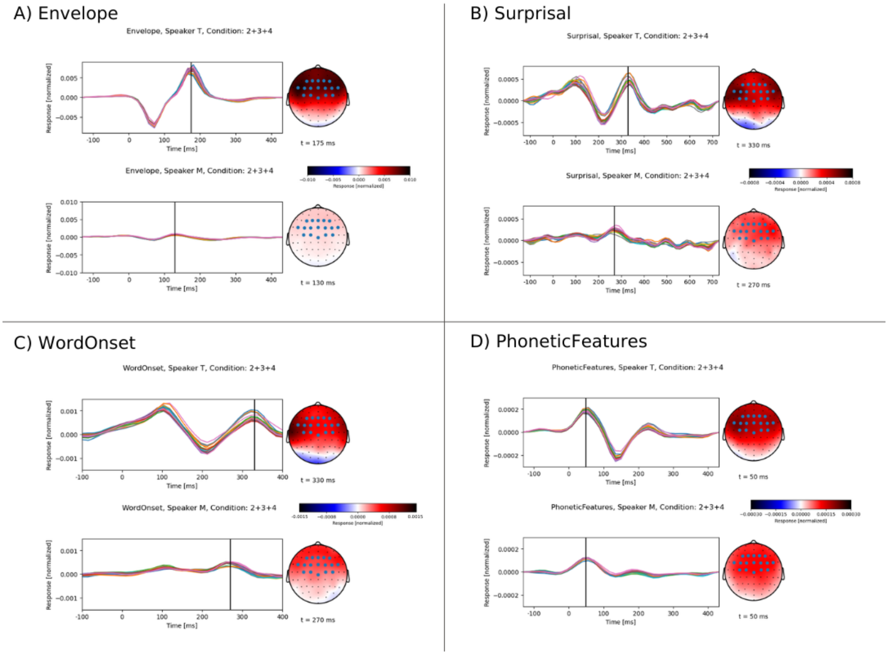


Figure 7.6 TRFs showing how the different one-predictor variable models (Target and Masker) capture the neural response to their respective speech features. The TRFs are averaged across all subjects and conditions, excluding condition 1. Each line represents one of the selected sensors (17 total), marked in blue in the sensor maps. The plots, which are on different scales, include A) Envelope, B) Surprisal, C) WordOnset, and D) PhoneticFeatures. The topographic maps indicate the situation at the time marked by the vertical lines in the corresponding plot.

A) Envelope. In Figure 7.6 A, for the Target model, the TRF revealed a negative peak at around 70 ms, followed by a positive peak at approximately 175 ms latency. These peaks likely correspond to the N100 and P200 components, respectively. The peak amplitudes were significantly lower for the Masker model, with the positive peak appearing somewhat earlier at approximately 130 ms latency. The lower amplitude indicates a weaker neural response, suggesting that the masker speech is suppressed. The topographic map, indicates concentrated positive correlation in the fronto-central and temporal regions of the brain for the Target model at 175 ms latency. The Masker Envelope model elicited a weaker and more diffuse response pattern than the target, indicating less neural reaction to the masker speech.

B) Surprisal. Surprisal responses were significantly weaker than the envelope responses, having roughly 10 times lower amplitudes. Two positive peaks were observed in the TRF of the Target model, one around 100 ms and another around 330 ms latency (see Figure 7.6 B), likely corresponding to the P100 and P300 components. Additionally, a negative peak at around 200 ms was noted. The peak amplitude for the Masker model TRF was lower, with less clear peaks. A potential positive peak could be distinguished at 300 ms, possibly a P300 component. In the topographic map corresponding to the peak at 330 ms (Target model) a strong positive response could be seen in the fronto-central and temporal brain regions. The response was weaker and more diffused in the topographic map corresponding to the Masker model.

C) WordOnset. In Figure 7.6 C), Two positive peaks appeared at approximately 100 ms and 300 ms for the Target model, and a negative peak can be seen at 200 ms, possibly corresponding to the P100, P300, and N200 components, respectively. This pattern was similar to the Surprisal model, reflecting the shared onset times. The peak amplitude was lower for the Masker model, with a positive peak at 270 ms, appearing earlier than the Target's P300 peak at 330 ms. The neural activity at the P300 peak was positively correlated in the fronto-central and temporal brain regions.

D) Phonetic Features. For the Target model TRF, an early positive peak was observed at 50 ms, possibly linked to the P100 component, followed by a negative peak at approximately 130 ms and a small positive peak around 230 ms, see Figure 7.6 D. The amplitude was lower for the Masker model TRF, with the first peak occurring at the same latency as the Target. Noticeable positive activity was present in the bilateral fronto-central brain regions during the initial positive peak in both the Target and Masker TRFs, although the response was weaker for the Masker model.

7.3 Word Surprisal and Semantic Processing

This section investigates the performance of word surprisal as a predictor variable in TRF models and examines how different preprocessing methods and hearing aid conditions affect these models. The analysis begins by comparing different preprocessing methods for surprisal TRFs to determine their impact on model performance. Following this, a comparative analysis between Surprisal and Word Onset

TRFs is conducted to explore the differences in neural encoding. Subsequently, the significance of various hearing aid conditions, including grouped ON/OFF conditions and individual conditions (1-4), is evaluated using statistical tests. The results are visualized using box plots, TRF plots, and topographic maps to illustrate differences in neural encoding of word surprisal across conditions.

Preprocessing and Modelling of Surprisal TRF

To evaluate how different EEG frequency ranges and processing methods affect the performance of word surprisal as a predictor variable in TRF models, different Surprisal models were trained using various versions of the stimulus and response in the training data (see Table 6.1 in Section 6.2). Subsequently, t-tests were conducted to examine whether the different Surprisal models significantly affected the correlation values with the EEG, hence testing the null hypothesis that the two models have identical average values.

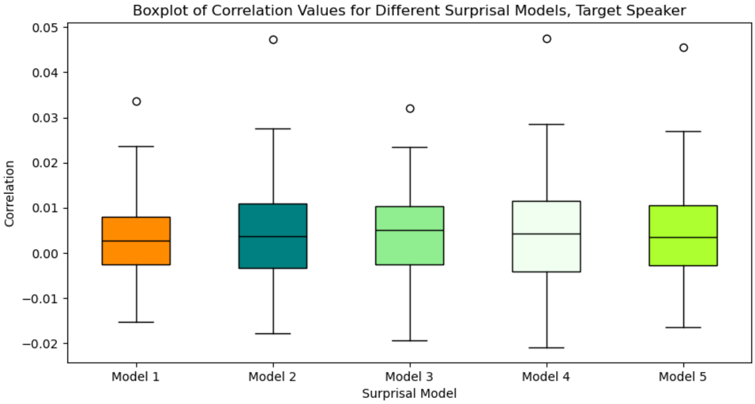


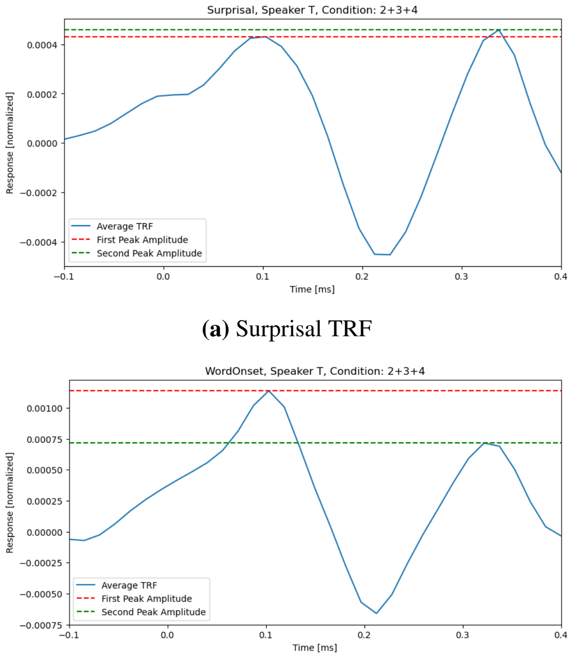
Figure 7.7 Boxplot illustrating mean correlation values for the five different word surprisal TRF models, for the target speaker as stimulus. The models are plotted in the same order as outlined in Table 6.1. No significant difference can be seen between the different methods to handle the surprisal features, models 1-5.

No significant differences in the correlation values were found between the Surprisal models for either the Target or Masker model. The similarities between the models can be observed in Figure 7.7. Therefore, for subsequent analyses, model number 2 was used, involving unfiltered word surprisals and a low-frequency EEG signal bandpass-filtered between 1-8 Hz, both sampled at 64 Hz. This choice was made to maintain consistency across the models, so that they were all trained on the same EEG signal, and to enhance reproducibility by utilizing unprocessed surprisals.

TRF Peak Analysis: Surprisal versus WordOnset

Although word surprisal is a scaled version of word onsets, examining these features separately allows for the identification of potential differences in neural encoding. Therefore, the differences in peak amplitudes between the P100 and P300 peaks in their respective TRFs, (see Figure 7.6 B and C) were compared. For this analysis, the TRFs were averaged across all channels to obtain a representative response for each model, as shown in Figure 7.8.

TRF Peak Comparison - WordOnset and Surprisal Target Models



(a) Surprisal TRF

(b) WordOnset TRF

Figure 7.8 Comparison of the TRF peak amplitude differences for the Surprisal and WordOnset models. The P300 peak is higher than the P100 peak in the Surprisal TRF. The opposite is observed for the WordOnset TRF.

When comparing the TRFs of Word Onset and Word Surprisal, a key observation is the relative amplitude differences between the P100 and P300 peaks. For the Word Surprisal model, the second peak (P300) is higher than the first (P100), whereas for the WordOnset model, the second peak is lower.

Significance of Condition for Surprisal Models

The different hearing aid conditions, including the grouped ON/OFF conditions as well as all condition separately, were evaluated for the Surprisal model to investigate their effect on the correlation for both Target and Masker Surprisal models. One-Way ANOVA tests, followed by pairwise Tukey's HSD (Honestly Significant Difference) post-hoc tests, were utilized to determine differences between conditions 1-4. Data from all subjects were used in the tests and the significance level was set to 0.05. In the ANOVA test, the correlation was the dependent factor, speaker was the fixed factor, and condition was the independent factor.

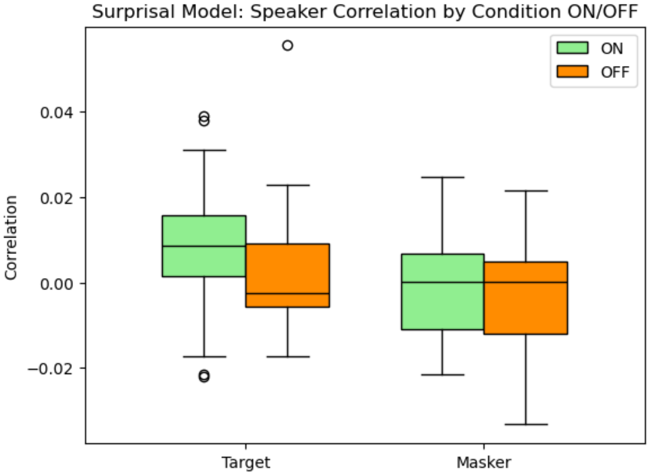


Figure 7.9 Boxplot illustrating the correlation of the Target and Masker Surprisal model based on data from all subjects. The data is divided by Target and Masker model, and grouped conditions ON and OFF.

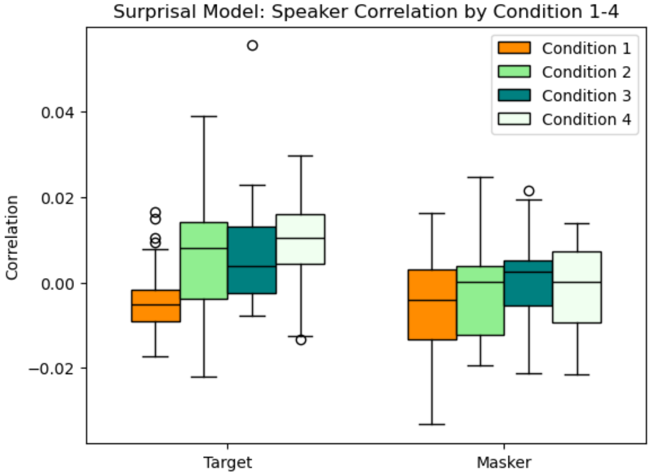


Figure 7.10 Boxplot illustrating the correlation of the Target and Masker Surprisal model based on data from all subjects. The data is divided by Target and Masker model, and conditions 1, 2, 3, and 4.

Significance of Condition: Surprisal Models

Condition:	ON/OFF	1-4
Target	0.011	0.0007
Masker	0.78	0.42

Table 7.1 One-Way ANOVA - Fixed effects: Model (Surprisal), Speaker; Independent factor: Condition; Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for Target and Masker. There was a significant difference in the Target models but not in the Masker models ($\alpha < 0.05$).

The findings from the one-way ANOVA show significant effects for the grouped ON/OFF conditions (Table 7.1 and Figure 7.9). Specifically, NR significantly impacts the correlation for the Target models, although not for the Masker models.

Additionally, a notable distinction was observed between condition 1 and the remaining three conditions, but no significant differences were detected among conditions 2, 3, and 4 for the Target models, see Figure 7.10. Post-hoc Tukey results are shown in Table 10.11 in Appendix 10.2. The conditions showed no significant impact for the Masker models, see Table 7.1. This suggests that condition 1 is the sole condition that diverges significantly from the others.

Surprisal Model Performance: Conditions 1 vs. 4

To investigate the effect of conditions 1 and 4 on correlation values for the Surprisal model, TRFs and topographic maps were plotted for the Target models using data from conditions 1 and 4. The results, averaged across all subjects, can be viewed in Figure 7.11.

Target Surprisal Model - Condition 1 vs. 4

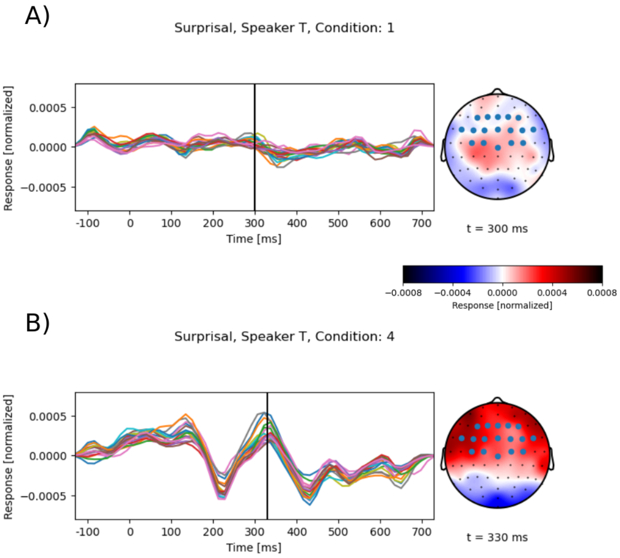


Figure 7.11 TRF plots and topographic maps of the neurological response to the word surprisal speech feature, comparing A) condition 1 (HA 1, NR OFF) and B) condition 4 (HA 2, NR ON).

In condition 4, there is a prominent positive peak around 300 ms (P300), and two negative peaks at approximately 200 and 400 ms (Figure 7.11 B). In condition 1, there are no prominent peaks and the response is more noise-like in behavior. This indicates a stronger and more coherent correlation between the Surprisal model and EEG in condition 4 compared to condition 1, as seen in the topographic maps, particularly in the frontal, central, and temporal regions.

7.4 Subject-Specific Neural Responses

This section explores the variability in neural responses to speech features across individual subjects and different hearing aid conditions. By analyzing mean correlation values for each subject, model, and condition, this section aims to highlight the individual differences in EEG signals.

All Subjects

Additional analysis to compare individual results was conducted. This involved illustrating bar plots, see Figures 7.12, 7.13, 7.14 and 7.15, showcasing the mean correlation for each subject, model, and condition.

Subject ID1030 was found to have two sessions recorded under condition 2, with no session documented for condition 1. This discrepancy is likely due to a documentation error, though it is unknown which condition is which. Consequently, the mean correlation from the two ’condition 2’ sessions was used, and condition 1 was omitted from the plots for this subject. Thus, the difference in correlation between condition 1 and condition 2 remains uncertain.

Mean Correlation by Subject and Condition: Target Envelope Model

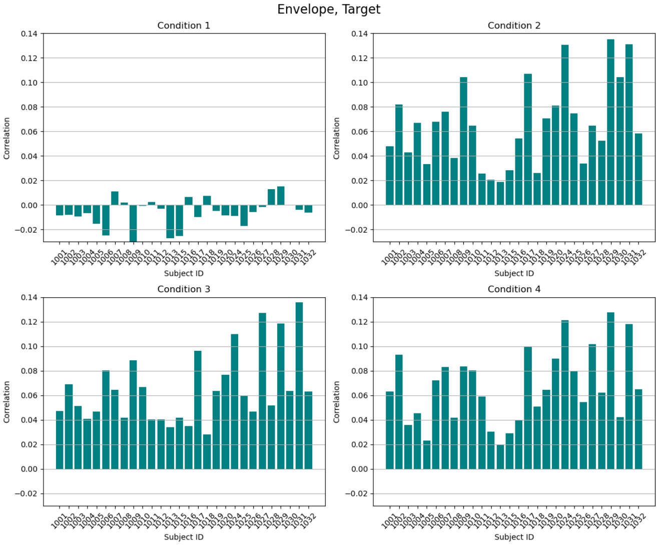


Figure 7.12 Mean correlation per subject and condition for the Target Envelope model. 27/28 subjects had higher mean correlation values in condition 2 compared to condition 1, i.e. when the NR for hearing aid 1 was ON vs. OFF. 18/28 subjects had higher mean correlation values in condition 4 compared to condition 3, i.e. when the NR for hearing aid 2 was ON vs. OFF. One subject (ID1030) had missing values in condition 1.

Mean Correlation by Subject and Condition: Target Surprisal Model

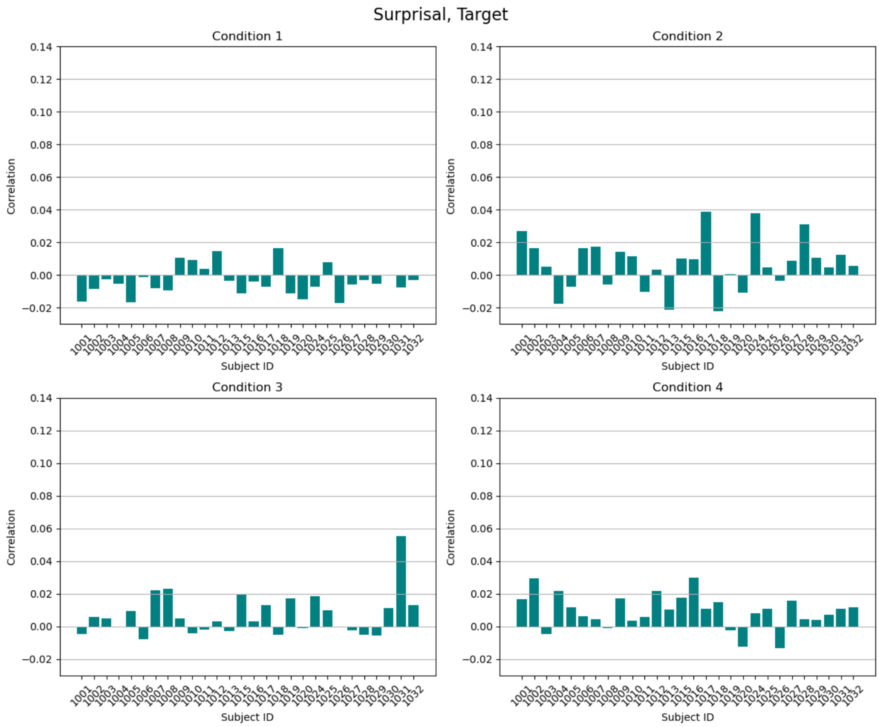


Figure 7.13 Mean correlation per subject and condition for the Target Surprisal model. 21/28 subjects had higher mean correlation values in condition 2 compared to condition 1, i.e. when the NR for hearing aid 1 was ON vs. OFF. 16/28 subjects had higher mean correlation values in condition 4 compared to condition 3, i.e. when the NR for hearing aid 2 was ON vs. OFF. One subject (ID1030) had missing values in condition 1.

Mean Correlation by Subject and Condition: Target WordOnset Model

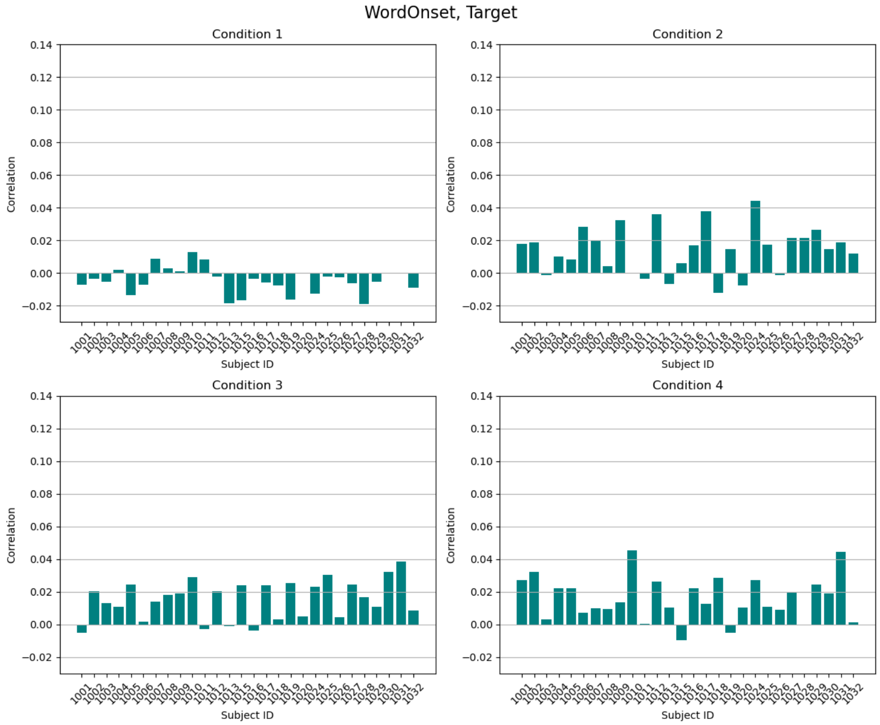


Figure 7.14 Mean correlation per subject and condition for the Target WordOnset model. 22/28 subjects had higher mean correlation values in condition 2 compared to condition 1, i.e. when the NR for hearing aid 1 was ON vs. OFF. 15/28 subjects had higher mean correlation values in condition 4 compared to condition 3, i.e. when the NR for hearing aid 2 was ON vs. OFF. One subject (ID1030) had missing values in condition 1.

Mean Correlation by Subject and Condition: Target PhoneticFeatures Model

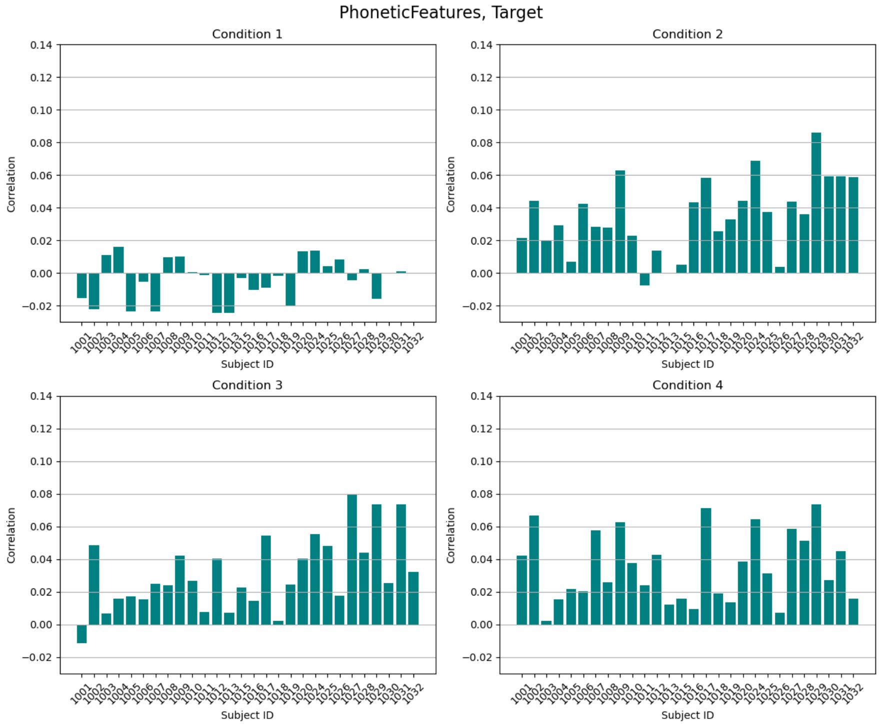


Figure 7.15 Mean correlation per subject and condition for the Target PhoneticFeatures model. 22/28 subjects had higher correlations in condition 2 compared to condition 1, i.e. when the NR for hearing aid 1 was ON vs. OFF. 15/28 subjects had higher mean correlation values in condition 4 compared to condition 3, i.e. when the NR for hearing aid 2 was ON vs. OFF. One subject (ID1030) had missing values in condition 1.

The Envelope model yielded the highest mean correlation values in condition 2, 3 and 4, across all subjects, with no negative values observed (see Figure 7.12). It was followed by the PhoneticFeatures model (see Figure 7.15), the WordOnset model (see Figure 7.14), and lastly, the Surprisal model (see Figure 7.13). Condition 1 displayed noticeably lower mean correlation values on an individual level compared to the other conditions. The gap between NR settings was larger for HA 1 (conditions 1 and 2) than for HA 2 (conditions 3 and 4), as condition 1 belongs to HA 1. For the Envelope model, the individual results were consistent across conditions 2-4, whereas the results varied more between conditions for the other models.

Comparing Strong and Weak Response Subjects

To analyze and illustrate the diversity in the dataset and how differently the subjects' brains process auditory input, three strong response (SR) subjects (with high correlation values) and three weak response (WR) subjects (with low correlation values) were selected based on their performance with the Target Envelope model in condition 4. The TRFs and corresponding topographic maps for each of these six subjects are shown in Figure 7.16, with the same scale used to facilitate magnitude comparison.

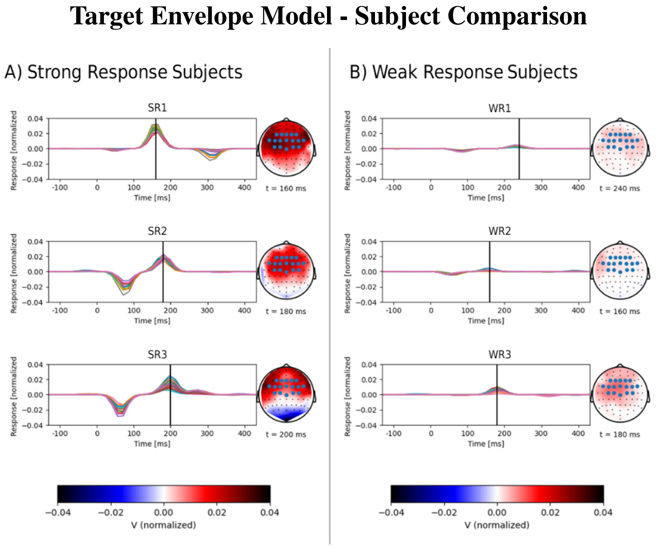


Figure 7.16 The figure shows the TRFs and corresponding topographic maps for A) three strong response subjects and B) three weak response subjects.

Figure 7.16 illustrates how different individuals' brains respond to the Target Envelope stimulus using each subject's Target Envelope model in condition 4. The SR subjects exhibit a prominent positive peak with a maximum magnitude within the 160-200 ms latency range. These strong positive responses are localized in the fronto-central region of the brain.

In contrast, the WR subjects display a smaller, less distinct peak with a maximum amplitude within the 160-240 ms latency range. The topographic maps of these subjects reveal a weaker and more diffuse response, yet still localized in the fronto-central brain region.

8

Discussion

8.1 Performance of all TRF and mTRF models

Evaluating grouped conditions ON/OFF

The significant speaker effect suggests that there are discernible differences in how the EEG represents the features of the Target compared to the Masker model. Similarly, the significant condition effect indicates that the activation of NR schemes has a notable impact on correlation across models for both Target and Masker speakers ($p<0.01$). The significant difference in correlation between the ON and OFF conditions for both Target and Masker models (Figures 7.2a, 7.2b, 7.3) suggests that the NR schemes, by suppressing background noise, increase the correlation between the neural activity and the speech features. This was expected as the cognitive load increases with the noise and disrupts both low- and high-level auditory processing [Mattys et al., 2012]. This increase in correlation for condition ON applies to both the Target and Masker model, which suggests that the neural tracking of the various speech features is more effective when the background noise is reduced. The enhancement of the masker speech may not be optimal for focusing on the target speaker, but it is important for the ability to switch targets in a conversation with multiple speakers.

For these results with the grouped conditions ON/OFF, we later in the analysis found that the correlation for OFF was largely influenced by condition 1 which had a lower performance than the other three conditions (Figure 7.4). The role of NR could therefore not be applied to both hearing aids as it only had an effect on HA 1.

Evaluating Conditions 1-4

Condition 1 (HA 1 NR OFF) consistently displayed lower correlation values compared to the other conditions across all models, for both the Target and Masker models (Figure 7.4). This suggests that the presence of background noise is particularly evident in condition 1 which impairs the subjects' ability to focus on and process the speech effectively. This finding implies that the absence of NR in HA 1 leads to reduced encoding of all speech features employed, compared to when it was switched on (condition 2). Interestingly, since condition 1 also yielded a significantly lower correlation compared to condition 3 (HA 2, NR OFF), it is suggested that hearing aid 2 better encodes the speech features and thus dampens background noise better in its OFF condition than hearing aid 1. Figure 7.5 demonstrates the improvement in mean correlation by removing condition 1 from the result.

There was no significant difference between NR ON and OFF for HA 2, which challenges the previous hypothesis that NR strengthens the encoding of the speech features in the EEG. This suggests that HA 2 has a robust baseline performance and does not benefit significantly from additional NR. This finding aligns with previous research suggesting that speech intelligibility is largely unaffected by background noise that does not overpower the speech signal's energy [Brungart, 2001]. About 50% of the subjects got higher correlation values when NR in hearing aid 2 was switched ON, see Figures 7.12, 7.13, 7.14 and 7.15. It is hard to asses whether this result is random or if it's related to the individual abilities of the subjects such as age, hearing loss, focus, etc.

When comparing the Masker models, the only conditions that were not significantly different were conditions 3 and 4, i.e., the settings for HA 2, which had higher correlations than both 1 and 2. This was not the case for the Target models, where no difference could be found between conditions 2, 3 and 4. These results suggest that hearing aid 2 is better at segregating the masker speech and could be beneficial for attention switching.

Comparison of Model Performance

The performance and range of mean correlation values of each TRF model were as expected, see all model's correlation values averaged across all conditions and models in Figure 7.1. The Envelope model was the best performing, indicating that the envelope is the best predictor for EEG. This aligns with previous research [Horton et al., 2014], which states that the neural response strongly encodes the envelope. Models based on word onset and phonetic features also showed significant representation in brain responses, with the PhoneticFeature model providing higher correlation values than the WordOnset model. The Surprisal model yielded the lowest correlations (mean of <1%).

An unexpected outcome was that adding more predictor variables to the envelope model did not improve the correlation. Previous research has stated that the predictive power of a TRF model increases when adding features like word onset and phonetic features to the model [Brodbeck et al., 2021]. One factor to consider is the sample size of the study; more robust results could be obtained by using a larger number of subjects. A plausible explanation for this lack of improvement could be the unique auditory processing characteristics of the HI subjects, who may process auditory input differently than those with normal hearing. This could then result in less benefit from additional predictors. Age also plays a role in auditory processing, with research suggesting that certain auditory processing peaks occur later in elderly populations.

Research has indicated that in situations of reduced speech intelligibility, the brain often relies more heavily on low-level speech features, such as the acoustic envelope, to comprehend speech [Mattys et al., 2012]. Due to the complex acoustic scene of the experiment and the strong correlation values obtained for the Envelope model, it seems plausible that this reliance on low-level features holds true for the population analyzed in this project.

Furthermore, it has previously been observed that HI individuals often exert greater cognitive effort during listening tasks compared to those with normal hearing [Reiss and Molis, 2021]. This heightened cognitive demand may lead to EEG signals that are more closely related to other cognitive processes, potentially overshadowing certain speech-related features analyzed in this project.

Neural Response of the Speech Features

As can be observed in the TRF plot in Figure 7.6, the chosen sensors (marked in blue in topographic maps) are relatively aligned both in timing and magnitude. Consistent for all TRF models is that the temporal dynamics of Target and Masker models are approximately the same, indicating similar processing timelines for both speech signals. Additionally, the neural responses to the Target model are consistently stronger compared to the Masker model. This further fortifies the notion that the subjects, despite their hearing impairments, can direct their attention to the target speaker in a cocktail party scenario.

The different predictors had unique impulse responses, indicating that they encode different parts of the auditory response and at different strengths. These highlight distinct aspects of neural encoding for different speech features. These differences can be detailed as follows:

Envelope. The acoustic envelope is the feature that has the highest TRF amplitude and thus shows the strongest encoded neural response across all TRF models. The Envelope TRF plot further displays a clear N1-P2-complex, see Figure 7.6 A, which is in line with previous research that shows that the acoustic envelope drives early auditory cortical responses [Oganian et al., 2023].

Word Surprisal. The Surprisal TRF (Figure 7.6 B) reveal two prominent positive peaks, likely related to the P100 and P300 peaks. Since the P300 peak is involved in attention and higher cognitive processes [Picton, 1992; Mueller et al., 2008], the presence of this peak suggest that the Surprisal model might in fact capture higher- level semantic processing. The N200 peak, which is related to attention and discrimination processes [Sur and Sinha, 2009], was also prominent, which indicates that the Surprisal model manages to capture some level of attention.

Word Onset. The pattern of the WordOnset TRF is very similar to that of the Surprisal but with a higher peak amplitude, which is expected as word surprisal is a scaled version of the word onsets. In Figure 7.6 C, there are a P100 and a P300 component, which indicates that the Word onsets trigger higher-level processes related to lexical integration. The presence of the N200 peak reinforces previous research that has stated that word onsets are correlated to this component, and that they are involved in attention and processing of phonologial information [Marslen-Wilson and Zwitserlood, 1989].

Phonetic Features. Phonetic features (D), being closely related to word onset times as it is also a fundamental element of speech, seem to be processed in early auditory regions but do not seem to require the broad cortical involvement seen with word onsets.

An interesting result is that for the PhoneticFeatures (see Figure 7.6 D), the amplitude of the first peak shows less variation between the masker and target models compared to the other models. This observation aligns with previous research, which indicates that early auditory processing, occurring before approximately 85 ms, is unable to distinguish between multiple speakers in individuals with normal hearing. [Alickovic et al., 2021]. The other two following peaks, beyond 85 ms, are more suppressed for the masker model. More analysis would be necessary to conclude that this also applies to hearing-impaired individuals, but it appears likely. Consequently, this finding supports the notion that the brain's ability to segregate and attend to a targeted speaker in a noisy environment is a higher-order process that occurs later in the auditory pathway.

8.2 Word Surprisal and Semantic Processing

Preprocessing and Modelling of Surprisal TRF

The mean correlation with EEG for the surprisal model did not improve despite using different variations of the surprisals and EEG parameters. This was unexpected, as the 0-30 Hz frequency band was anticipated to capture more surprisal encoding compared to the narrower 0-8 Hz range [Zion Golumbic et al., 2013]. However, these results suggest that the 0-8 Hz range is sufficient to capture the essential surprisal information in the data from the hearing aid users analyzed in this thesis.

Filtering the word surprisals to match the EEG signal also had no effect. Interestingly, filtering the signal introduced additional effects, altering its shape. The surprisals are represented as discrete instances between 0 and 1, though they likely represent a continuous variable. Therefore, it was expected that filtering or interpolation would improve the correlation. Alternatively, surprisals could be modeled as a step function, considering the end time of each word.

Word Surprisal as a Predictor Variable

One possible explanation for the low mean correlation for the Surprisal model (<1%) could be that word surprisal is a higher-level linguistic feature that requires more complex cognitive processing. The adverse hearing conditions of the experiment may have increased the cognitive demand on the hearing-impaired individuals to the extent that they did not comprehend enough of the target speech to encode word surprisal effectively. Additionally, HI individuals might rely more on lower-level acoustic features, such as the acoustic envelope, due to difficulties in processing higher-level linguistic information.

It was found that there was no significant difference between the masker models for word surprisal (Table 7.1 and Figure 7.11), unlike the other models where condition 1 had lower correlations for the masker models compared to the other conditions (Figure 7.4). This indicates that even in conditions 2, 3, and 4, there is minimal to no encoding of the masker stimuli. This result is expected, as word surprisal is a semantically processed feature and should not elicit a response if the subject is not attending to it.

The low correlation between surprisal and EEG, see Figure 7.1 and 7.13, might be attributed to the limitations of the linear TRF models. Exploring more advanced models, potentially incorporating nonlinearities or deep neural networks, could be a promising avenue for future research. A more complex model could offer valuable insights, pointing toward further investigation beyond the scope of this thesis.

Furthermore, analyzing the Surprisal model performance for the subjects individually (see Figure 7.13), it appeared more random compared to the other TRF models (see Figures 7.12, 7.14, and 7.15). This highlights that the level of word surprisal encoding captured by the Surprisal model is highly individual. Cognition and semantics are highly individual by nature with many factors shaping our thought patterns, such as the context of the situation, past experiences, culture, personality, and state of mind.

Another important factor for Surprisal models is how the predictor variable is created. There are different methods for calculating surprisal values, and some may yield higher correlation values than others. These calculations depend heavily on the LLM and how well it has been fine-tuned. This study was conducted using the Danish language, which is relatively small and less represented in technology. The GPT-2 model used [KennethTM, 2021] was adapted from the English medium version, which presents a challenge due to the limited Danish training data available. As a result, the performance of the Surprisal model might be constrained. In contrast, the same Surprisal model might be more successful if applied to English, which has access to larger and more diverse training datasets. Alternatively, using a language model trained on a more extensive Danish corpus could improve the accuracy and reliability of surprisal calculations for Danish speakers. Another option for creating predictor variables related to word surprisal would be to use the semantic embeddings from the LLM instead of calculating word probability.

Comparing Surprisal and WordOnset

Since word surprisals and word onsets occur the same time, there was concern that surprisal values may not carry meaningful information. Further analysis was conducted to determine whether word surprisals uniquely influenced the neural response or were a weaker version of word onsets.

TRF Estimation . The mean correlation and the neural response strength are consistently lower for the Surprisal model than the WordOnset model, see Figure 7.1 and Figure 7.6 (B and C). This can partially be explained by their model characteristics. During model training, the predictor variables were scaled relative to the EEG signal. Word onsets are binary events, occurring either 100% or 0% of the time. In contrast, surprisal values vary, reaching 100% only at their highest levels. Consequently, words with low surprisal elicit weaker responses, which reduces the overall amplitude of the Surprisal TRF.

Another important factor to consider is the time window selected in the boosting algorithm (basis). The time window for the Surprisal model ranges from -100 to 700 ms, whereas the window for the WordOnset model is from 100 to 400 ms latency. This difference was intentional to capture the semantic encoding in response to word surprisal values, which is known to occur later in auditory processing stages. However, this larger time window for the Surprisal model also means it is likely to capture more noise.

TRF Comparison. The pattern of both model's TRFs in Figure 7.6 (B and C) were similar and the later peak, P300, was of particular interest. The pronounced presence of the P300 peak highlights the model's potential to reflect the brain's response to the unexpectedness and contextual significance of words, thereby encoding complex semantic information beyond basic auditory features. Figure 7.8 illustrates that the Surprisal TRF exhibited a higher P300 peak compared to its P100 peak. In contrast, the WordOnset TRF shows the opposite pattern, with the P300 peak being lower than the P100 peak. The stronger P300 response in the Surprisal TRF indicates a deeper involvement in cognitive functions such as attention, memory, and semantic integration. This suggests that word surprisal, as a feature, encapsulates not only the timing of word onsets but also the brain's reaction to the contextual predictability of words, thereby providing a more comprehensive representation of speech processing.

Shuffled Surprisal. A shuffled surprisal model could be created and trained on several sets of shuffled surprisal values as training data to further investigate whether surprisal is encoded in the brain response and is significantly different from noise. This shuffled model would represent random noise and could be compared to the Surprisal model. This approach was planned for inclusion in this thesis but had to be omitted due to time constraints.

Significance of Condition for Target and Masker Models

The findings from the statistical analysis of the Surprisal models differ from those of the other models. Specifically, the ON or OFF condition did not significantly affect the masker Surprisal model; the correlation values between the two conditions showed no statistically significant difference (Table 10.6). This suggests that NR strategies do not influence how masker speech surprisal is encoded in the brain response of hearing-impaired subjects.

Similar to the other models, the target Surprisal model showed a significant difference between condition 1 and the other conditions, with no differences observed among conditions 2, 3, and 4. The conditions had no effect on the masker models, which could be due to little to no semantic encoding of the unattended speech. The difference in the TRF and topographic map between conditions 1 and 4 can be seen in Figure 7.11. The weaker and fragmented neural response in condition 1 suggests that the auditory processing system is struggling to synchronize and process the speech features effectively. It is evident that there is a significant difference in response between the two conditions, with condition 4 encoding information better than condition 1, which behaves similar to noise. This indicates that the Surprisal models for conditions 2, 3, and 4 model information that is distinct from random noise.

8.3 Subject-Specific Neural Responses

Consistency and Variability Across Subjects

Examining the individual correlation values for the target models (Figure 7.12, 7.13, 7.14, and 7.15) reveals that the results are relatively consistent across subjects, and that extreme outliers do not significantly influence the mean correlation values. The Envelope model demonstrated the highest and most consistent correlation values across all conditions, except for condition 1. In this condition, the correlations were mostly negative, a pattern observed in the other feature models as well. This further supports the notion that the Envelope is the highest correlated model and that condition 1 has the lowest correlation of the conditions.

It might seem reasonable to expect that subjects would perform consistently across all sessions; however, the individual nature of hearing loss means that preferences for different hearing aids and the use of NR are also highly individual. Mental fatigue is another important factor; subjects are likely to be more fatigued during session 4 than session 1. The adverse conditions in the experimental set up is designed to increase the cognitive load and thus the fatigue. Additionally, practice effects may play a role, with subjects potentially performing better in later sessions due to increased familiarity with the task. The significance of session order has not been explored further in the analysis but is an interesting subject for future work.

Although the different models are trained on the same EEG data, a strong correlation for a subject in one predictor variable and condition does not necessarily guarantee a strong correlation in other models. This variability might be due to the predictor variables being present at varying levels across different stages of speech processing [Mattys et al., 2012]. As the cognitive load increases, due to attentional resources and background noise, it is more difficult to distinguish word and phoneme onsets. The brain might reallocate its resources and rely more on contextual than acoustic information in these situations [Mattys et al., 2012]. The envelope will be encoded regardless of whether the subject can understand what is being said or not. [Tezcan et al., 2023].

Strong versus Weak Response Subjects

The comparison between high and low response subjects provides valuable insights into the variability of auditory processing within the dataset due to the three different HA conditions.

The initial negative peaks observed in Figure 7.16 A in the SR subjects SR2 and SR3 (around 70 ms latency) could be early N100 peaks, which typically appear around 100-160 ms. The prominent positive peak observed in the SR subjects is likely the P2 peak, usually expected around 200-300 ms. Together they form an N1-P2 complex. These strong positive responses are localized in the fronto-central region of the brain, indicating efficient auditory processing and strong neural encoding of the target speech features. The clear responses observed in the TRFs and topographic maps of the SR subjects align with the theories of neural tracking. The N100 and P200 peaks, generally associated with early sensory processing of auditory stimuli and attention discrimination, are prominently featured in these subjects. The prominence of these peaks suggests a more efficient initial auditory processing and better cognitive processing of the target stimuli compared to the WR subjects.

Within this comparison, a late N200 peak (typically expected around 200-300 ms) is only prominent in subject SR1 (at approximately 300 ms) among all subjects, both weak and strong. The N200 peak is often linked to attentional allocation and cognitive control. Therefore, the presence of this peak in SR1 indicates a better allocation of attention and cognitive resources towards processing the auditory stimuli. A similar pattern was observed for the WR subjects shown in Figure 7.16 B: two subjects, WR1 and WR2, exhibited an N100 peak, while the third subject, WR3, displayed a late N200 peak. The response amplitude was however lower than for the SR subjects.

The differences between the strong and WR subjects indicate significant variability in how effectively hearing-impaired individuals' brains can process and respond to auditory stimuli. The observation that only one subject exhibits a visible N200 peak, even among those selected based on their high correlation values, highlights the diversity among hearing-impaired subjects.

Another factor to consider is the diversity within the population. The subjects had varying ages and severity of hearing loss. For further research, it would be interesting to consider these factors as they affect the peak amplitude and latency [Mueller et al., 2008].

9

Conclusion

This thesis explored the neural encoding of speech features in hearing aid users, utilizing TRF and mTRF models. The primary objective was to investigate how the different speech features acoustic envelope, phonetic features, word onsets and word surprisals correlate with EEG signals during speech processing in a competing speech scenario. Additionally, the study aimed to assess the impact of hearing aid settings on these correlations.

The TRF and mTRF Models

Key findings were, consistent with previous research, that the Envelope models were the best predictors of EEG signals as they provided the highest mean correlation value. Following them in performance were the PhoneticFeatures models, succeeded by WordOnset and Surprisal models. The mTRF models, combining the envelope and the other features as predictor variables, did not improve the correlation. This suggests that, for individuals with hearing impairment, in high-cognitive demand scenarios, low-level acoustic features are likely the most critical for speech processing. Moreover, distinct neural encoding was observed for all speech features. Early processing stages correlated with the acoustic envelope and phonetic features, while later neural responses were influenced by word onset and word surprisal.

Word Surprisal

The Surprisal model yielded the lowest correlations, potentially due to the higher-level cognitive processing required for the encoding of word surprisal, which might be less accessible to individuals with hearing impairments under adverse listening scenarios. Some representation of word surprisal could however be found in the late neural responses, specifically the P100, N200, and P300 components. Differences between the Surprisal models and the WordOnset models suggest that the word surprisal could be used to predict some semantic encoding in the neural response.

Effect of Speaker (target vs masker)

Except for condition 1 (HA 1, NR OFF), all models and conditions showed significant differences for both Target and Masker models. This indicates that attended speaker could be distinguished from the EEG response. Interestingly, the Surprisal model was the only Masker model where condition 1 did not significantly differ from the other conditions. This suggests that word surprisal from the masker speaker is less represented in the neural response compared to other speech features.

Effect of Hearing Aid Conditions

Condition 1 (HA 1, NR OFF) consistently showed lower correlations across all models. There were no significant differences in correlation between the other three conditions. In other words, the NR had a reinforcing effect on the correlation for hearing aid 1 but not for hearing aid 2. The significant difference between condition 1 and condition 3 (HA 2, NR OFF) suggests that hearing aid 2 may offer more effective noise suppression even in its OFF state, indicating potential variations in device performance.

Need of Individualised Hearing Aids

The comparison of TRFs and topographic maps for SR and WR subjects underlines the importance of individual differences in neural processing of auditory stimuli among hearing-impaired individuals. Recognizing these differences can guide the development of more effective, personalized hearing aid technologies that cater to the specific neural processing characteristics of each user.

Implications for Future Research

The findings in this thesis suggest that incorporating more advanced models, e.g. non-linearity or deep neural networks, could improve the predictive power, especially in capturing the encoding of word surprisal. Additionally, using LLMs trained on larger and more diverse datasets could enhance the reliability of linguistic feature encoding in the neural responses when extracting word surprisal. This is particularly important when dealing with less represented languages like Danish.

Furthermore, this area of research could benefit from further investigation involving subjects using hearing aids, with larger sample sizes and more speech features such as cepstrum analysis and embeddings from the LLM. It would also be interesting to investigate individual differences between subjects such as age and severity of hearing loss.

Our results indicate the significance of speech features in comprehending the speech processing of individuals using hearing aids. This insight could pave the way for the future advancement of hearing aids, which are not only customized to individual needs but also adapted to their distinct neural responses.

Bibliography

Alickovic, E., T. Lunner, F. Gustafsson, and L. Ljung (2019). “A tutorial on auditory attention identification methods”. *frontiers in Neuroscience* **13**. DOI: `10.3389/fnins.2019.00153`.

Alickovic, E., T. Lunner, D. Wendt, L. Fiedler, R. Hietkamp, E. H. N. Ng, and C. Graversen (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**, p. 846.

Alickovic, E., C. F. Mendoza, A. Segar, M. Sandsten, and M. Skoglund (2023a). “DECODING AUDITORY ATTENTION FROM EEG DATA USING CEPSTRAL ANALYSIS”. In: *2023 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING WORKSHOPS, ICASSPW*. IEEE.

Alickovic, E., C. F. Mendoza, A. Segar, M. Sandsten, and M. A. Skoglund (2023b). “Decoding auditory attention from eeg data using cepstral analysis”. In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, pp. 1–5.

Alickovic, E., E. H. N. Ng, L. Fiedler, S. Santurette, H. Innes-Brown, and C. Graversen (2021). “Effects of hearing aid noise reduction on early and late cortical representations of competing talkers in noise”. *Frontiers in Neuroscience* **15**, p. 636060. DOI: `10.3389/fnins.2021.636060`.

Anderson, A. J., C. Davis, and E. C. Lalor (2023). “Context and attention shape electrophysiological correlates of speech-to-language transformation”. *bioRxiv (Cold Spring Harbor Laboratory)*. DOI: `10.1101/2023.09.24.559177`.

Biesmans, W., N. Das, T. Francart, and A. Bertrand (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**:5, pp. 402–412.

Bourisly, A. K. and A. Shuaib (2018). "Neurophysiological effects of aging: a p200 erp study". *Translational Neuroscience* **9**, pp. 61–66. DOI: 10.1515/tnsci-2018-0011.

Brodbeck, C., P. Das, J. P. Kulasingham, S. Bhattasali, P. Gaston, P. Resnik, and J. Z. Simon (2021). "Eelbrain: a python toolkit for time-continuous analysis with temporal response functions". *eLife* **12**. DOI: 10.1101/2021.08.01.454687.

Brodbeck, C., L. E. Hong, and J. Z. Simon (2018a). "Rapid transformation from auditory to linguistic representations of continuous speech". *Current Biology* **28**:24, pp. 3976–3983.

Brodbeck, C., L. E. Hong, and J. Z. Simon (2018b). "Rapid transformation from auditory to linguistic representations of continuous speech". *Current Biology* **28**, 3976–3983.e5. DOI: 10.1016/j.cub.2018.10.042.

Broderick, M. P., A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor (2018). "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech". *Current Biology* **28**:5, pp. 803–809.

Brungart, D. S. (2001). "Informational and energetic masking effects in the perception of two simultaneous talkers". *The Journal of the Acoustical Society of America* **109**, pp. 1101–1109. DOI: 10.1121/1.1345696.

Chalehchaleh, A., M. M. Winchester, and G. Di Liberto (2024). "Robust assessment of the cortical encoding of word-level expectations using the temporal response function". *bioRxiv*, pp. 2024–04.

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**, pp. 975–979. DOI: 10.1121/1.1907229.

Crosse, M. J., G. M. Di Liberto, A. Bednar, and E. C. Lalor (2016). "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli". *Frontiers in Human Neuroscience* **10**. DOI: 10.3389/fnhum.2016.00604.

Di Liberto, G. M., J. A. O'sullivan, and E. C. Lalor (2015). "Low-frequency cortical entrainment to speech reflects phoneme-level processing". *Current Biology* **25**:19, pp. 2457–2465.

Di Liberto, G. M., C. Pelofi, R. Bianco, P. Patel, A. D. Mehta, J. L. Herrero, A. De Cheveigné, S. Shamma, and N. Mesgarani (2020). "Cortical encoding of melodic expectations in human temporal cortex". *Elife* **9**, e51784.

Folstein, J. R. and C. Van Petten (2007). "Influence of cognitive control and mismatch on the n2 component of the erp: a review". *Psychophysiology* **0**, 070915195953001–??? DOI: 10.1111/j.1469-8986.2007.00602.x.

Friederici, A. D. (2011). "The brain basis of language processing: from structure to function". *Physiological Reviews* **91**, pp. 1357–1392. DOI: 10.1152/physrev.00006.2011.

Geirnaert, S., S. Vandecappelle, E. Alickovic, A. De Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand (2021). "Electroencephalography-based auditory attention decoding: toward neurosteered hearing devices". *IEEE Signal Processing Magazine* **38**:4, pp. 89–102.

Gillis, M., J. Van Canneyt, T. Francart, and J. Vanthornhout (2022). "Neural tracking as a diagnostic tool to assess the auditory pathway". *Hearing Research* **426**, p. 108607. DOI: 10.1016/j.heares.2022.108607.

Gillis, M., J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck (2021). "Neural markers of speech comprehension: measuring eeg tracking of linguistic speech representations, controlling the speech acoustics". *Journal of Neuroscience* **41**:50, pp. 10316–10329.

Hafter, E. R. (2010). "Is there a hearing aid for the thinking person?" *Journal of the American Academy of Audiology* **21**, pp. 594–600. DOI: 10.3766/jaaa.21.9.5.

Heilbron, M., K. Armeni, J.-M. Schoffelen, P. Hagoort, and F. P. de Lange (2022). "A hierarchy of linguistic predictions during natural language comprehension". *Proceedings of the National Academy of Sciences* **119**. DOI: 10.1073/pnas.2201968119.

Hickok, G. and D. Poeppel (2007). "The cortical organization of speech processing". *Nature Reviews Neuroscience* **8**, pp. 393–402. DOI: 10.1038/nrn2113.

Holmes, E., P. Folkeard, I. S. Johnsrude, and S. Scollie (2018). "Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment". *International Journal of Audiology* **57**, pp. 483–492. DOI: 10.1080/14992027.2018.1432901.

Horton, C., R. Srinivasan, and M. D'Zmura (2014). "Envelope responses in single-trial eeg indicate attended speaker in a 'cocktail party'". *Journal of Neural Engineering* **11**, p. 046015. DOI: 10.1088/1741-2560/11/4/046015.

Johnsrude, I. and J. M. Rodd (2015). "Factors that increase processing demands when listening to speech". *The Journal of the Acoustical Society of America* **137**, pp. 2211–2211. DOI: 10.1121/1.4920048.

Karunathilake, I. D., C. Brodbeck, S. Bhattasali, P. Resnik, and J. Z. Simon (2024). "Neural dynamics of the processing of speech features: evidence for a progression of features from acoustic to sentential processing". *bioRxiv*, pp. 2024–02.

Kegler, M., H. Weissbart, and T. Reichenbach (2022). "The neural response at the fundamental frequency of speech is modulated by word-level acoustic and linguistic information". *Frontiers in neuroscience* **16**, p. 915744.

KennethTM (2021). *Gpt-2 medium danish*. https://huggingface.co/KennethTM/gpt2-medium-danish. Accessed: 2024-05-24.

Kutas, M. and K. D. Federmeier (2011). “Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp)”. *Annual Review of Psychology* **62**, pp. 621–647. DOI: 10.1146/annurev.psych.093008.131123.

Lowder, M. W., W. Choi, F. Ferreira, and J. M. Henderson (2018). “Lexical predictability during natural reading: effects of surprisal and entropy reduction”. *Cognitive Science* **42**, pp. 1166–1183. DOI: 10.1111/cogs.12597.

Marslen-Wilson, W. and P. Zwitserlood (1989). “Accessing spoken words: the importance of word onsets.” *Journal of Experimental Psychology: Human Perception and Performance* **15**, pp. 576–585. DOI: 10.1037/0096-1523.15.3.576.

Mattys, S. L., M. H. Davis, A. R. Bradlow, and S. K. Scott (2012). “Speech recognition in adverse conditions: a review”. *Language and Cognitive Processes* **27**, pp. 953–978. DOI: 10.1080/01690965.2012.705006.

Mirkovic, B., M. G. Bleichner, M. De Vos, and S. Debener (2016). “Target speaker detection with concealed eeg around the ear”. *Frontiers in neuroscience* **10**, p. 206084.

Mueller, V., Y. Brehmer, T. von Oertzen, S.-C. Li, and U. Lindenberger (2008). “Electrophysiological correlates of selective attention: a lifespan comparison”. *BMC Neuroscience* **9**. DOI: 10.1186/1471-2202-9-18.

Näätänen, R. and I. Winkler (1999). “The concept of auditory stimulus representation in cognitive neuroscience”. *Psychological Bulletin* **125**, pp. 826–859. DOI: 10.1037/0033-2909.125.6.826.

Näätänen, R. and T. Picton (1987). “The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure”. *Psychophysiology* **24**, pp. 375–425. DOI: 10.1111/j.1469-8986.1987.tb00311.x.

O’sullivan, J. A., A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor (2015). “Attentional selection in a cocktail party environment can be decoded from single-trial eeg”. *Cerebral cortex* **25**:7, pp. 1697–1706.

Oganian, Y., K. Kojima, A. Breska, C. Cai, A. Findlay, E. F. Chang, and S. S. Nagarajan (2023). “Phase alignment of low-frequency neural activity to the amplitude envelope of speech reflects evoked responses to acoustic edges, not oscillatory entrainment”. *The Journal of Neuroscience* **43**, pp. 3909–3921. DOI: 10.1523/jneurosci.1663-22.2023.

OpenAI (2019). *Better language models and their implications*. Accessed: 2024-04-05. openai.com. URL: https://openai.com/research/better-language-models.

Paulraj, M. P., K. Subramaniam, S. B. Yaccob, A. H. B. Adom, and C. R. Hema (2015). “Auditory evoked potential response and hearing loss: a review”. *The Open Biomedical Engineering Journal* **9**, pp. 17–24. DOI: 10.2174/1874120701509010017.

Peelle, J. E., J. Gross, and M. H. Davis (2012). “Phase-locked responses to speech in human auditory cortex are enhanced during comprehension”. *Cerebral Cortex* **23**, pp. 1378–1387. DOI: 10.1093/cercor/bhs118.

Picton, T. W. (1992). “The p300 wave of the human event-related potential”. *Journal of Clinical Neurophysiology* **9**, p. 456. URL: https://journals.lww.com/clinicalneurophys/abstract/1992/10000/the_p300_wave_of_the_human_event_related_potential.2.aspx.

Podury, A., N. T. Jiam, M. Kim, J. I. Donnenfield, and A. Dhand (2023). “Hearing and sociality: the implications of hearing loss on social life”. *Frontiers in Neuroscience* **17**. DOI: 10.3389/fnins.2023.1245434.

Polich, J. (2007). “Updating p300: an integrative theory of p3a and p3b”. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, pp. 2128–48. DOI: 10.1016/j.clinph.2007.04.019.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever (2023). “Robust speech recognition via large-scale weak supervision”. In: *International Conference on Machine Learning*. PMLR, pp. 28492–28518.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). *Language models are unsupervised multitask learners*. Accessed: 2024-04-14. Semantic Scholar. URL: https://api.semanticscholar.org/CorpusID:160025533.

Reiss, L. A. and M. R. Molis (2021). “An alternative explanation for difficulties with speech in background talkers: abnormal fusion of vowels across fundamental frequency and ears”. *Journal of the Association for Research in Otolaryngology* **22**, pp. 443–461. DOI: 10.1007/s10162-021-00790-7.

Schneider, B. A., C. Rabaglia, M. Avivi-Reich, D. Krieger, S. R. Arnott, and C. Alain (2022). “Age-related differences in early cortical representations of target speech masked by either steady-state noise or competing speech”. *Frontiers in Psychology* **13**. DOI: 10.3389/fpsyg.2022.935475.

Scott, S. K., S. Rosen, C. P. Beaman, J. P. Davis, and (2009). “The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes”. *National Library of Medicine* **125**, pp. 1737–1743. DOI: 10.1121/1.3050255.

Seabold, S. and J. Perktold (2010). *Statsmodels: econometric and statistical modeling with python*. Accessed: 2024-06-05. URL: https://www.statsmodels.org/.

Siuly, S., Y. Li, and Y. Zhang (2016). *EEG Signal Analysis and Classification*. Springer International Publishing. DOI: 10.1007/978-3-319-47653-7.

Sur, S. and V. Sinha (2009). “Event-related potential: an overview”. *Industrial Psychiatry Journal* **18**, p. 70. DOI: 10.4103/0972-6748.57865.

Team, T. (2024). *Textgrid.* https://www.textgrid.org/. Accessed: 2024-06-05.

Tezcan, F., H. Weissbart, and A. E. Martin (2023). “A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension”. *eLife* **12**. DOI: 10.7554/elife.82386.

The MathWorks Inc (n.d.). *Envelope extraction - matlab simulink - mathworks nordic*. Accessed: 2024-05-06. se.mathworks.com. URL: https://se.mathworks.com/help/signal/ug/envelope-extraction-using-the-analytic-signal.html.

Tomé, D., F. Barbosa, K. Nowak, and J. Marques-Teixeira (2014). “The development of the n1 and n2 components in auditory oddball paradigms: a systematic review with narrative analysis and suggested normative values”. *Journal of Neural Transmission* **122**, pp. 375–391. DOI: 10.1007/s00702-014-1258-3.

Van Canneyt, J., J. Wouters, and T. Francart (2021). “Neural tracking of the fundamental frequency of the voice: the effect of voice characteristics”. *European Journal of Neuroscience* **53**:11, pp. 3640–3653.

Vander Werff, K. R., C. E. Niemczak, and K. Morse (2021). “Informational masking effects of speech versus nonspeech noise on cortical auditory evoked potentials”. *Journal of Speech, Language, and Hearing Research* **64**, pp. 4014–4029. DOI: 10.1044/2021_jslhr-21-00048.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush (2020). “Transformers: state-of-the-art natural language processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

World Health Oragnization (2024). *Deafness and hearing loss*. Accessed: 2024-03-20. WHO. URL: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

Zion Golumbic, E. M., N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder (2013). “Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party””. *Neuron* **77**, pp. 980–991. DOI: 10.1016/j.neuron.2012.12.037.

10

Appendix

10.1 All TRF and mTRF model performances

Model effect

Table 10.1 One-Way ANOVA - Fixed effects: **Condition (ON)**, Speaker (T); Independent factor: Model (TRFs); Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for that particular model combination.

Model	Phonetic Features	Surprisals	Word Onsets
Envelope	2e-08	2e-23	8e-20
Phonetic Features		7e-27	2e-28
Surprisals			3e-5

Table 10.2 One-Way ANOVA - Fixed effects: **Condition (4)**, Speaker (T); Independent factor: Model (TRFs); Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for that particular model combination.

Model	Phonetic Features	Surprisals	Word Onsets
Envelope	3e-5	3e-13	5e-11
Phonetic Features		7e-15	7e-15
Surprisals			6e-10

Table 10.3 One-Way ANOVA - Fixed effects: **Condition (2+3+4)**, Speaker (T); Independent factor: Model (TRFs); Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for that particular model combination.

Model	Phonetic Features	Surprisals	Word Onsets
Envelope	7e-13	2e-35	7e-30
Phonetic Features		4e-15	2e-23
Surprisals			9e-07

Table 10.4 One-Way ANOVA - Fixed effects: Condition (2+3+4), Speaker (T); Independent factor: Model (Envelope versus mTRFs); Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for that particular model combination.

mTRF models	Envelope
Envelope+PhoneticFeatures	0.75
Envelope+Surprisal	0.90
Envelope+WordOnset	0.92
Envelope+PhoneticFeatures+WordOnset	0.84
Envelope+Surprisal+PhoneticFeatures	0.74
Envelope+Surprisal+WordOnset	0.11
AllPredictors	0.79

Speaker Effect

Table 10.5 One-Way ANOVA - Fixed effects: Model, Condition; Independent factor: Speaker; Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for that particular model and condition combination. $\alpha < 0.05$

Model Condition	ON	OFF	C1	C2	C3	C4	All
Envelope	5e-21	0.002	0.010	7e-11	7e-10	1e-11	4e-16
PhoneticFeatures	3e-12	9e-5	0.497	4e-07	2e-07	3e-6	9e-13
Surprisal	4e-5	0.101	0.751	0.020	0.048	4e-4	5e-5
WordOnset	3e-09	0.002	0.933	4e-5	1e5	2e-5	7e-10
Envelope+PhoneticFeatures	8e-18	3e-5	0.824	3e-09	1e-09	7e-10	3e-17
Envelope+Surprisal	2e-21	4e-4	0.227	1e-10	1e-09	3e-12	3e-18
Envelope+WordOnset	8.e-22	8e-4	0.053	3e-11	3e-10	6e-12	2e-17
Envelope+PhoneticFeatures+WordOnset	3e-18	3e-5	0.816	2e-09	1e-09	34e-10	2e-17
Envelope+Surprisal+PhoneticFeatures	6e-19	9e-5	0.560	4e-09	3e-09	3e-11	4e-17
Envelope+Surprisal+WordOnset	1e-23	6e-4	0.177	4e-12	8e-11	5e-13	2e-18
AllPredictors	6e-19	1e-4	0.498	3e-09	2e-09	3e-11	5e-17

Condition Effect

Table 10.6 One-Way ANOVA - Fixed effects: Speaker; Independent factor: Condition (ON/OFF); Dependent factor: Correlation. P-values are presented in the table, indicating the statistical significance of the correlation for Target and Masker.

Condition:	**ON/OFF**
Target	6e-42
Masker	5e-10

Table 10.7 **Target** models: Post-hoc Tukey's results for the effect of Condition between Conditions 1, 2, 3, and 4. TRUE (green) indicates a significant difference between the combination of conditions, FALSE (red) indicates no significant difference. $\alpha < 0.05$

Condition	2	3	4
1	TRUE	TRUE	TRUE
2		FALSE	FALSE
3			FALSE

Table 10.8 **Masker** models: Post-hoc Tukey's results for the effect of Condition between Conditions 1, 2, 3, and 4. TRUE (green) indicates a significant difference between the combination of conditions, FALSE (red) indicates no significant difference. $\alpha < 0.05$

Condition	2	3	4
1	TRUE	TRUE	TRUE
2		TRUE	TRUE
3			FALSE

Mean Correlation Values for Target and Masker Models (All Conditions)

Table 10.9 **Target** models: Mean correlation values for all models (Pearson correlation)

Model	**Correlation**
Envelope	0.049
PhoneticFeatures	0.025
WordOnset	0.010
Envelope+WordOnset	0.049
Surprisal	0.005
Envelope+Surprisal+PhoneticFeatures+WordOnset	0.047
Envelope+PhoneticFeatures	0.047
Envelope+Surprisal	0.049
Envelope+PhoneticFeatures+WordOnset	0.048
Envelope+Surprisal+WordOnset	0.042
Envelope+Surprisal+PhoneticFeatures	0.047

Table 10.10 **Masker** models: Mean correlation values for all models (Pearson correlation)

Model	**Correlation**
Envelope	0.011
PhoneticFeatures	0.003
WordOnset	-0.010
Envelope+WordOnset	-0.010
Surprisal	-0.002
Envelope+Surprisal+PhoneticFeatures+WordOnset	0.007
Envelope+PhoneticFeatures	0.008
Envelope+Surprisal	0.009
Envelope+PhoneticFeatures+WordOnset	0.008
Envelope+Surprisal+WordOnset	0.008
Envelope+Surprisal+PhoneticFeatures	0.007

10.2 Surprisal models

Condition effect

Table 10.11 **Target** models: Post-hoc Tukey's results for the effect of Condition between Conditions 1, 2, 3, and 4. TRUE (green) indicates a significant difference between the combination of conditions, FALSE (red) indicates no significant difference. $\alpha < 0.05$

Condition	2	3	4
1	TRUE	TRUE	TRUE
2		FALSE	FALSE
3			FALSE

Lund University
Department of Automatic Control
Box 118
SE-221 00 Lund Sweden

Document name
MASTER'S THESIS

Date of issue
June 2024

Document Number
TFRT-6243

Author(s)
Klara Almgren
Annie Mentzer

Supervisor
Emina Alickovic, Eriksholm Research Centre
Martin Skoglund, Eriksholm Research Centre
Bo Bernhadsson, Dept. of Automatic Control, Lund University
Pontus Giselsson, Dept. of Automatic Control, Lund University, Sweden (examiner)

Title and subtitle
Neural Speech Tracking in EEG: Integrating Acoustics and Linguistics for Hearing Aid Users

Abstract

This master thesis explores the neural encoding of speech features for hearing aid users. The study utilizes electroencephalography (EEG) and audio data from an experiment that stimulates a Cocktail Party scenario. This is a complex auditory scene, especially difficult for individuals with hearing impairments. The primary objective of the study is to investigate how different acoustic and linguistic speech features are represented in the brain response and how these representations are influenced by hearing aid settings. The speech features analyzed are the acoustic envelope, phonetic features, word onset, and word surprisal. The word surprisal values were derived from GPT-2. Temporal Response Functions (TRFs) and multivariate TRFs (mTRFs) were employed to examine the correlation between these features and EEG signals during speech processing in both attended (target) and unattended (masker) speech scenarios.

The TRFs were estimated by training a forward model using a boosting algorithm. In this process, the speech features serve as the predictor variables (X-data), and the EEG signals serve as the response variables (Y-data). The boosting algorithm iteratively improves the model by combining multiple weak learners to better predict the EEG responses based on the given speech features.

The study found that target and masker speech are significantly distinguishable using TRF models trained on these features. It also revealed that hearing aid conditions impact their encoding. Among the features analyzed, the acoustic envelope had the highest correlation with neural responses. Adding other predictor variables to the Envelope model did not improve the correlation. Further, all speech features were found to have unique neural encoding. The acoustic envelope and phonetic features could be correlated to early processing, while word onset and word surprisal are reflected in later neural responses.

Our findings suggest that speech features are important in understanding how hearing aid users process speech, which could lead to future development of hearing aids that are not only fitted to the user's needs but also tailored to their unique neural responses.

Keywords

Classification system and/or index terms (if any)

Supplementary bibliographical information

ISSN and key title
0280-5316

ISBN

Language English	*Number of pages* 1-84	*Recipient’s notes*
Security classification		

http://www.control.lth.se/publications/