Title	音声知覚における脳内の音声エンコーディング・デコーディングのプロセスの研究
Author(s)	周, 迪
Citation	
Issue Date	2023-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/18781
Rights	
Description	Supervisor: 鵜木祐史, 先端科学技術研究科, 博士

Doctoral Dissertation

Speech encoding and decoding processes in the brain during speech perception

Di Zhou

Supervisor　Unoki Masashi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2023

Abstract

In recent years, there has been a growing trend towards using naturalistic setups in speech research. These setups involve presenting participants with continuous speech, such as stories or conversations, instead of isolated words or sentences. This provides a more realistic understanding of speech processing and comprehension in real-world conditions.

However, fMRI studies have found that naturalistic speech processing elicits more widespread brain activation compared to traditional setups. This highlights a challenge for traditional neurocognitive models of language, which are based on isolated words or simple sentences and have a strong left hemisphere bias. These models fail to explain how naturalistic speech is processed in the brain.

To address this limitation, our study aims to describe the encoding of speech in the brain using a linear time-invariant (LTI) model. We focus on the process of semantic processing during speech perception and investigate the extraction of the temporal amplitude envelope (TAE) from the speech signal. This TAE carries crucial semantic information and is encoded by specific brain regions involved in semantic processing.

We used EEG signals to measure brain activity and traced the origins of activity with our proposed hyper-alignment method under natural language paradigms. Our goal was to identify the brain regions responsible for semantic processing, determine the semantic representations in the brain's output, and explore the possibility of recovering the original semantic information from this representation. In an experiment, participants were exposed to normal speech and time-reversed speech. Then, we used the hyper-alignment method to map EEG signals from the scalp to the cortex level to overcome the spatial limitation of EEG and obtain precise information about the semantic processing in the brain.

Our findings reveal that semantic processing during naturalistic paradigms involves a widespread distribution of brain regions beyond the traditional

temporal and frontal areas. We observed the involvement of the cingulate area and a significant role played by the right hemisphere in semantic processing, challenging the conventional left hemisphere bias. Using multivariate autoregressive modeling, we captured the dynamic characteristics of brain activity and found a top-down predictive mechanism where higher-level semantic processing areas assist in capturing upcoming acoustic features.

Through the reverse decoding process, we successfully reconstructed the TAEs of speech from brain activity and recovered semantic information using noise-vocoded speech (NVS). While speech intelligibility restoration on unknown data remains a challenge due to noise and inter-subject variability, we achieved perfect fitting of TAEs and restored semantic information to a certain level of speech intelligibility in the known training dataset.

In summary, our study provides a unique perspective on the brain's natural language processing by combining both temporal and spatial dimensions. We have developed an innovative methodology to estimate encoding functions across various brain regions, which was previously difficult to achieve with fMRI and EEG research methods. By challenging traditional models, we emphasize the extensive involvement of multiple brain regions and the dynamic nature of their encoding capabilities. Our findings suggest that desynchronization between different subnetworks, especially within the frontal and temporal areas, plays a crucial role in the brain's semantic information processing mechanism. Our research also involves reconstructing speech TAEs and recovering semantic content, which deepens the understanding of language processing. Furthermore, our findings may potentially lead to advancements in speech-brain interface technologies in the future.
Keywords: electroencephalography, speech encoding/decoding, temporal response function, source localization, neural entrainment.

Acknowledgment

Firstly, I would like to express my heartfelt gratitude to my research supervisors, Prof. Jianwu Dang and Prof. Masashi Unoki, for their invaluable assistance and guidance throughout my research journey. Their expertise and support have been instrumental in helping me explore the fascinating field of cognitive science and successfully complete this paper.

I am also deeply grateful to my minor research supervisor, Prof. Gaoyan Zhang, for their insightful feedback and contributions to my research. Their perspectives and expertise have added significant value to my work.

I extend my sincere appreciation to my research teams at His-lab in JAIST and the teams at the Tianjin Key Laboratory of Cognitive Computing and Applications at Tianjin University. The stimulating discussions during our team meetings have played a crucial role in enhancing my research progress. I have cherished many wonderful memories with my teammates, and their camaraderie has been invaluable.

Furthermore, I would like to express my heartfelt thanks to my girlfriend. Her unwavering support during the challenging times of research has been a constant source of strength. Together, we have faced difficulties head-on, providing each other with encouragement and collaborating to overcome obstacles.

Lastly, none of this would have been possible without the unwavering support of my family. They have been a constant source of inspiration and motivation, uplifting me in times of doubt. Their encouragement and guidance have propelled me forward, and I am forever grateful for their unwavering belief in me.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Background

Speech perception is the process of organizing, identifying, and interpreting communicative information conveyed by speech sounds, which involves the linkage of auditory signals and the nervous system [1]. During speech perception, human auditory system first transforms sound waves into electrical signals, and the brain then decodes these signals to perceive the sound. Speech processing is considered one of the most complex and abstract systems in human cognitive systems, and its specific mechanisms and foundations are not yet fully understood. To better understand the neural mechanisms underlying speech processing in the brain, researchers use theories and methods from psychology, cognitive neuroscience, and linguistics. Among these methods, neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), have become important tools in recent decades. They have been widely used to investigate brain regions and brain responses involved in language processing and have advanced our understanding of these mechanisms [2–4].

In past decades, researchers tried to use isolated words or simple sentences to investigate speech comprehension in the human brain. In those studies, subjects were asked to participate in specific tasks such as identifying whether the perceived word is a real word or a pseudoword [1,5], or assessing whether a word in a sentence is congruent or incongruent with the rest of the sentence [6]. With this kind of well-designed paradigm, researchers can use a statistical analysis method (such as t-tests, analysis-of-variance) to estimate

the mechanism of speech processing by comparing neural behaviors between different conditions. For example, in an experiment designed to study the N400 amplitude and word expectancy, it was found that N400 responses to sentences that were inconsistent with word expectancy and common sense (e.g., The bill was due at the end of the hour) were significantly larger compared to those elicited by sentences that were consistent with expectations and common sense (e.g., The bill was due at the end of the month) [7]. However, such a task is far away from human speech comprehension in daily life.

In recent years, there has been a notable shift in research towards expanding the controlled experimental paradigm beyond traditional isolated words or simple sentences. Researchers have begun to embrace more naturalistic experimental settings that involve scenarios where participants engage in listening tasks with continuous speech and complete storylines [8]. Through these naturalistic experimental settings, researchers have discovered that natural language processing engages a broader range of brain regions compared to traditional isolated words or simple sentences [9]. Moreover, naturalistic language processing involves widespread activation across the entire brain, including the right hemisphere. These findings suggest that traditional results based on isolated words or simple sentences may not fully apply to naturalistic experimental settings [8, 10]. Therefore, it is necessary to further investigate the process of language processing within naturalistic paradigms.

1.2 Research significance

The research significance of the study lies in its exploration of the connection between language, cognition, and brain activity during the process of speech perception. By understanding how the brain converts speech into meaningful representations during encoding and then retrieves these representations during decoding, valuable insights can be gained into speech comprehension.

This study has significant implications for the fields of neuroscience, psychology, and linguistics. It provides a deeper understanding of the

intricate mechanisms that are involved in speech processing and lays a foundation for further advancements in these disciplines. By exploring the neural processes that underlie speech comprehension, researchers can uncover fundamental principles of language processing and cognition.

Furthermore, this research has practical applications and potential benefits for individuals with hearing loss. The insights gained from the study can be used to develop advanced technologies that enhance speech perception for people with hearing difficulties. This has the potential to improve communication and overall quality of life for people with hearing impairments, underscoring the significance of the research in the development of future assistive technologies.

Overall, the study's significance lies in its contribution to our understanding of speech perception, its interdisciplinary nature encompassing neuroscience, psychology, and linguistics, and its potential to impact the development of technologies for individuals with hearing loss.

1.3 Research motivation

My interest in investigating the relationship between the auditory perception system and the brain system during speech perception and comprehension has motivated my research. Although it is difficult to directly observe the brain processes during speech perception, understanding these mechanisms is crucial, and the study aims to achieve this.

From an engineering perspective, informed decisions or interventions are challenging to make without observation. The theory of systems provides a framework to evaluate the brain system's functioning by estimating its encoding function, which involves the conversion of speech signals into neural representations.

The study's motivation lies in the possibility of estimating the encoding function by analyzing observed speech signals and brain activity. By understanding the relationship between these signals and neural responses, researchers can gain insights into how the brain processes speech information. This understanding can inform the development of technologies and

interventions aimed at enhancing speech perception and comprehension.

In summary, my research is motivated by the desire to uncover the mechanisms underlying speech perception and comprehension by estimating the encoding/decoding function through the analysis of observed speech signals and brain activity.

1.4 Research challenges

The research challenges can be summarized as follows:

Capturing dynamic information: The comprehension of natural language triggers a series of interconnected processes that unfold concurrently and overlap in time. While fMRI provides excellent spatial resolution, its temporal resolution is insufficient for capturing the rapid temporal dynamics of speech processing. This poses a challenge in accurately assessing the encoding function during natural language processing stages.

Spatial resolution limitations of EEG: EEG, with its high temporal resolution, is often used to estimate the encoding function in natural language processing. However, its lower spatial resolution means that the signals recorded from EEG electrodes represent a mixture of many source components, making it challenging to precisely identify the cortical origins of the underlying processes involved in speech comprehension.

Integration of spatial and temporal information: To gain a comprehensive understanding of speech comprehension, it is crucial to capture both spatial and temporal information simultaneously. Combining the strengths of fMRI and EEG could provide a more comprehensive view of the encoding function across different brain regions, but it presents a challenge due to the disparate spatial and temporal resolutions of the two techniques.

Due to the current limited availability of non-invasive neuroimaging techniques that can simultaneously provide both temporal and spatial information, estimating the encoding function at the brain cortex level remains challenging. Addressing these research challenges requires innovative approaches and methodologies that can effectively capture and integrate spatial and temporal information, ultimately enabling a more accurate estimation of

the encoding function in different brain regions during the process of speech comprehension.

1.5 Research goal

This study aims to investigate the encoding and decoding processes of natural language in the brain from both spatial and temporal aspects. Specifically, we aim to address the following questions: (1) Which brain regions play a crucial role in the comprehension of natural language? (2) Whether it can find the speech semantic representation in these brain regions? (3) If we can find the speech semantic representation in the brain, can we decode the original speech semantic information from these brain's activity? To answer these questions, several steps need to be taken, as outlined below:

1.5.1 Crucial brain regions for speech comprehension

To identify the brain regions involved in natural language comprehension, our study involved mapping and studying the neural activity within the brain during speech processing tasks. We conducted an EEG experiment that compared natural language stimuli with time-reversed speech. During the experiment, participants were exposed to both natural and time-reversed speech stimuli, while their brain activity was measured using EEG. Time-reversed speech is difficult to comprehend, and its processing can be assumed to be unrelated to language comprehension. By contrasting the neural activity associated with natural language and time-reversed speech, we aimed to determine which brain regions contribute more significantly to the language comprehension process. We specifically looked for brain regions that exhibited differential activity between the two conditions. Our goal was to identify brain regions that showed greater activation or connectivity patterns specific to natural language comprehension. By comparing these differential neural responses, we aimed to pinpoint the brain regions that are more strongly associated with language comprehension. These regions are likely to play a critical role in processing the semantic aspects of natural

language.

This approach allowed us to distinguish brain activity related to language comprehension from general auditory processing or acoustic features that are present in both natural language and time-reversed speech. By doing so, we gained valuable insights into the specific brain regions and networks involved in the comprehension of natural language. This, in turn, contributes to a better understanding of the neural mechanisms underlying language processing.

1.5.2 Semantic representation in brain cortex

To understand how semantic information is represented and processed in the identified brain regions, we will employ advanced analytical methods, such as common spatial pattern (CSP) [11] and multivariate autoregressive (MVAR) [12]. These techniques can provide valuable insights into the neural coding of semantic information, as well as describe the dynamic changes in brain networks during the processing of semantic information.

By examining the similarity or dissimilarity of neural representations across different stimuli or conditions, we can gain insights into how the brain encodes and discriminates semantic information. Furthermore, the processing of semantic information in the brain involves the dynamic interplay of multiple brain regions and networks. Functional connectivity analysis can be employed to examine the temporal correlations and interactions between different brain regions during the processing of semantic information. This analysis can reveal the dynamic changes in functional connectivity patterns and identify the key network nodes involved in semantic processing.

1.5.3 Decode speech from brain activities

Read out speech directly from measured neural activity could enable natural conversations and improve quality of life, particularly for the individuals who suffer from neurological diseases. Using the collected brain activity data, we will explore the possibility of decoding the original speech information.

This can involve reconstructing speech sounds or identifying specific linguistic features from the neural signals using machine learning approaches.

1.5.4 Interpretability of our results

Our findings from the above analyses will be interpreted and discussed in the context of existing literature on language processing. By following these steps, we aim to shed light on the neural mechanisms of natural language processing, including the brain regions involved, the representation of semantic information, and the potential for decoding speech information from brain activity.

1.6 Research novelty

The research novelty lies in its integration of spatial and temporal perspectives to investigate natural language processing comprehensively. The study aims to simultaneously consider both spatial and temporal aspects, providing a deeper understanding of the neural mechanisms involved in natural language comprehension.

One crucial objective is to identify specific brain regions that play a critical role in understanding natural language. By comparing neural responses to natural speech and time-reversed speech, the research aims to identify brain regions that are strongly associated with speech comprehension. This exploration offers valuable insights into the neural basis of language processing.

Another focus is on investigating how semantic information is represented and processed within the identified brain regions. By employing advanced analytical methods, the study aims to uncover the neural encoding of semantic information and explore the potential for directly decoding speech information from measured neural activity. These findings could have significant implications for natural conversations and improving the quality of life for individuals with neurological conditions.

1.7 Research originality

The current study is unique in multiple aspects. Firstly, it challenges the traditional belief that language processing occurs only in specific brain areas. Instead, the study explores the natural language processing process from a new perspective of brain networks. This theoretical contribution provides new insights into the potential mechanisms underlying language processing.

Secondly, the study presents methodological innovations by integrating spatial and temporal perspectives, which offer a clearer understanding of how natural language is processed in the brain. This integration allows for a more comprehensive analysis of language processing.

Moreover, the study addresses the limitations associated with applying source localization techniques to natural language and proposes advancements in this area. It overcomes the challenges that arise when studying natural language comprehension.

Lastly, the study explores the reconstruction of original speech information from brain activity using noise-vocoded speech (NVS) techniques [13]. This technological innovation holds potential applications in speech prosthetics and communication assistance.

Overall, the study's originality lies in its challenge to traditional theories by adopting a brain network perspective. It also showcases methodological advancements in integrating spatial and temporal perspectives, improving source localization techniques for natural language, and exploring the reconstruction of speech information from brain activity. Our study enhances the understanding of natural language processing and holds great promise for practical applications in the field of speech assistance.

1.8 Dissertation outline

The organization of this dissertation is illustrated in Fig. 1.1. Apart from this introductory chapter, the remainder of the dissertation comprises five chapters. The literature review will provide an overview of the current advancements in natural language paradigm research, followed by methodology,

and a comprehensive presentation of our data collection. Subsequently, we will analyze our results from both the encoding and decoding processes of speech. Finally, we will interpret and discuss our findings in light of existing research, while also highlighting future directions.

Chapter 2 will primarily concentrate on the development, advantages, and distinctions in analysis methods between natural language paradigms and traditional well-designed experimental paradigms. Additionally, we will explain why research on natural language paradigms faces challenges when extending to source space and the reasons behind the limitations of traditional source localization procedures in accurately localizing cortex level activity in natural language paradigms.

Chapter 3 will introduce our research philosophy and the methodology in our study.

Chapter 4 is the design and procedure of our experimental.

Chapter 5 will employ the aforementioned data analysis methods to investigate the representation of speech semantic information in the brain and the associated brain regions, focusing on the encoding perspective. Utilizing the proposed source localization procedures, we will present the dynamic process of language processing in the brain from both temporal and spatial dimensions.

Chapter 6 will delve into the decoding aspect by examining the feasibility of reconstructing speech temporal amplitude envelopes (TAEs) from recorded neural signals. Additionally, leveraging knowledge from noise-vocoded speech (NVS), we will attempt to non-invasively reconstruct the semantic information of the original speech using scalp EEG data.

Chapter 7 is a general discussion. We discussed the benefits of transitioning from the scalp level to the cortex level, which allows for more precise encoding and decoding functions. By incorporating spatial and temporal information, we can construct dynamic brain networks during semantic processing. Additionally, we compared the decoding differences between natural speech and time-reversed speech. Due to the lack of semantic information in time-reversed speech, its decoding accuracy is significantly lower than that of natural speech.

Chapter 8 highlights the research contributions, limitations, and future directions of the study.

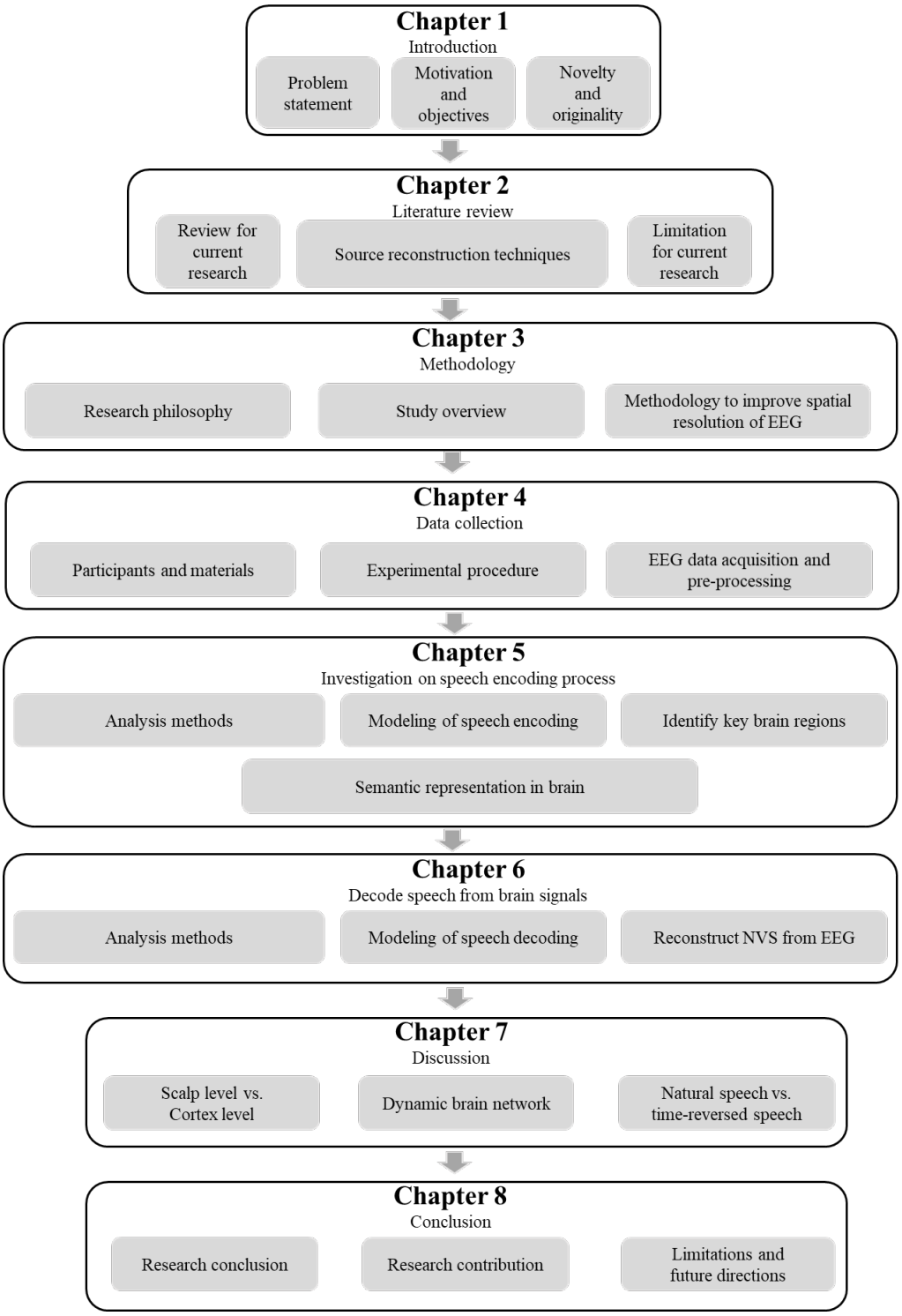

Figure 1.1: Thesis organization.

Chapter 2

Literature review

This chapter begins by introducing EEG and fMRI studies that explore language comprehension using traditional isolated words or simple sentences paradigms. It then discusses the significance of shifting from traditional paradigms to naturalistic experimentation and the differences in analysis approaches compared to isolated words or simple sentences. The focus then shifts to recent advancements in high-temporal-resolution EEG techniques within the context of naturalistic experimentation and their limitations. Additionally, the chapter discusses the challenges of applying EEG source localization techniques to naturalistic experimentation paradigms.

2.1 Investigate speech processing based on traditional well-designed paradigm

2.1.1 Insights from high spatial resolution

When utilizing neuroimaging methods to investigate speech comprehension, researchers must decide whether to prioritize temporal or spatial resolution for language-related tasks. In simple terms, functional magnetic resonance imaging (fMRI) provides sufficient spatial resolution to investigate the location of semantic processing in the brain. Figure 2.1 shows these brain areas for semantic processing base on fMRI.

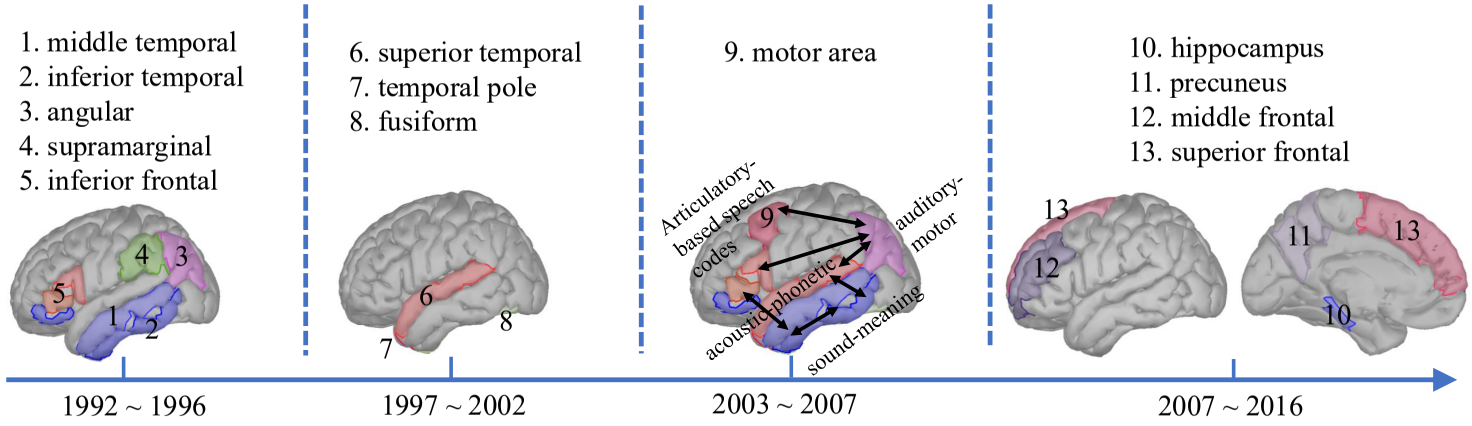

Figure 2.1: Brain areas for semantic processing base on traditional well-designed paradigm.

Around the period of 1996, through the comparison of semantic processing on real words and pseudowords, research revealed widely distributed activation in several brain areas. Specifically, the left middle and inferior temporal gyri, angular gyri, superior temporal gyri, supramarginal gyri, and inferior frontal areas were found to be activated during semantic processing [14–16].

Around 2002, an increasing number of studies reaffirmed the key role of the previously identified brain areas, including the temporal regions and Broca's area of inferior frontal areas (BA44, BA45 and BA47), in semantic processing [17–19]. Notably, when participants processed complex and semantically incongruent phrases or sentences, the activation in these regions became particularly evident [20]. Furthermore, other studies also highlighted the importance of the superior temporal gyrus, temporal pole, and fusiform gyrus in language comprehension [21, 22].

Around 2007, Hickok integrated the brain regions associated with speech semantics and comprehension and proposed the influential dual-stream model [23, 24]. According to this model, there are two main processing pathways involved in speech comprehension. The ventral pathway is responsible for mapping the acoustic information of speech to its meaning. It begins in the superior temporal gyrus and progresses ventrally into the middle temporal gyrus and inferior temporal gyrus. This pathway is primarily involved in sound-to-meaning mapping and is crucial for semantic processing in speech comprehension. In contrast, the dorsal pathway is involved in the processing of speech sounds for articulation and phonological information. It starts in the superior temporal gyrus and extends dorsally into the superior parietal lobule and posterior frontal areas, including Broca's area. This pathway is responsible for mapping speech sounds to articulatory representations and plays a role in phonological processing during speech comprehension. Overall, the dual-stream model provides a framework for understanding the functional organization of the brain regions involved in speech semantics and comprehension, highlighting the distinct processing streams involved in sound-to-meaning mapping and articulation.

In recent years, an increasing number of studies have started to investigate the brain regions involved in language comprehension by using more complex

sentences or stories. In addition to the previously mentioned brain regions, such as the precuneus, hippocampus, and middle and superior frontal areas, a broader range of regions have been reported to participate in semantic processing [2, 25]. Particularly in fMRI experiments using longer stories as stimuli, as shown in Fig. 2.2, this naturalistic paradigm has led to extensive activation in both hemispheres of the brain [9, 10, 26, 27]. These findings challenge the traditional view that semantic processing is predominantly associated with the left hemisphere. Moreover, the widespread activation across the entire brain seems to defy the explanatory power of the classical dual-stream model, which is primarily focused on specific regions in the left hemisphere. Although fMRI is useful in determining the specific brain regions involved in processing various aspects of language, such as phonology, semantics, and syntax, researchers are also interested in understanding the timeline and mechanisms of integrating these different aspects, from syllable sequences to phrases, and ultimately forming a comprehensible sentence. However, as language processing in the brain occurs within 100 milliseconds, and the BOLD signal generated by fMRI is based on fluctuations in blood oxygen level resulting from neuronal activity changes in distinct regions of the brain, it is not sensitive enough to capture such rapid neural activity.

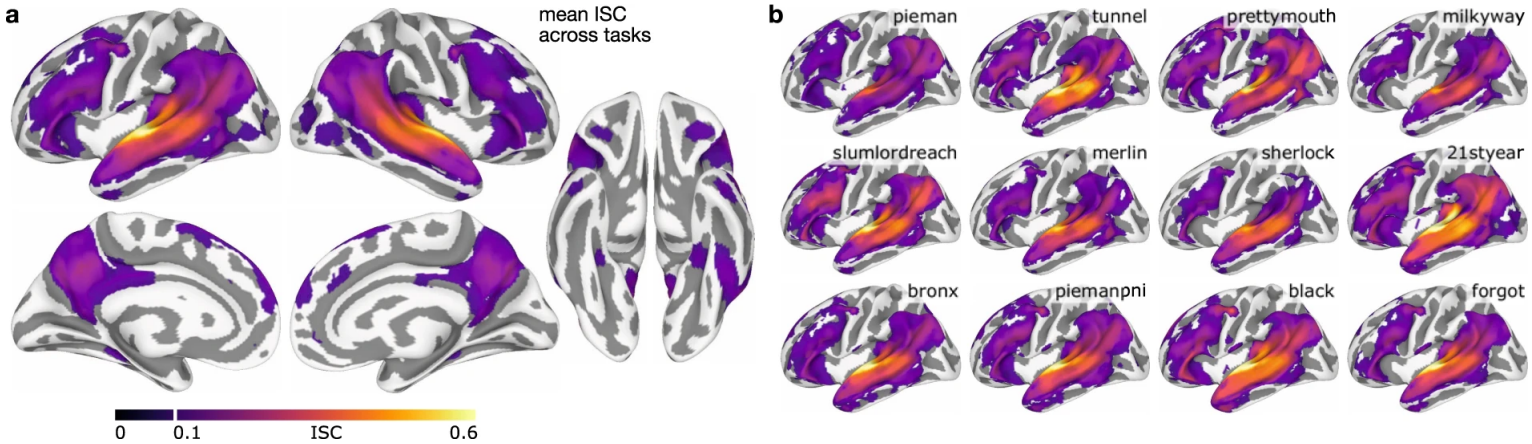

Figure 2.2: Naturalistic paradigm reveal more widespread responses to the speech comprehension process [26].

2.1.2 Insights from high temporal resolution

As mentioned earlier, language processing is a remarkably rapid process, with words being perceived and integrated into ongoing discourse in less than 600 milliseconds [28]. EEG, with its high temporal resolution, is a more suitable tool for capturing these fast and dynamic events compared to fMRI. Researchers commonly use event-related potentials (ERPs) derived from EEG to investigate language processing in the brain [4]. However, ERPs are not easily visible in raw EEG recordings due to their small amplitude [29]. Therefore, they are typically extracted by averaging multiple trials of the same stimulus from the continuous EEG recording. Figure 2.3 illustrates the analysis process involved in obtaining the N400 ERP component [30]. The N400 response is often observed when semantic incongruence occurs. During an experiment, semantic incongruent phrases are presented to subjects multiple times, and averaging the brain responses across these trials helps to reduce noise that is unrelated to the stimuli, allowing us to identify and study these N400 responses.

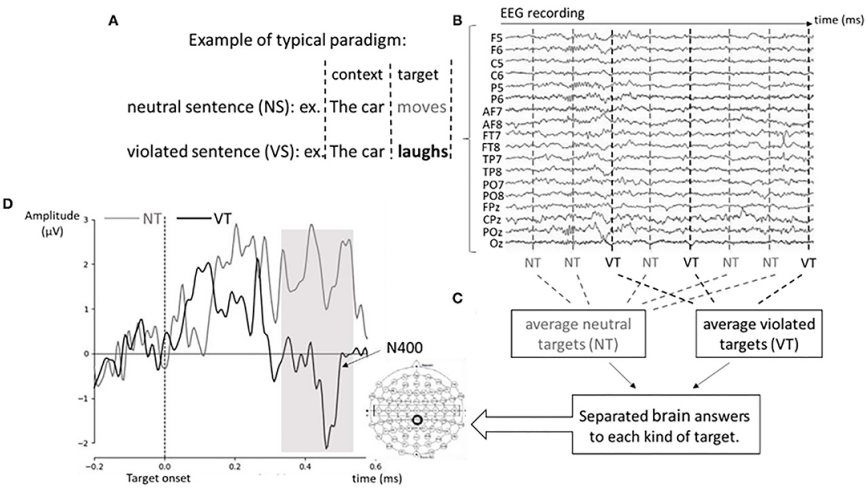

Figure 2.3: Analysis processes involved in obtaining the N400 ERP [30].

Based on ERPs, previous research has found some systematic responses to speech processing. Evoked responses occur around 50 ms after the onset of speech stimuli, and are modulated by low-level spectro-temporal information

of speech [31,32]. The following responses occur at 100 ms, which map acoustic information into phonetic features [33]. Once the phonetic features of words have been identified, their meaning can be retrieved from our memory. The N400 is a typical EEG response that corresponds to language semantic feature processing [34]. P600 responses are often reported in sentence-level processing, which may correspond to syntax information [35, 36].

Although EEG can provide a clear temporal profile of when different aspects of language are processed by the brain, it lacks sufficient spatial information to delineate which brain regions are involved in processing these language properties.

2.2 Beyond ERP and toward naturalistic experimentation

The shift towards naturalistic experimental settings has emerged from the recognition that language is rarely encountered in isolation but rather in rich and contextually meaningful contexts. As described in section 2.1.1, an increasing number of fMRI studies have shown that naturalistic experiments reveal different results compared to traditional isolated word research. By incorporating continuous speech and complete storylines into experiments, researchers can better capture the complexities of real-world language processing. This approach allows for a more ecologically valid examination of how individuals comprehend and make sense of language in everyday situations. However, this approach comes with significant challenges.

Traditional well-designed experimental paradigms often involve constructing a set of controlled experiments and using statistical techniques such as t-tests to compare differences between different control conditions and determine the mechanisms of the brain when processing different types of data. However, this approach is ineffective in most naturalistic stimulus experiments because they cannot control confounding or correlated variables. Moreover, it has been acknowledged by researchers that the response elicited by natural language at a given moment can encompass a sequence of processes

that are initiated at distinct and overlapping time points. Consequently, traditional ERP analysis methods are not suitable for naturalistic experimental paradigms [28].

Therefore, natural language experiments often employ alternative statistical techniques, with the most commonly used being encoding models in current mainstream research. These models typically have numerous free parameters estimated using a dataset called the training dataset. The free parameters are then fixed, and the encoding model is used to predict brain responses in a validation dataset that was not used during parameter estimation. The performance of the encoding model can be evaluated by comparing the predicted responses in the retained validation dataset with the actual responses (e.g., using Pearson correlation) [10]. Recently, encoding models have been applied to various language-related questions in EEG research [32, 37].

In encoding models, we can extract features from experimental stimuli using prior knowledge or some manually or automatically labeled methods. For example, in auditory experiments, temporal modulation information of speech signals can be extracted by leveraging the auditory mechanisms of the cochlea. These features can then be combined to construct a linear regression model of brain responses, which is also known as a multivariate temporal response function (TRF) [38]. Linearized models have demonstrated the existence of strong spectrotemporal and phonetic feature representations in brain regions such as the superior temporal gyrus and motor cortex [32,37,39].

Overall, the use of naturalistic experimental paradigms presents challenges for traditional analysis methods but has led to the development of encoding models as a popular approach in current research. These models allow researchers to extract features from stimuli and build regression models to predict brain responses. By leveraging these techniques, researchers can gain insights into the neural representation of language in the brain, considering the temporal and spectrotemporal aspects of language processing.

2.3 Significance of naturalistic experimentation

While the naturalistic paradigm presents various challenges in analysis methods compared to traditional experimental paradigms, it also addresses several limitations of previous research. Firstly, scientific findings are most valuable when they can be applied to broader contexts. Current results based on well-designed paradigms, such as isolated words or sentences, fail to generalize to the neural mechanisms underlying natural language stimuli [8, 9]. Secondly, in terms of experimental efficiency, well-designed paradigms often require controlling variables by keeping other factors constant, limiting the investigation of neural mechanisms to specific conditions. In contrast, naturalistic language experiments typically do not start with specific hypotheses. Instead, they extract relevant features from speech stimuli based on prior knowledge or assumptions and correlate them with brain signals. The use of continuous speech in experimental designs allows for the exploration of various linguistic phenomena, including prosody, intonation, and discourse structure, which play essential roles in conveying meaning and guiding comprehension [40]. Moreover, the development of statistical and signal-processing methods, along with advances in natural language processing tools, such as parsers and forced alignment systems, have made it easier than ever to annotate and analyze naturalistic language data [41]. While controlled laboratory experiments remain valuable, incorporating naturalistic designs into language research allows for a more comprehensive understanding of language processing in its full complexity.

In addition to enhancing our understanding of language processing, naturalistic studies have practical implications. They inform the development of more effective language learning interventions, speech recognition systems, and natural language processing algorithms. By incorporating realistic language contexts and considering the dynamic nature of communication, researchers can create robust models and systems that align better with human language understanding.

Overall, the shift towards naturalistic experimental settings in language research is a valuable development. By transcending isolated words and simple sentences, researchers can capture the complexities of language processing in real-world contexts. These studies provide insights into how individuals comprehend continuous speech, integrate prior knowledge, and utilize linguistic cues to extract meaning. With implications for both theoretical understanding and practical applications, naturalistic studies significantly contribute to advancing our knowledge of language comprehension.

2.4 Current research based on naturalistic experimentation

Based on the naturalistic experiment design. Previous research have got many exciting results. In this section, it will introduce these research and summarize the shortcomings of these research.

2.4.1 Encoding function TRF reflect various aspects of speech processing

In the study by Shamma et al. (2003), the estimation of encoding function TRF in the auditory cortex of ferrets revealed the connection between TRF and cognitive processing [42]. Subsequent research by Luo and Poeppel (2007) demonstrated that the phase pattern of theta band responses in the human auditory cortex tracked and synchronized with spoken sentences [43]. Ding (2012) further identified that these neural responses were influenced by the acoustic modulations of the temporal envelope of speech [37]. These findings set the stage for exploring the relationship between neural responses and various speech features.

Liberto (2015) utilized forced alignment tools to extract phonemes and phonetic features (Figure 2.4) from speech, highlighting the categorical nature of phonetic processing in the brain's neural responses [32]. Broderick (2018) extended the investigation by encoding speech features with EEG sig-

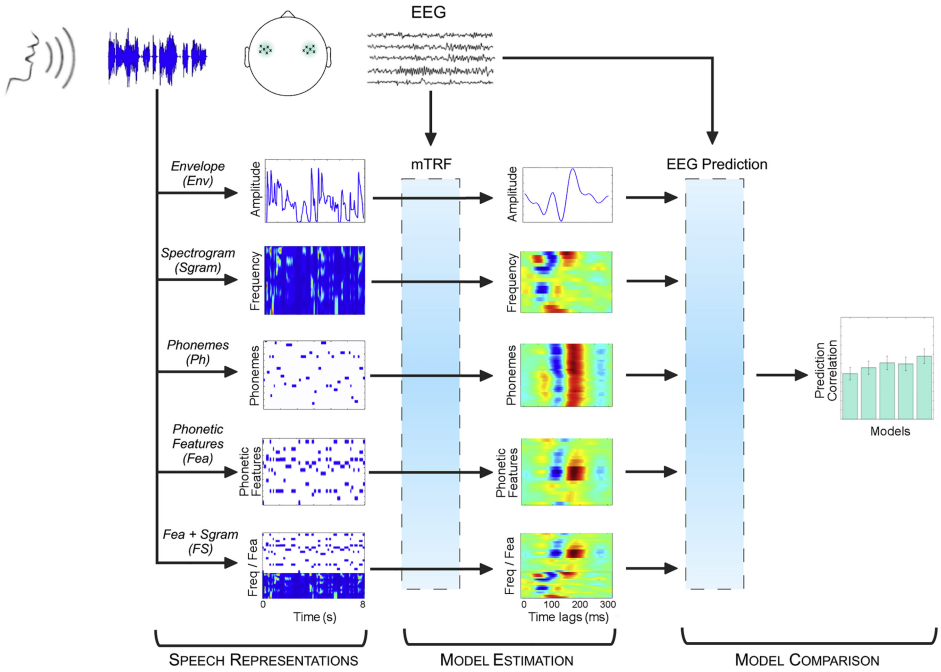

Figure 2.4: Encoding between speech phonetic feature and neural responses [32].

nals using word vectors (Figure 2.5), demonstrating that TRF derived from EEG signals could reflect the semantic processing of continuous speech [44]. Moreover, statistical probability models have been employed to represent linguistic-level features, such as word surprisal and semantic prediction, indicating that TRFs also capture linguistic-level processing. These findings collectively contribute to a deeper understanding of how the brain translates sound input into meaningful speech processing [45–47].

In summary, studies utilizing EEG have shown that neural responses reflect various aspects of speech processing, including the acoustic features of speech, phonetic processing, semantic processing, and even higher-level linguistic processing. These findings enhance our understanding of how the brain processes speech stimuli and pave the way for developing improved language learning interventions, speech recognition systems, and natural language processing algorithms.

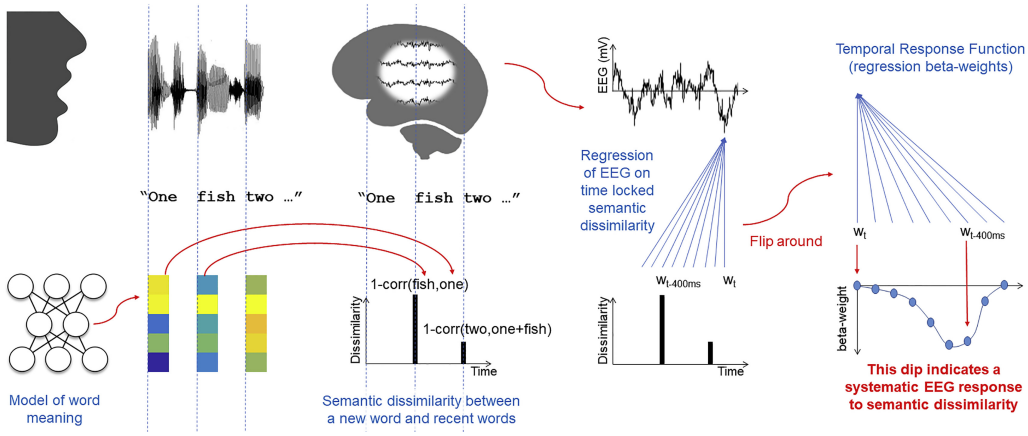

Figure 2.5: Encoding process from word vectors to EEG signals [44].

2.4.2 Role of neural oscillations in tracking speech

Recent research based on encoding function (TRF) has accumulated compelling evidence supporting the synchronization of brain oscillations with various speech features during the perception of speech. Notably, investigations have revealed that low-frequency neural oscillations in the theta and delta range exhibit synchronization with the dynamics of the speech envelope, corresponding to syllabic and phrasal rates, respectively [43, 48]. Conversely, high-frequency neural activity in the gamma range aligns with the fine-grained temporal dynamics associated with phonetic features [49]. These findings emphasize the role of neural oscillations in tracking and aligning with the fundamental properties of speech.

Moreover, experimental studies have shed light on the involvement of neural oscillations in the semantic processing of speech. Notably, it has been observed that neural entrainment to speech is stronger when the speech is easy to understand [39, 50]. This suggests that the synchronization of neural oscillations with speech enhances the comprehension and processing of semantic information.

The analysis of encoding function has provided valuable insights into the cortical dynamics underlying semantic processing and speech comprehension [36, 51]. Specifically, during sentence comprehension, the desynchronization

of neural oscillations in the alpha and beta frequency ranges has been linked to the engagement of task-relevant brain regions in supporting sentence-level processing. This desynchronization is thought to facilitate the efficient processing of semantic information within the sentence context.

Moreover, investigations into brain network functioning have uncovered intricate patterns of functional connectivity between the left inferior frontal and temporal cortex during the comprehension of sentences. Notably, granger causality analysis has revealed that alpha activity facilitates the transfer of information from the temporal to frontal regions, while beta activity supports information transfer in the opposite direction [52]. Additionally, synchronized beta and low-gamma oscillations have been observed between the left frontal and temporal regions, particularly when processing unexpected sentence-final words [53]. Furthermore, cross-frequency connectivity has been reported, demonstrating interactions between gamma power in the left prefrontal region and alpha power in the left temporal region, particularly during the processing of anticipated sentence-final words [51].

Overall, these findings point out the significance of the left inferior frontal and temporal cortex in speech processing. However, more research is necessary to systematically explore the mechanisms underlying inter-regional communication in the brain. Additionally, investigating the specific patterns of synchronization, such as amplitude synchronization, phase-locking and phase-amplitude coupling across various frequency bands, is crucial for a comprehensive understanding of these processes [28].

In summary, the investigation of oscillatory activity has significantly advanced our understanding of semantic processing, uncovering the intricate cortical dynamics and connectivity patterns that underlie speech comprehension. By elucidating the role of neural oscillations in tracking speech features and facilitating semantic processing, these studies provide valuable insights into the neural mechanisms underlying language processing.

2.5 Source Reconstruction Techniques

EEG/MEG is a non-invasive technique used to record brain activities by measuring voltage fluctuations on the scalp (or magnetic field changes in MEG), which directly reflect the biophysical phenomena of populations of neurons [54]. However, the signals measured on the scalp represent a mixture of many cortical responses, making it challenging to determine the specific cortical origins underlying speech processing. To investigate speech encoding and decoding processes at the cortex level, source reconstruction techniques can be employed to estimate the original cortical responses from scalp recordings. In this section, we will introduce the source reconstruction technique and explain why standardized low-resolution electromagnetic tomography (sLORETA) is chosen as our source reconstruction method.

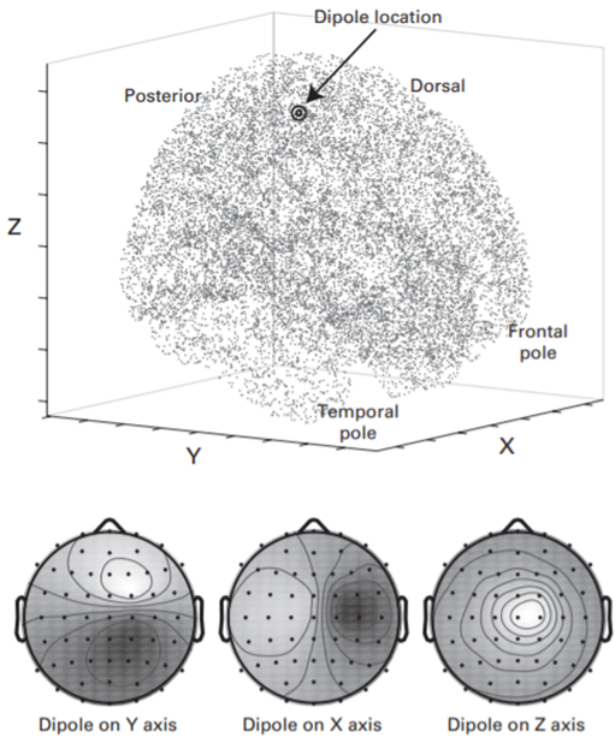

Figure 2.6: An example of a forward model [54].

2.5.1 Forward Solution

Accurate source localization heavily relies on the forward head model. To illustrate this, let's consider placing several small probes into the brain that can transmit a radio signal. If these probes are activated and their signals are recorded on the scalp, what would the recorded signals look like? The answer to this question lies in the forward solution. The shape of the head and the conductivity of the skull strongly influence the forward head model. Therefore, an accurate MRI-derived boundary head model and precise electrode positions are essential for obtaining a reliable forward solution [55, 56]. This enables the study and understanding of brain activity with better spatial resolution.

It's important to note that these probes (dipoles) do not transmit the radio signal in a specific direction but rather in all directions unevenly. Thus, different orientations of the dipoles are modeled in many forward solutions. Typically, three fixed orientations that are perpendicular to each other are considered. Figure 2.6 shows an example of topographical maps on the scalp for three different dipole orientations [54]. A brain source head model consisting of 15,028 locations (each gray dot represents a location) was constructed from an MRI image. The forward model from brain sources into 64 channel electrodes was computed for three orientations at each source location. The figure demonstrates how the scalp signal appears from different dipole orientations when the dipole is activated.

2.5.2 Inverse Solution

The inverse solution aims to determine the locations, magnitudes, and orientations of dipoles within the head based on the observed scalp signals. It is the inverse problem of the forward solution. Figure 2.7 illustrates the relationship between the inverse and forward solutions [57]. Similar to most inverse problems, there is no unique solution for this inverse model, making it an ill-posed problem from a theoretical standpoint. Estimating the states of tens of thousands of brain sources from just a few hundred scalp electrodes becomes particularly challenging as the number of brain sources exceeds the

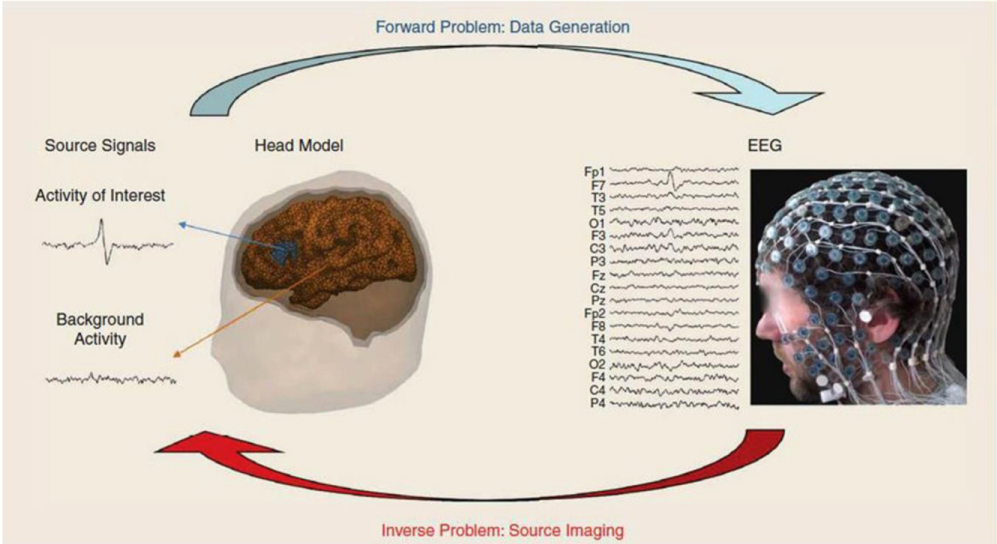

Figure 2.7: An illustration of the relationship between the forward and inverse solutions [57].

available measurements. At a practical level, all methods for estimating the inverse solution require several parameter selections that impact the source reconstruction results.

2.5.3 Common Source Reconstruction Methods

2.5.3.1 Dipole Fitting

Dipole fitting aims to estimate a small number of discrete dipoles in the brain that can explain the maximum amount of topographical potential. Once the dipole locations, orientations, and magnitudes are estimated, the dipole activity can be calculated for all electrodes. Dipole fitting is commonly used in event-related potential (ERP) data analysis. However, before employing dipole fitting, it is necessary to determine the number of dipoles to be estimated. Estimating too many dipoles can lead to suboptimal results due to the potential of getting "stuck" in local minima and poor source reconstruction [54]. Furthermore, since the inversion results are not unique, similar topographical distributions can be produced even with a different number of dipoles. Figure 2.8 illustrates the situation where one dipole and

two dipoles produce similar patterns [54].

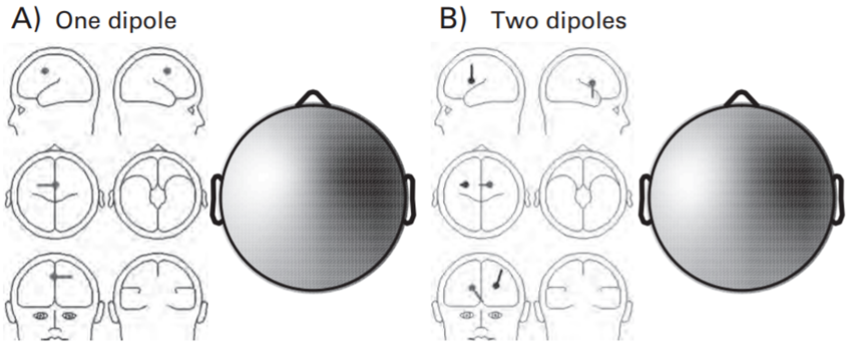

Figure 2.8: Similar topographical distributions produced by one dipole or two dipoles [54].

2.5.3.2 Distributed-Source Imaging

Distributed-source imaging differs from dipole fitting as it involves estimating a large number of fixed location and orientation dipoles, while only the magnitudes of these dipoles need to be estimated. Non-adaptive distributed-source imaging methods establish the mapping between electrodes and dipoles based solely on electrode locations, rendering the electrode signals independent of the source reconstruction results. The advantage of non-adaptive methods lies in their quick computation and stable results since only a few parameters need to be determined. Adaptive distributed-source imaging, on the other hand, also considers the electrode signals (such as frequency and amplitude) in the mapping between electrodes and dipoles. This adaptation to the data enables the method to reflect spectro-temporal information, experimental conditions, or subject-specific factors, providing more information for source reconstruction results. However, adaptive distributed-source imaging is susceptible to noise in the data, which can impact the source reconstruction outcomes.

2.5.3.3 Why choose sLORETA

In this study, we employed a naturalistic experimental design, which is prone to unexpected noise. Therefore, adaptive distributed-source imaging is not suitable. Additionally, dipole fitting is limited in investigating whole-brain level responses. For these reasons, we applied the standardized low-resolution electromagnetic tomography (sLORETA) method to obtain plausible EEG source estimates [58]. sLORETA has been widely used, especially in clinical applications, over the past decade [59–64]. Although sLORETA has a lower spatial resolution, it provides smooth and accurate localization with minimal errors [65].

2.6 Limitation in previous research

Although EEG is an effective and non-invasive technique for investigating the neural mechanisms behind auditory processing, previous studies on natural spoken language have primarily focused on the electrode/sensor space (scalp-level) due to the low spatial resolution of EEG/MEG [32,44,45,66]. However, the mixture of source components at the scalp level makes it challenging to explain the cortical origins of underlying natural spoken language processes. With advancements in EEG signal processing, it has been demonstrated that with a sufficient number of sensors or an accurate individual head model, EEG source localization can provide precise enough information to reflect the cortical origins of language processing [67]. Moreover, exploring brain functions in response to continuous speech explicitly benefits from studying EEG signals in the source space, as the generators of neural activity cannot be unambiguously interpreted from sensor-level data alone [68].

Recent studies using EEG source localization techniques have shown exciting results in both speech production and speech perception [1, 68, 69]. However, most of these studies have been based on the ERP paradigm [70]. In the context of the natural speech paradigm, where stimuli are typically long segments from lectures or stories presented only once to avoid priming effects, there are two key challenges that need to be addressed for single-trial

analysis before source localization. Firstly, the generated electrical fields are susceptible to contamination from external noise (e.g., eye movement, head movement) during the transmission from the neural population to the scalp. Naturalistic experiments often involve more complex and varied stimuli, leading to increased signal variability and potential artifacts. These factors can negatively impact the accuracy and reliability of source reconstruction results. EEG signals recorded during naturalistic tasks are typically noisier compared to controlled laboratory experiments, making it more challenging to obtain accurate source estimates. Reconstructing a single source from a single trial and fitting encoding function TRF directly to cortical sources may be affected by the unexpected noise, thereby impacting the accuracy and interpretability of the encoding function. Additionally, most of the existing source localization techniques have been developed for the ERP paradigm, assuming spatiotemporal sparsity [65, 71–74]. However, the natural speech paradigm does not allow for additive averaging across repeated trials.

In this paper, we aim to explore the speech encoding and decoding processes at the cortex level by EEG using proposed hyper-alignment methods. By examining encoding function throughout the entire brain, we can gain insights into how different cortical regions are involved in the comprehension of spoken language. This approach will provide a more comprehensive understanding of the neural mechanisms underlying language processing, shedding light on the distributed nature of speech comprehension across the cortex.

2.7 Summary

In this chapter, we have reviewed the literature on naturalistic experimentation in the study of language processing and comprehension. We highlighted the shift from traditional isolated word or simple sentence paradigms to more ecologically valid experimental settings that involve continuous speech and complete storylines. This shift has been motivated by the recognition that language is encountered in rich and meaningful contexts, and investigating language processing in such naturalistic settings provides a more compre-

hensive understanding of how individuals comprehend and make sense of language.

We discussed the challenges posed by naturalistic experiments, including the inability to control confounding variables and the limitations of traditional analysis methods such as ERP analysis. To overcome these challenges, researchers have turned to encoding models, which involve extracting features from stimuli and building regression models to predict brain responses. These models have been successful in capturing various aspects of speech processing, including acoustic features, phonetic processing, semantic processing, and higher-level linguistic processing.

We also discussed the significance of naturalistic experimentation, emphasizing the need for findings that can be applied to broader contexts and the practical implications for language learning interventions, speech recognition systems, and natural language processing algorithms. By incorporating realistic language contexts and leveraging advancements in statistical and signal-processing methods, researchers can better understand the complexities of language processing and improve the alignment between computational models and human language understanding. Furthermore, we addressed the limitations of previous research, particularly the focus on the electrode/sensor space and the challenges associated with source localization in naturalistic experiments. We highlighted the importance of studying EEG signals at the cortex level to better understand the cortical origins of language processing.

In summary, naturalistic experimentation provides valuable insights into the neural mechanisms underlying language processing. By incorporating continuous speech and complete storylines, researchers can capture the complexities of real-world language comprehension. Future research should focus on addressing the challenges of source localization in naturalistic experiments and exploring the contributions of different cortical regions to speech processing. Additionally, efforts should be made to develop robust and reliable methods for analyzing and interpreting EEG signals at the cortex level.

Chapter 3

Methodology

3.1 Research philosophy

The research philosophy underlying our study is guided by the following principles, which is shown in Fig. 3.1. Firstly, we focus on the process of speech comprehension, aiming to understand the intricate mechanisms involved in how the human brain comprehends speech. We recognize the initial processing of speech by the auditory system, which leads to the extraction of temporal amplitude envelopes (TAEs). These TAEs contain important information for subsequent semantic processing [13].

Our hypothesis posits that the semantic information within the TAEs is encoded by specific brain areas responsible for semantic processing. We propose a linear time-invariant system to explain this encoding process, where the semantic content of speech is transformed into the system output. In our study, we employ EEG to record the system output (neural response).

By leveraging the linearity of the encoding process, we propose the possibility of reversing the transformation and recovering the original speech TAEs from the observed semantic representation. This enables us to decode the semantic information and reconstruct what was heard solely from brain activity, even in the absence of direct auditory perception.

By accurately estimating the encoding function, we aim to decode and recover the semantic information, providing insights into the neural mechanisms underlying speech comprehension. This research philosophy allows us to explore the possibility of inferring speech content solely from brain activity, expanding our understanding of speech processing in the brain.

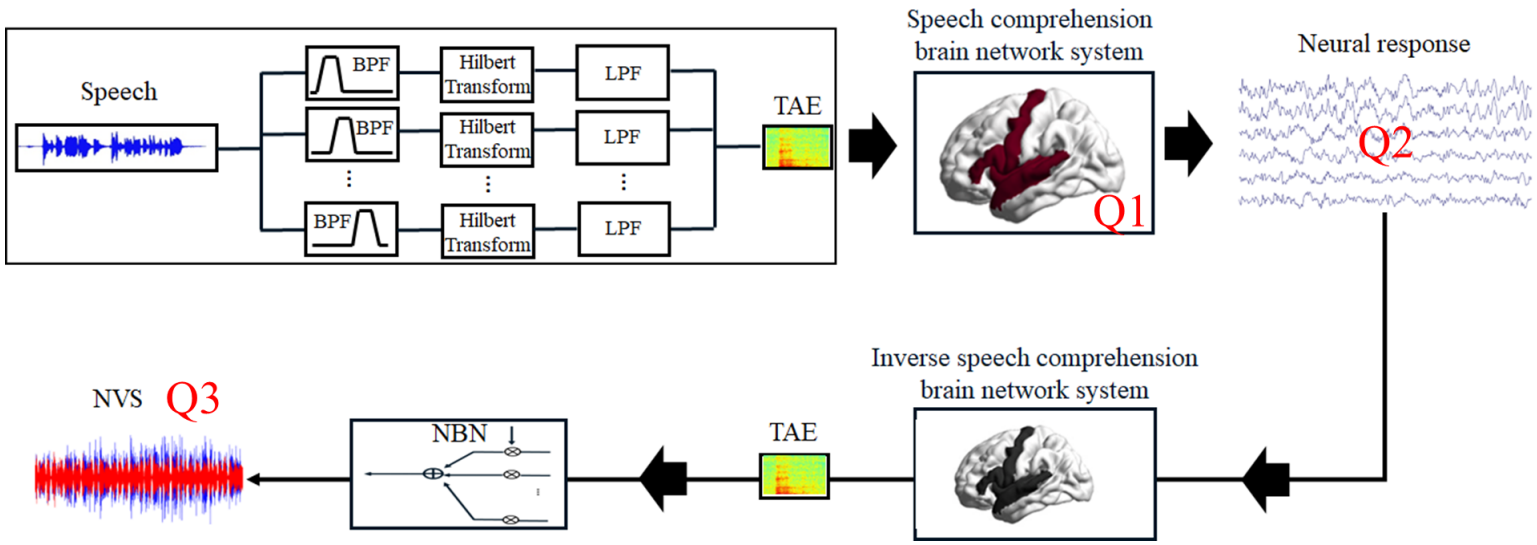

Figure 3.1: Research philosophy in this study.

To validate this research philosophy, our study aims to address three proposed key questions in section 1.5:

- Localization of the brain system for semantic processing: We want to identify the specific brain regions that play a role in processing semantic information during speech comprehension (Q1).
- Identification of the semantic representation in the system output: We seek to ascertain whether the semantic content of speech can be discerned in the output of the encoding system (Q2).
- Recovery of semantic information from the semantic representation: We aim to assess the viability of accurately retrieving the original semantic information from the observed semantic representation, thereby demonstrating the efficacy of the encoding and decoding process (Q3).

By addressing these questions, our study aims to provide valuable insights into the neural underpinnings of speech comprehension and shed light on the potential for decoding semantic information from brain activity.

3.2 Overview of the study

During speech perception, information is transmitted rapidly and continuously, necessitating methods with high temporal resolution to capture the temporal dynamics between speech signals and brain activity. To address this requirement, we first conducted a speech perception experiment and collected EEG data to model the process of speech encoding and decoding, as illustrated in Fig. 3.2. In Chapter 4, we provide a detailed account of the experimental design and EEG data acquisition procedure.

As previously discussed, the mixture of source components at the electrode/sensor level in EEG signals presents challenges when elucidating the cortical origins of natural spoken language processes. Therefore, we proposed a method to transfer EEG signals from the scalp level to the brain cortex level. This approach aimed to overcome this challenge and enhance our understanding of cortical involvement in speech processing.

At the cortex level, we used LTI models to explore the relationship

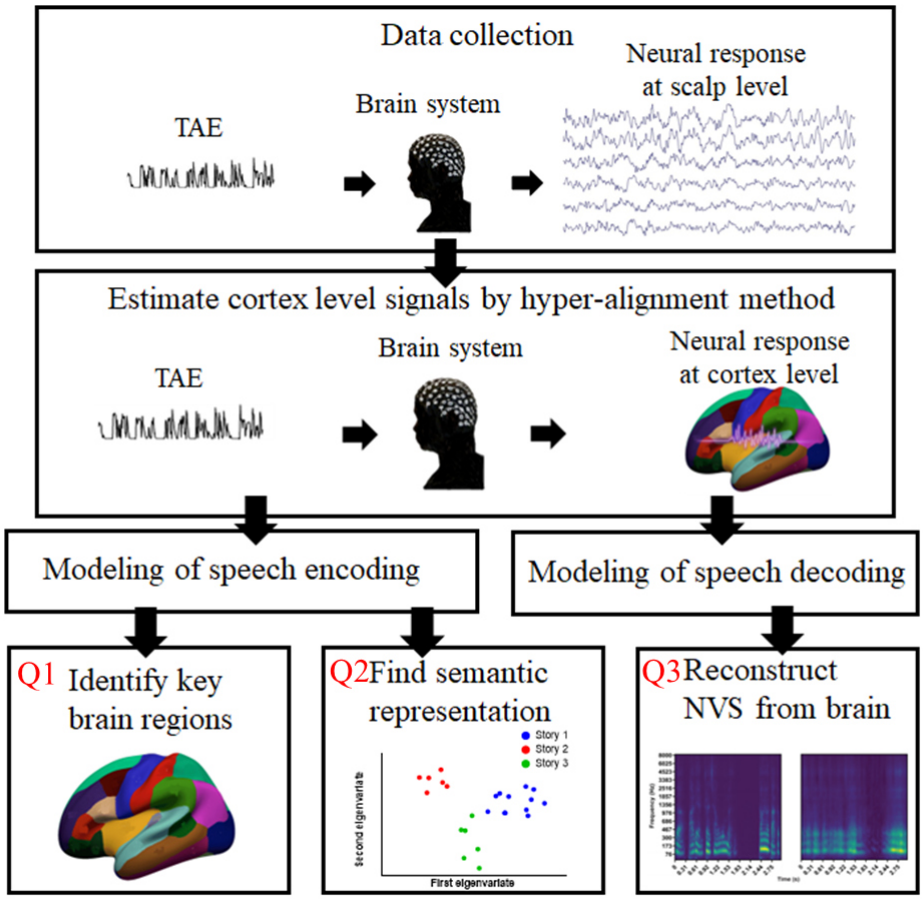

Figure 3.2: Overview of the procedure in this study.

between speech temporal amplitude envelopes (TAEs) and cortex signals. In the speech encoding process, we identified key brain regions involved in semantic processing, addressing the first question (Q1) regarding the localization of the brain system responsible for semantic processing. Additionally, we aimed to demonstrate the existence of semantic representation within these brain regions, which addresses the second question (Q2). Our study's results related to these inquiries will be presented in Chapter 5.

During the speech decoding process, our main goal was to reconstruct the neural vocabulary semantics (NVS) from brain signals, with the objective of recovering semantic information from the brain activations. This approach addresses the third question (Q3) regarding the ability to retrieve semantic

information from the semantic representation. The results obtained from this investigation will be discussed in detail in Chapter 6.

By systematically addressing these questions and implementing rigorous methodologies, our study aimed to provide valuable insights into the neural underpinnings of semantic comprehension and the potential for decoding semantic information from EEG signals. These findings have the potential to contribute to the development of neurocognitive models of language processing and advance the field of brain-computer interfaces for natural language communication.

3.3 Methodology to improve spatial resolution of EEG

3.3.1 Noise Reduction for EEG

External noises resulting from eye movements, heartbeat, electrical interference, and other sources can contaminate EEG signals during their transmission from the neural population through the brain tissue and skull [54]. These noises are often treated as random and can adversely affect the accuracy of source reconstruction. In ERP analysis, researchers commonly employ an averaging operation across multiple trials of the same task to mitigate these noise sources and improve the accuracy of source reconstruction.

In our study, we adopt a similar approach to reduce noise by applying additive averaging to EEG signals elicited by the same stimulus material across all subjects. Assuming that the brain functions for speech processing are consistent across individuals, a similar neural response can be expected from different subjects for the same speech stimulus. In contrast, external noise, involuntary breathing, and attentiveness differ from individual to individual, and such noises can be suppressed by averaging the neural signals of the same stimuli for all subjects.

To apply additive averaging across subjects, it is important to account for individual differences, such as head shape, cortical location, and setup

positions of the electrodes. These factors can introduce variability in the EEG signals and potentially impact the accuracy of source localization. To address this problem, we propose a functional hyper-alignment method for soft calibration. This method aims to reduce the mismatch caused by individual experiment settings and improve the accuracy of source localization. By aligning the functional data across subjects, we can better account for individual differences and obtain more accurate source estimates. It uses a well-designed spatial filter to align the setup positions of electrodes by minimizing the distance of the signal features among the subjects. For example, due to the lack of methods that account for subjects' differences in the setup stage of the electrodes, the position of an electrode n for subject i may not be the same as that of subject j. Thus, the additive average over the EEG data $x_i(t,n)(i = 1, 2, \ldots, I)$ cannot be used to perform denoising properly, where I is the subject number. For this reason, we propose using a functional hyper-alignment method for eliminating this effect. The main idea of the functional hyper-alignment is to rotate $x_i(t,n)$ and $x_j(t,n)$ $(i \neq j \in [1, 2, \ldots, I])$ to maximize their correlation among subjects. So far, several methods have been proposed for this purpose, such as group task-related component analysis (gTRCA) [75] and multi-set canonical correlation analysis (MCCA) [76]. We choose MCCA to maximize the data correlation among subjects, which satisfies the requirement of our study.

Here we first briefly review the canonical correlation analysis (CCA). Consider the EEG data X_1 and X_2 from two subjects for the same stimulus, the size X_1 and X_2 are $T \times N$ where T is data length and N is the number of channels. For simplicity, all data are assumed to have zero average value. Assuming vector (spatial filters) ω_1 and ω_2 exist, which can linear transform the X_1 and X_2 to $\widetilde{X}_1$ and $\widetilde{X}_2$ by

$$\widetilde{X}_1 = {\omega_1}^T X_1, \widetilde{X}_2 = {\omega_2}^T X_2. \tag{3.1}$$

Since X_1 and X_2 are the neural responses for the same stimulus, they should be almost the same if they were obtained in the same location of the scalp, and thus they should have higher correlation. Therefore, the goal of CCA

attempt to find optimal spatial filters ω_1 and ω_2 to maximize correlation coefficient ρ of $\widetilde{X}_1$ and $\widetilde{X}_2$. When average value of X_1 and X_2 is zero, the correlation coefficient ρ of $\widetilde{X}_1$ and $\widetilde{X}_2$ can be calculated by

$$\rho\left(\widetilde{X}_1, \widetilde{X}_2\right) = \frac{E(\widetilde{X}_1\widetilde{X}_2)}{\sqrt{E(\widetilde{X}_1^2)}\sqrt{E(\widetilde{X}_2^2)}} = \frac{{\omega_1}^T V_{x_1x_2}\omega_2}{\sqrt{{\omega_1}^T V_{x_1x_1}\omega_1}\sqrt{{\omega_2}^T V_{x_2x_2}\omega_2}}, \tag{3.2}$$

where $V_{x_1x_2}, V_{x_1x_1}, V_{x_2x_2}$ is the variance-covariance matrix of X_1 and X_2. It is no problem to normalize the denominator of ${\omega_1}^T V_{x_1x_1}\omega_1 = {\omega_2}^T V_{x_2x_2}\omega_2 = 1$, therefore the solution for maximize the $\rho\left(\widetilde{X}_1, \widetilde{X}_2\right)$ changed to the quadratic programming with equality constraints, where

$$\begin{aligned} \arg\max_{\omega_1,\ \omega_2}\ & {\omega_1}^T V_{x_1x_2}\omega_2 \\ s.t.\ & {\omega_1}^T V_{x_1x_1}\omega_1 = 1 \\ & {\omega_2}^T V_{x_2x_2}\omega_2 = 1. \end{aligned} \tag{3.3}$$

Lagrange multiplier can be used to solve this quadratic programming problem, where

$$L\left(\omega_1, \omega_2, \lambda_{\omega_1}, \lambda_{\omega_2}\right) = {\omega_1}^T V_{x_1x_2}\omega_2 + \lambda_{\omega_1}\left(1 - {\omega_1}^T V_{x_1x_1}\omega_1\right) + \lambda_{\omega_2}\left(1 - {\omega_2}^T V_{x_2x_2}\omega_2\right), \tag{3.4}$$

the differential of ω_1 and ω_2 is

$$\begin{aligned} \frac{\partial L}{\partial \omega_1} &= V_{x_1x_2}\omega_2 - 2\lambda_{\omega_1} V_{x_1x_1}\omega_1 = 0, \\ \frac{\partial L}{\partial \omega_2} &= {V_{x_1x_2}}^T\omega_1 - 2\lambda_{\omega_2} V_{x_2x_2}\omega_2 = 0. \end{aligned} \tag{3.5}$$

According to the differential of ω_1 and ω_2 in Eq. 3.5, we multiply both sides of the equation by ${\omega_1}^T$ and ${\omega_2}^T$, it can get

$$\begin{aligned} {\omega_1}^T V_{x_1x_2}\omega_2 - 2\lambda_{\omega_1}{\omega_1}^T V_{x_1x_1}\omega_1 &= 0, \\ {\omega_2}^T {V_{x_1x_2}}^T\omega_1 - 2\lambda_{\omega_2}{\omega_2}^T V_{x_2x_2}\omega_2 &= 0. \end{aligned} \tag{3.6}$$

It's known that $({\omega_1}^T V_{x_1x_2}\omega_2)^T = {\omega_2}^T {V_{x_1x_2}}^T \omega_1$, and ${\omega_1}^T V_{x_1x_1}\omega_1 = {\omega_2}^T V_{x_2x_2}\omega_2 = 1$ according to Eq. 3.3. It is easy to get $2\lambda_{\omega_1} = 2\lambda_{\omega_2} = \lambda$. Therefore, Eq. 3.5 can be be summarized into a generalized eigenvalue problem, where

$$\begin{bmatrix} O & V_{x_1x_2} \\ {V_{x_1x_2}}^T & O \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} = \lambda \begin{bmatrix} V_{x_1x_1} & O \\ O & V_{x_2x_2} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}. \tag{3.7}$$

According to the solution of generalized eigenvalue problem [77], we can know that the eigenvalue λ for generalized eigenvalue problem is the correlation ρ of $\widetilde{X}_1$ and $\widetilde{X}_2$. The corresponding eigenvector is the spatial filters ω_1 and ω_2 which can linear transform the X_1 and X_2 to $\widetilde{X}_1$ and $\widetilde{X}_2$.

This kind of idea can be extended to multi-subjects. The goal of MCCA is to find projection vectors ω that maximize the correlation between multiple data sets $X_i, i = 1, 2, \ldots, I$. The correlation ρ of all data sets can be calculated as the ratio of the summations of the between-set covariance $V_{x_ix_j}$ over the within-set covariance $V_{x_ix_i}$,

$$p(\widetilde{X}_1, \widetilde{X}_2, ..., \widetilde{X}_i, ..., \widetilde{X}_I) = \frac{1}{N-1} \frac{\sum_{i=1}^{I} \sum_{j=1, i\neq j}^{I} \omega_i^T V_{x_ix_j} \omega_j}{\sum_{i=1}^{I} \omega_i^T V_{x_ix_i} \omega_i}, \tag{3.8}$$

where

$$V_{x_ix_j} = (X_i - \bar{X}_i)^T (X_j - \bar{X}_j), \tag{3.9}$$

$$V_{x_ix_i} = (X_i - \bar{X}_i)^T (X_i - \bar{X}_i). \tag{3.10}$$

$\bar{X}_i$, $\bar{X}_j$ are the means for set i and set j. $\frac{1}{N-1}$ ensures that the correlation ρ scales between 0 and 1. Altogether, the above equation can be summarized into a generalized eigenvalue problem,

$$B\omega = \lambda R\omega, \tag{3.11}$$

where

$$B = \begin{bmatrix} O & V_{x_1x_2} & \cdots & V_{x_1x_I} \\ V_{x_2x_1} & O & \cdots & V_{x_2x_I} \\ \vdots & \vdots & \ddots & \vdots \\ V_{x_Ix_1} & V_{x_Ix_2} & \cdots & O \end{bmatrix}, R = \begin{bmatrix} V_{x_1x_1} & O & \cdots & O \\ O & V_{x_2x_2} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & V_{x_Ix_I} \end{bmatrix}. \quad (3.12)$$

B is a matrix combining all between-set covariance $V_{x_ix_j}$, and R is a diagonal matrix that contains all within-set covariance $V_{x_ix_i}$. ω is a spatial vector set for an entire data set $\omega = [\omega_1^T, \omega_2^T, \ldots, \omega_I^T]$. Finally, the spatial filter for aligning the positions of the electrodes is reduced to solve the generalized eigenvalue problem.

3.3.2 Source reconstruction based on hyper-alignment EEG data

After the hyper-alignment, the EEG data are used to estimate their cortical source activations in the brain. In this study, the forward and reverse models for source localization were calculated by the Brainstorm toolbox [78]. The finite element method (FEM) as implemented in DUNEuro was used to compute the forward head model using Brainstorms default parameters with a MNI MRI template (ICBM152) [79, 80]. The FEM models provide more accurate results than the spherical forward models and more realistic geometry and tissue properties than the boundary element method (BEM) methods [81]. For source estimation, the number of potential sources (grid on the cortex surface) is set to 15,002. And the option of constrained dipole orientations was selected, which means dipoles are oriented perpendicular to the cortical surface. [78]. We then apply the method of standardized low-resolution electromagnetic tomography (sLORETA) [58] to obtain plausible EEG source estimates. Although the spatial resolution of sLORETA is low, sLORETA can provide smooth and good localization with few localization errors [65]. Finally, according to the Desikan-Killiany Atlas (DKA), the cortical surface is divided into 68 anatomical regions of interest (ROIs) [82]. The time series of each ROI is calculated from the average value of all dipoles

in the respective region. As a result, we obtain a series of brain areas (sources) that are activated during speech comprehension, providing sufficient spatial resolution for our study.

3.4 Methodology to identify brain regions involved in semantic processing

In our experiment, participants were presented with two types of auditory signals: normal natural speech and time-reversed speech. The aim was to investigate participants' ability to understand the content of the normal speech signal, which contained semantic information, compared to the time-reversed speech signal, which lacked semantic information.

We hypothesized that participants would be able to comprehend the content of the natural speech signal but would struggle to understand the time-reversed speech due to the absence of semantic information. By examining the brain encoding processes associated with these two types of speech signals, we aimed to identify the specific brain regions involved in semantic processing.

Analyzing the brain encoding patterns allowed us to determine the regions that showed differential activation or response to the natural speech signal compared to the time-reversed speech. These identified regions would provide valuable insights into the neural mechanisms underlying semantic processing during speech comprehension.

3.5 Methodology to validate semantic representation in brain responses

To determine the presence of semantic representation in brain responses, we used a methodology based on the bidirectional encoder representations from transformers (BERT) language model [83]. We transformed speech stimuli into word vectors that contain semantic information using BERT. Since

different speech stimuli carry different semantic information, different stories or narratives would be distributed across distinct clusters in the semantic space.

This study hypothesized that neural responses would exhibit similar distribution patterns in the spatial domain since brain responses are influenced by these semantic information stimuli. Thus, we formally hypothesized that these semantic representations could be found within the brain responses. To validate this hypothesis, we performed the following steps.

We initially preprocessed the speech stimuli by encoding them into word vectors using the BERT language model. This process captured the semantic information contained within the speech and transformed it into a numerical representation (word vectors). Using the word vectors obtained from the speech stimuli, we mapped the semantic information onto a spatial representation. This allowed us to visualize the distribution of different stories or narratives in the semantic space.

This study then analyzed the participants' brain responses by examining the EEG signals and identifying neural patterns associated with semantic processing. We applied statistical techniques, such as multivariate pattern analysis or machine learning algorithms, to detect the presence of semantic representation within the brain responses. To ensure the reliability and generalizability of our findings, we performed cross-validation procedures. This involved dividing the data into training and testing sets and evaluating the performance of our models on independent datasets. Cross-validation helped validate the robustness of the observed semantic representations in the brain responses.

We compared the distribution patterns of semantic information in the brain responses with the distribution patterns in the semantic space. This allowed us to assess the similarity and consistency between the two representations, further supporting the presence of semantic representation in the brain responses.

3.6 Methodology to recover the semantic information from brain responses

Once we have identified the brain regions involved in semantic processing and confirmed the presence of semantic representation within these regions, we can leverage this knowledge to recover the semantic information from brain responses.

Based on our assumption that the transformation from TAEs to neural responses is a linear process, we can engineer a reverse system that reconstructs the original speech TAEs from the input of semantic representations in these brain areas. This reverse system incorporates statistical models or machine learning algorithms to infer the TAEs from the given input.

Using the engineered reverse system, we reconstructed the TAEs by leveraging the semantic representations present in the targeted brain areas. The NVS, derived from the brain responses, provided the necessary information to recover the original speech information with high intelligibility [84]. To assess the quality and intelligibility of the recovered speech, we conducted subjective evaluations for speech intelligibility ratings. These evaluations helped validate the effectiveness of the reverse system in recovering semantic information from brain responses.

By following this methodology, we aimed to demonstrate the feasibility of recovering the semantic information from brain responses and reconstructing the original speech TAEs. The reverse system, utilizing the observed semantic representations in the identified brain areas, allowed us to bridge the gap between brain activity and meaningful speech content.

3.7 Analysis methods in encoding process

3.7.1 Extraction of TAEs

According to the human peripheral system, the TAEs are obtained from a gammatone filterbank followed by a power law [85–87].

Speech signals were divided into several frequency bands by using a band-pass filterbank. The band-pass filters were determined based on the ERB_N (Equivalent Rectangular Bandwidth) and ERB_N-number scale, which corresponds to the distance scale along the basilar membrane. In this study, the numbers of channels of the band-pass filterbank were 16. Then, the TAE was extracted from each filter's output using Hilbert transformation and performing a 64 Hz low-pass filter [84]. For the modeling of encoding process, the TAE is then decimated to the same sampling rate as the source signal, enabling us to relate its dynamics to the source signal.

3.7.2 Extraction of semantic representation

BERT is pre-trained on large-scale corpora and can generate high-quality word embeddings that capture the semantic and syntactic properties of words. These embeddings can be used to extract semantic representations from text data. BERT has been shown to achieve state-of-the-art performance in various language understanding and generation tasks, including text classification, named entity recognition, and question answering [83]. To extract the semantic information from the speech stimuli in our study, we utilized pre-trained Chinese BERT [88]. We employed final layer of BERT at the sentence level to extract 768-dimensional word vectors. The average word vector across all sentences within a trial was taken as the representation of the semantic information for that specific trial. This approach allowed us to obtain a comprehensive semantic representation for each trial based on the extracted word vectors.

3.7.3 Modeling of speech encoding process

In this study, a linear model is used to discribe the speech encoding process. The main principle is to treat the brain as a linear time-invariant (LTI) system where the output (neural response) of the system is the convolution of the input and an encoding function (TRF) of the brain, which is shown in Fig. 3.3. The encoding function can be considered a filter that linearly transfers the continuous speech TAE to the dynamic neural response. Let

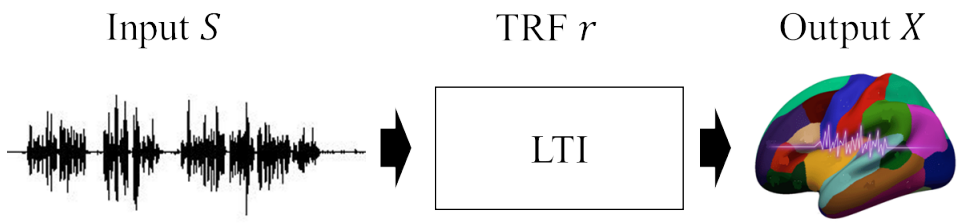

Figure 3.3: LTI model for speech encoding.

$r(\tau, ROI_n)$ be the encoding funcyion of a brain region ROI_n for an input speech TAE $s(t)$, the neural response signal $x(t, ROI_n)$ of the source ROI_n can be described as follows.

$$x(t, ROI_n) = \sum_{\tau} r(\tau, ROI_n) s(t - \tau). \tag{3.13}$$

The optimal encoder $r\left(\tau, ROI_n\right)$ is acquired by minimizing the mean square error (MSE) between the original source signal $x(t, ROI_n)$ and predicted source signal $\hat{x}(t, ROI_n)$, where

$$\arg\min_{r} \sum_{t} [x(t, ROI_n) - \hat{x}(t, ROI_n)]^2, \tag{3.14}$$

which is a linear regression problem. According to [89], the solution of $r\left(\tau, ROI_n\right)$ can use the following matrix operations

$$r = \left[S^T S\right]^{-1} S^T X, \tag{3.15}$$

where S is the time-lag series of the speech stimulus, is defined as

$$\begin{bmatrix} s(0) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ s(\Delta t) & s(0) & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & s(\Delta t) & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \cdots & s(0) & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & s(\Delta t) & s(0) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & s(\Delta t) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & s(0) \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & s(\Delta t) \\ s(T) & s(T-\Delta t) & \cdots & s(T-i\Delta t) & s[T-(i+1)\Delta t] & \cdots & s(T-\tau_{max}) \end{bmatrix}, \tag{3.16}$$

where the value τ_{max} represent the range $[0, \tau_{max}]$ of time lags τ. Δt is the sample period. Variable X is a matrix containing all the neural response data. The r is a $\tau_{max} \times N$ matrix, where N is the number of ROI, each column represents the univariate mapping from s to the neural response at each brain source. The range for τ is from 0 to 800 ms in this study, as most common ERP components in language research are within 800 ms [4].

3.7.4 Common spatial pattern analysis

To identify which parts of the brain are crucial for semantic processing, we used a method called the common spatial pattern (CSP) algorithm [90]. The CSP algorithm is designed to find the brain regions that can best distinguish between two categories, based on a weighted scoring system. By applying this method, our goal was to identify the brain regions that play a key role in distinguishing between natural speech and time-reversed speech. These regions are important for speech processing and semantic understanding.

3.7.5 Brain network analysis based on encoding functions

The brain network can be characterized as a community structure. Therefore, community detection is often used in exploring the brain network during a given task [91]. To do so, we first need to define the nodes of the brain and links of the network [92]. In large-scale brain networks, nodes usually represent brain regions, and links represent anatomical, functional, or effective connections [93]. The pre-defined spatial regions of interest (ROIs) assessed by anatomical atlases are one of the most popular methods for defining brain nodes [94]. This study uses the 68 nodes (brain region) that were defined in the DKA, and it uses Pearson correlation to describe the functional link among the nodes [95]. This would result in 2278 ($= C_{68}^{2}$) edges if linking all pairwise nodes for each trial. Differing from the previous studies, the link weights (temporal correlations) here are calculated using the encoding function of each node, but not the source neural signal. As a result, we obtain a preliminary brain network that consists of all of the brain regions and pairwise links with a weighted edge. Subsequently, we apply statistical testing methods to determine the statistical significance of the connections in order to ascertain their validity.

3.8 Analysis method in decoding process

3.8.1 Modeling of speech decoding process

Similar to the speech encoding process, the speech decoding approach can be modeled using a decoder function $r^{-1}(\tau, ROI_n)$, which is the inverse function of $r(\tau, ROI_n)$. The optimal decoder $r^{-1}(\tau, ROI_n)$ is acquired by minimizing the MSE between the original and predicted speech stimuli, and n denotes the number of regions. Thus, the input speech stimulus $s(t)$ can be decoded from the source neural signal $x(t, ROI_n)$ using the decoder function $r^{-1}(\tau, ROI_n)$.

This can be expressed as follows:

$$s(t) = \sum_{n} \sum_{\tau} r^{-1}(\tau, ROI_n) x(t - \tau, ROI_n). \tag{3.17}$$

Accordingly, the optimal decoder $r^{-1}(\tau, ROI_n)$ can be acquired by

$$r^{-1} = \left[X^T X\right]^{-1} X^T S, \tag{3.18}$$

where X is the time-lag series of the cortex response, is defined as

$$\begin{bmatrix} x(0, ROI_n) & 0 & \cdots & 0 \\ x(\Delta t, ROI_n) & x(0, ROI_n) & \cdots & 0 \\ \vdots & x(\Delta t, ROI_n) & \ddots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & x(0, ROI_n) \\ \vdots & \vdots & \cdots & x(\Delta t, ROI_n) \\ x(T, ROI_n) & x(T - \Delta t, ROI_n) & \cdots & x(T - \tau_{max}, ROI_n) \end{bmatrix}, \tag{3.19}$$

where the value τ_{max} represent the range $[0, \tau_{max}]$ of time lags τ. Δt is the sample period. Variable S is a matrix represents the speech stimuli. The r^{-1} is a $\tau_{max} \times F$ matrix, where F is the column number of S, each column represents the univariate mapping from the neural response to speech stimuli. The range for τ is also be set from 0 to 800 ms in the decoding process.

3.8.2 Reconstruct noise-vocoder speech from brain response

The auditory peripheral system, from the cochlea to the primary auditory cortex, plays a crucial role in decomposing speech into time-frequency representations. This process can be computationally modeled as band-pass filtering and TAE extraction [84]. Each band-pass filter in the filterbank

can be seen as generating a TAE with a carrier signal (temporal fine structure). NVS is created by replacing these carriers with band-limited noise. Interestingly, studies have shown that NVS with only a few bands is sufficient for effective sentence recognition [13]. This suggests that humans can perceive linguistic information primarily through the TAEs of the speech signal.

In our study, as illustrated in Figure 3.4, we utilized the predicted TAEs derived from brain cortex responses to reconstruct the NVS. Initially, we extracted the original TAEs using several band-pass filters. Subsequently, these TAEs was used to model the speech decoding process with the brain cortex signals. Based on this model, we predicted the TAEs for the test set from the cortex responses. Finally, the predicted TAE for each channel was used to modulate the amplitude of band-limited noise, generated by filtering white noise at the same boundary frequency. The resulting amplitude-modulated narrow band-limited noises (NBN) were summed to generate the NVS stimulus. Through the reconstruction of the NVS, our objective was to assess the extent of information retrieval from the cortex response, concerning the original speech.

By implementing this approach, we aimed to examine the efficacy of utilizing the cortical response to reconstruct speech and assess the amount of semantic information that can be recovered from the cortex response.

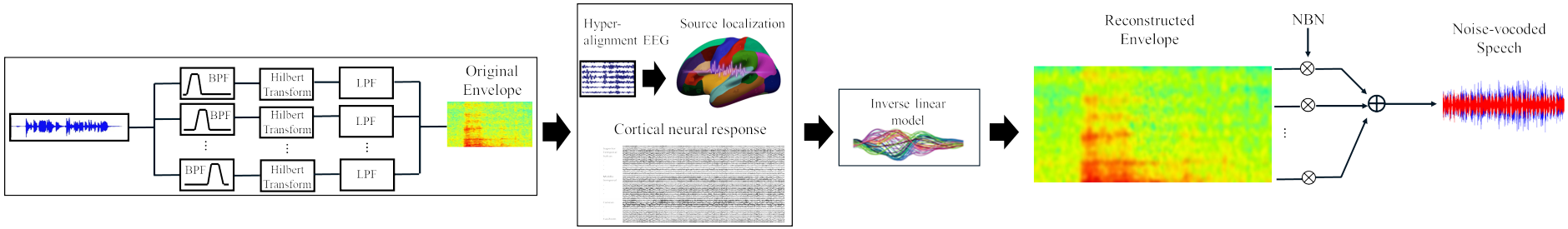

Figure 3.4: TAEs reconstruction and NVS generated procedure.

Chapter 4

Data collection

When individuals perceive sounds, different regions of the brain process auditory features and encode them into various representations. To explore how speech is processed in the brain, we employed a linear time-invariant system to model the encoding process from speech to brain signals. Additionally, we included a control group where the original natural speech was time-reversed, maintaining the acoustic features but reversing the temporal order and removing semantic information. Our hypothesis suggests that brain regions responsible for acoustic features exhibit similar encoding functions for both natural and time-reversed speech, while regions involved in semantics display distinct encoding functions. By comparing the encoding functions of natural and time-reversed speech in various brain regions, we can infer which regions primarily process acoustic features and which regions are responsible for semantic information. Therefore, we designed the following experiment.

4.1 Participants

Twenty-four healthy Mandarin Chinese speakers (mean ± standard deviation age, 22 ± 2.4 years; nine males; right-handed) were recruited from Tianjin University and Tianjin University of Finance and Economics. The experiments were conducted in accordance with the Declaration of Helsinki [96] and were approved by the local ethics committee. The subjects signed informed consent forms before the experiment and were paid for their participation afterward. All the subjects reported no history of hearing impairment or neurological disorders.

4.2 Materials

For our study, we carefully selected three short stories written by Shinichi Hoshi: "The Illusory Princess," "The Grand Plan," and "The Golden Parrot." These stories were chosen for their diverse content and engaging narrative structures, making them ideal for investigating speech comprehension. To ensure consistency and control over the stimuli, the stories were translated into Chinese and recorded by a male Chinese announcer. The recording took place in a soundproof room to minimize any external noise interference.

To maintain appropriate listening durations for each trial, we divided the three stories into 24 non-repetitive segments. Each segment had a duration of approximately 60 seconds. The segmentation process was conducted meticulously while maintaining coherence and continuity within itself. This approach allowed us to present meaningful and manageable units of speech to the participants during the experiment, avoiding excessively long listening periods.

In addition, we included 24 trials where the same story segments were played in time-reverse. All stimuli were mono-speech with a sampling rate of 44.1 kHz, and the stimulus amplitudes were normalized to have the same root mean square (RMS) intensity. The 48 trials, consisting of both forward and time-reverse segments, were randomly presented to the participants. Furthermore, all speech segments were modified to truncate silence gaps to less than 0.5 seconds, ensuring a more streamlined listening experience for the participants [97].

4.3 Experimental procedure

The experiments were carried out in an electronically and magnetically shielded soundproof room. Speech sounds were presented to subjects through Etymotic Research ER-2 insert earphones (Etymotic Research, Elk Grove Village, IL, USA) at a suitable volume (around 65 dB). As shown in Fig. 4.1, during each trial, subjects were instructed to focus on a crosshair mark

in the center of the screen to minimize head movements and other bodily movements. There was a five-second interval between each trial, and the subjects were given a five-minute break every ten trials.

After each story trial, subjects were asked immediately to answer multiple-choice questions about the content of the story to ensure that they focused on the auditory task. For example, during the auditory stimulus, the participants would hear the following passage:

"The king does not have a queen yet. It's time for him to have one. However, this matter must be approached with caution. It would not be good to hastily marry and regret it later. The king must find a beautiful and elegant woman because he is the ruler of a nation. But how should he proceed? With these thoughts in mind, the king summoned a magician—a magician who had long resided in the forest."

Following the presented audio stimulus mentioned above, the screen displayed the following question:

"How does the king plan to find a queen?"

The subjects were then required to choose the correct option from the provided choices:

1. **Launch an attack on another country.**
2. **Stumble upon someone in an ancient forest.**
3. **Seek help from a magician.**
4. **Enlist the assistance of his ministers.**

These multiple-choice questions served as a means to assess the subjects' comprehension of the story content and their engagement in the auditory task of original natural speech.

For the time-reversed speech, we embedded unique tones in some trials to draw more of the subjects' attention to the reversed stimuli. Subjects were requested to detect the tones and indicate how many times they appeared after the trial. The EEG data corresponding to the embedded tones was removed in further analysis.

The accuracy of the answers of these questions were 88.25 $\pm$ 4.62%, indicating that the subjects were attentive and focused on the speech stimuli during the experiment.

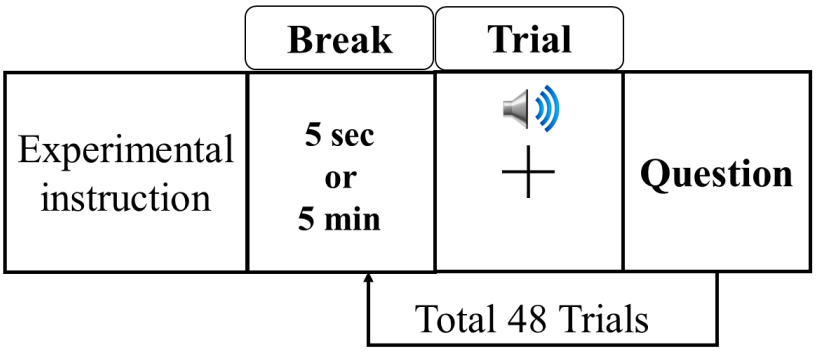

Figure 4.1: Experimental procedure.

4.4 EEG data acquisition and pre-processing

Scalp EEG signals were recorded with a 128-channel Neuroscan SynAmps system (Neuroscan, USA) at a sampling rate of 1000 Hz. Six of the channels were used for recording a vertical electrooculogram (VEOG), a horizontal electrooculogram (HEOG), and two mastoid signals. The impedance of each electrode was kept below 5 kΩ during data acquisition. Three subjects' data were discarded in further analysis because they did not give a proper answer to the multiple-choice questions or the electrodes detached during the EEG data recording. The raw EEG data were pre-processed using the EEGLAB toolbox (https://sccn.ucsd.edu/eeglab/index.php) in MATLAB (MathWorks) [98]. This involved removing sinusoidal (i.e., line) noise and bad channels (i.e., low-frequency drifts, noisy channels, short-time bursts) and repairing the data segments [99, 100]. Then, the EEG data was down-sampled to 128 Hz, 1-Hz high-pass filtering was performed to remove linear drift. Adaptive mixture independent component analysis (AMICA) [101] and ICLabel [102] were used to automatically identify and remove artifact components.

Chapter 5

Investigation on speech encoding process

In this chapter, we primarily focused on addressing the first two questions: identifying the brain regions involved in semantic processing and determining whether we can find semantic representations within these regions. To investigate these questions, we initially estimated the encoding functions of different brain regions using the Linear Time-Invariant (LTI) system. By identifying the optimal brain regions that differentiate between natural and time-reversed speech, we were able to pinpoint the regions implicated in semantic processing. Subsequently, we constructed patterns within these brain regions that resemble the semantic features extracted by BERT, aiming to confirm the presence of semantic representations in the brain.

5.1 Identification of key brain regions for speech processing

5.1.1 Estimated encoding function for natural and time-reversed speech

According to the proposed methods, this study gets the reconstructed cortex signals of EEG. Then, the encoding function (TRF) of each brain area is estimated by the LTI system. To assess the accuracy of the encoding process, a leave-one-out cross-validation procedure was employed. Specifically, out of the 24 trials, 23 trials were used for training the encoding function, while the

remaining trial was used for testing the accuracy of the encoding process. This cross-validation procedure was repeated 24 times, once for each trial, for both natural speech and time-reversed speech conditions.

Figure 5.1 show the examples of encoding functions for superiortemporal sulcus (STS) and middletemporal gyrus (MTG) for natural speech and time-reversed speech. One can see that the patterns of the peaks and troughs for STS (Fig. 5.1A) show a significant difference at time lags between 300 and 450 ms of the encoding function (paired t test, $p = 5.2 \times 10^{-5}$; effect size d = 1). In Fig. 5.1B, the encoding function patterns for MTG show a significant difference between 150 and 450 ms (paired t test, $p = 2.1 \times 10^{-14}$; effect size d = 2).

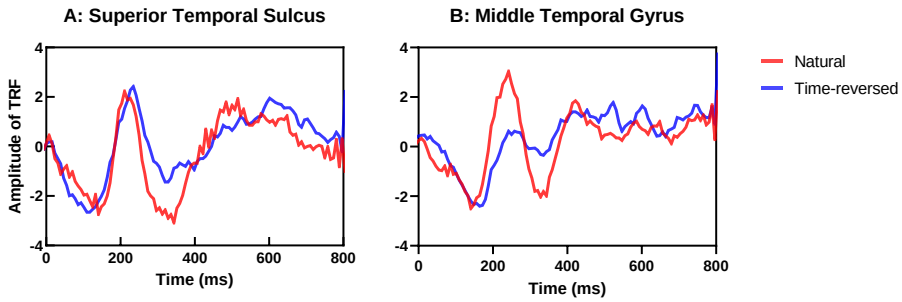

Figure 5.1: Encoding functions for natural and time-reversed speech for STS (A) and MTG (B).

During the experiment, the intelligibility was evaluated on a numerical rating scale from 1 to 5 by the subjects, where "very easy to understand" was scored 5, and "completely incomprehensible" was scored 1. The speech intelligibility was 4.74 ± 0.45 for the normally played natural story but was 1.46 ± 0.81 for the time-reversed one. This means that speech was not comprehended in the time-reversed case since there was little semantic information. For these reasons, functional brain networks are expected to be separated into two clusters. One cluster is for semantically driven brain activation, and the other is for non-semantically driven audio processing. Since time-reversed speech is not understood, there is a lack of a top-down

modulation mechanism to assist in encoding the speech. Therefore, we hypothesize that these differences in the encoding function are attributed to top-down semantic processing.

To validate this expectation, we then examined these encoding functions from a brain functional perspective and constructed functional brain networks based on them. The strength of the connection between two brain regions was quantified using Pearson correlation, ranging from -1 to 1. A higher correlation indicated a greater similarity between the two regions, while a lower correlation indicated less similarity. Then, we employed t-distributed stochastic neighbor embedding (t-SNE) [103] to visualize the brain networks in a two-dimensional representation and determine whether semantically-driven brain activation could be distinguished from non-semantically-driven activation. Initially, we transformed the connection matrix of size 68 $\times$ 68 into a vector with 2278 dimensions, capturing the pairwise connections between all brain nodes for each trial. With a resulting matrix size of 48 $\times$ 2278 for 48 trials, t-SNE analysis was applied. We then utilized the k-means algorithm [32] to cluster the t-SNE results, setting the cluster number to 2 and performing 1000 repetitions with random initial states. Figure 5.2 illustrates a scatter plot of the semantically and non-semantically-driven brain networks in two dimensions. The functional connections displayed distinct clusters for natural and time-reversed speech. We computed the F1-scores between the actual groupings and the k-means clusters for all repetitions, yielding an average F1-score of 0.92 across the 1000 repetitions. These findings indicate that the differences observed in the encoding functions between natural speech and time-reversed speech are indeed influenced by semantic processing.

5.1.2 Key brain regions for natural speech

According to the results of k-means clustering, it was found that the encoding functions of these brain regions can effectively distinguish between natural and time-reversed speech. To investigate which brain regions are more important for this distinction, the importance of these regions was sorted

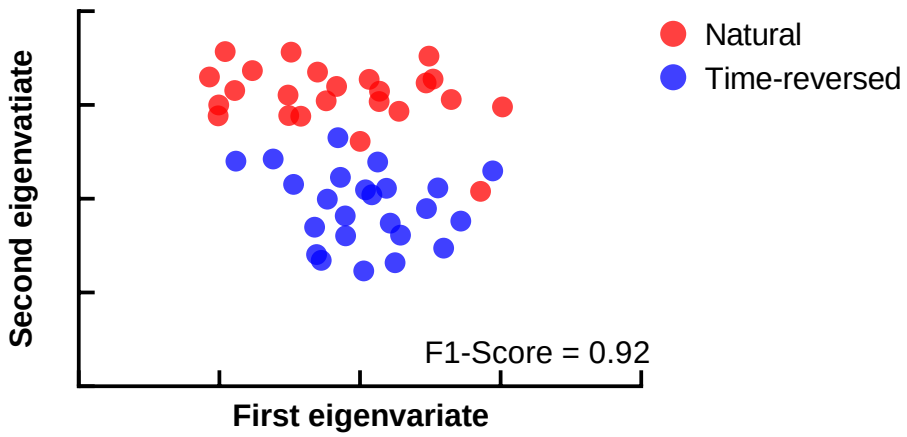

Figure 5.2: K-means clustering of t-SNE embedded distributions for natural and time-reversed speech.

based on the weights of the CSP classification algorithm. Furthermore, using a one-sample t-test ($p < 0.05$), 40 brain regions were selected based on their significant weighted scores, and their spatial locations were plotted. Figure 5.3 displays the weighted score of each brain region. We consider these regions to be important for semantic processing. From the figure, it can be observed that the majority of brain regions involved in distinguishing between natural and time-reversed speech are located in the frontal, temporal, and cingulate cortex. The results are highly consistent with traditional fMRI studies [26, 104].

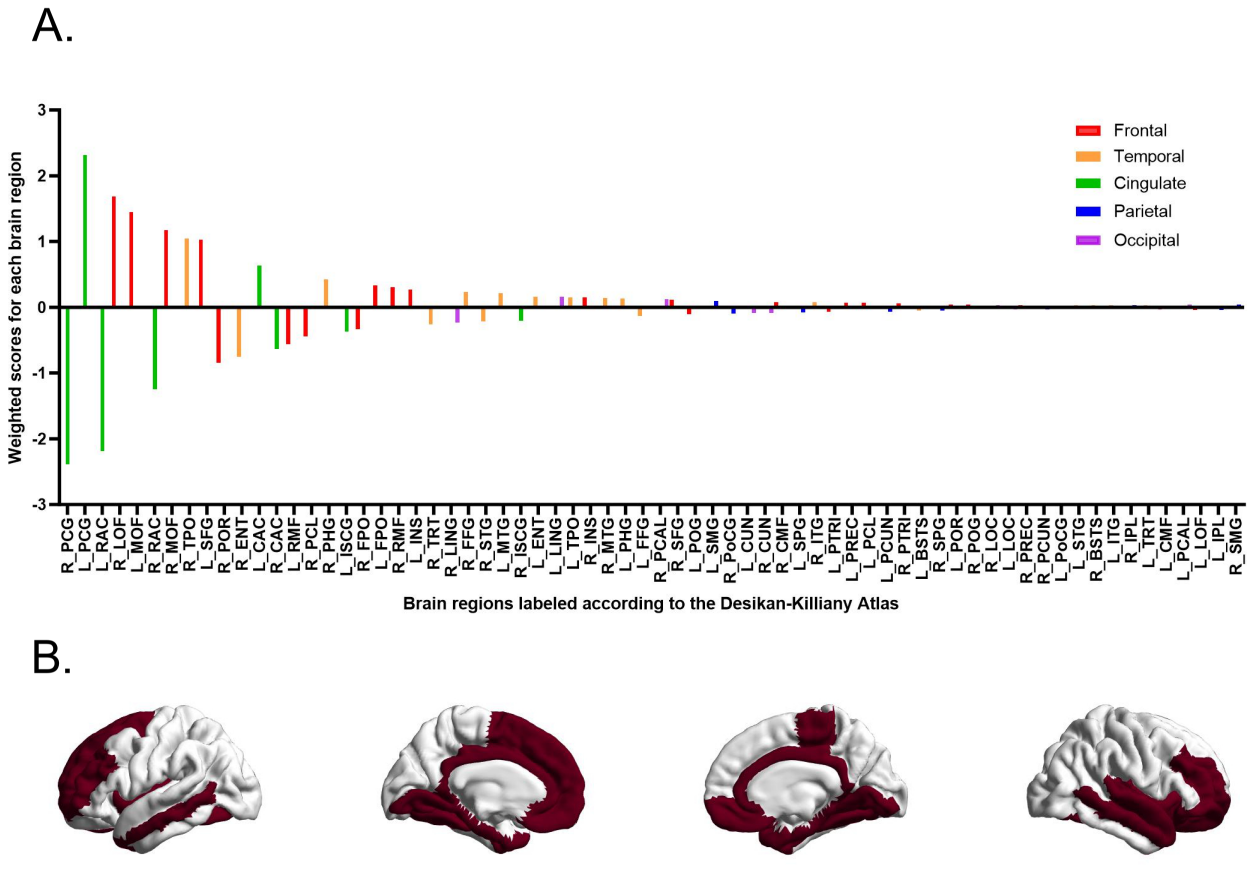

Figure 5.3: Key brain regions for distinguishing between natural and time-reversed speech.

5.2 Semantic representations in brain cortex

Having identified the crucial brain regions involved in semantic processing, our next objective was to investigate whether we could find representations of semantic information within these regions. By inputting the speech text data into the pre-trained Chinese BERT model, we obtained 768-dimensional word vectors that captured the underlying semantic content of the speech. These representations served as a reference for identifying similar patterns in the brain regions associated with semantic processing. Subsequently, we aimed to identify patterns in the brain regions that were similar to the extracted semantic representations. By comparing the patterns of neural activity in these brain regions with the semantic representations extracted from BERT, we sought to uncover whether the brain regions exhibited similar patterns that corresponded to the semantic information encoded in the speech stimuli.

5.2.1 Extracted Semantic information from speech

In our experiment, the dataset comprised 24 natural speech trials, which consisted of three stories, with each story having a complete storyline. Utilizing the pre-trained Chinese BERT model, we extracted the semantic information from these 24 natural speech trials and visualized the corresponding semantic representations. Figure 5.4 illustrates the two-dimensional representation of the semantic information.

Based on the t-SNE results, we observed distinct clusters in the semantic information of word vectors that corresponded to the different stories. We hypothesized that these semantic representations would also be present in the brain and categorized into three classes. Therefore, our next step was to explore and identify similar patterns of representation from brain activity.

5.2.2 Extracted semantic representations from brain

To explore this further, we extracted brain signals from these critical brain regions and examined whether their distribution patterns were similar to the BERT-extracted semantic information. These regions include the left

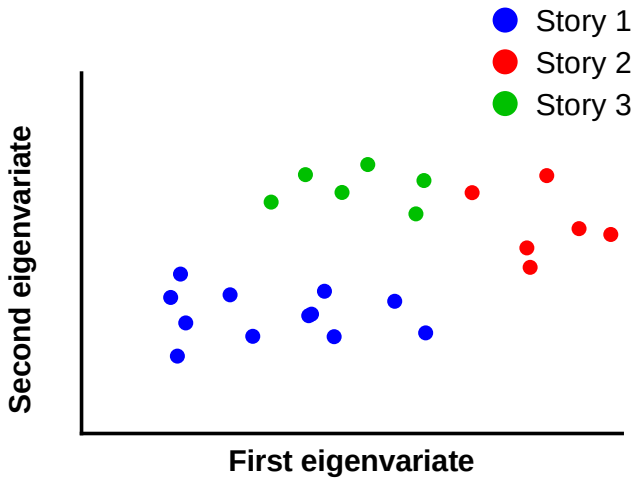

Figure 5.4: T-SNE embedded distributions obtained from semantic information based on BERT.

and right posterior cingulate, which have been implicated in various complex cognitive processes such as memory, navigation, and narrative comprehension [105]. We utilized the CSP analysis to separate the semantic representations of the three stories from the brain signals and transformed them into a two-dimensional representation using t-SNE.

Figure 5.5 displayed the distribution patterns of the brain signals in a 2D space. It was evident that, similar to the BERT-extracted semantic information, the semantic representations in the brain were also categorized into three distinct classes corresponding to the different stories. This finding provides evidence for the existence of semantic representations in these brain regions, further supporting the validity and consistency of the brain activity-based representations.

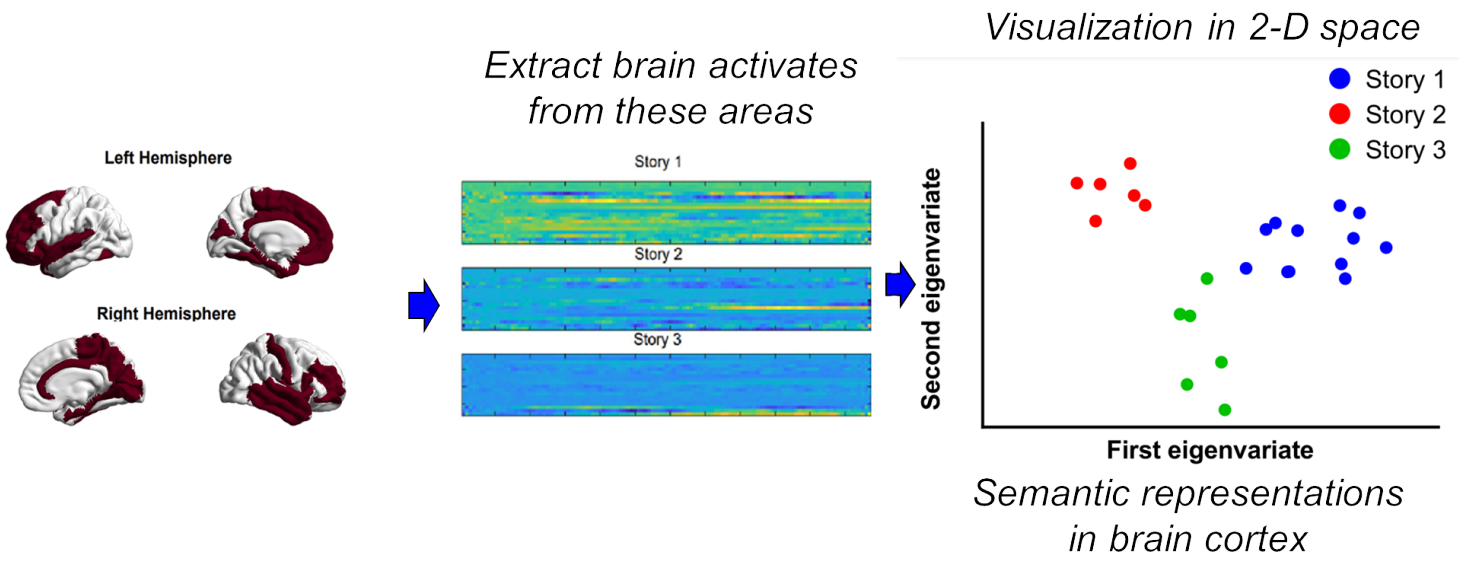

Figure 5.5: Semantic representations in brain cortex.

Since we hypothesized that semantic representations in the brain can be categorized into three classes that correspond to the different stories, we initially conducted a three-classification task based on the brain responses across different EEG frequency ranges in these brain areas. The classification results, presented in Table 5.1, demonstrate the classification accuracy across different frequency bands for these semantic representations. To ensure result stability, the final classification accuracy is based on 100 iterations of random training processes. In each iteration, 50% of the data is used for training, while the remaining 50% is used for testing. To account for the varying number of trials in each story class and maintain data balance, an equal number of test samples from each class are used in each iteration of the evaluation process. This approach helps to mitigate potential biases arising from imbalanced class distributions and provides robust and reliable accuracy estimates. It is evident that brain oscillations in the delta and gamma frequency ranges are closely associated with speech semantic processing. This finding is consistent with previous research reports indicating that brain oscillations in the delta and gamma range synchronize with the lexical aspects of spoken sentences [106–108].

frequency bands	**1-4 Hz**	**4-8 Hz**	**8-12 Hz**	**12-30 Hz**	**30-40 Hz**
Accuracy (%)	83.98(0.11)	67.96(0.10)	69.61(0.09)	58.01(0.10)	70.17(0.09)

[1] The values in parentheses represent the standard deviations of the accuracy.

Table 5.1: Classification accuracy for different semantic categories in the brain.

According to the results, we can conclude that the estimated encoding functions can provide more detailed insights into the brain's distinct states while processing different semantic information. These findings suggest that the brain network dynamically adapts to process different semantic contexts, which is reflected in the encoding functions across various frequency bands.

5.3 Summary

In this chapter, our focus lies in addressing two key questions: (1) which brain regions contribute the most to speech comprehension, and (2) can we find the

semantic representations in these brain areas. To answer these questions, we initially employed the proposed functional hyper-alignment method to map raw EEG electrode data to the source space of the brain through source reconstruction.

In addressing the first question, we utilized common spatial pattern (CSP) analysis to identify the key brain regions that contribute the most to speech comprehension. By differentiating between natural speech and time-reversed speech, we were able to highlight the significant role of these brain regions in the comprehension of speech.

For the second question, we successfully extracted representations of semantic information from brain activity. This allowed us to investigate how speech stimuli are processed within these brain regions and understand the underlying mechanisms involved in speech comprehension. In addition to the classical language processing brain regions located in the temporal and frontal cortex, our results also indicate the significant involvement of the cingulate cortex in semantic processing. Interestingly, based on the weighted score analysis, the activation of the cingulate cortex appears to consistently exhibit desynchronization patterns with the activation of the temporal and frontal cortex. This desynchronization observed in the cingulate cortex may play a critical role in discerning between natural and time-reversed speech, as well as in processing distinct semantic information.

Chapter 6

Decode speech from brain signals

The speech decoding process is the inverse of speech encoding, aiming to reconstruct the original speech TAEs from brain cortex responses. Neural signal-based speech prosthetics aim to provide a natural means of communication for individuals who are unable to listen or speak due to physical or neurological impairments [109]. By decoding speech directly from measured neural activity, it is possible to enable natural conversations and improve the quality of life, particularly for individuals who suffer from neurological diseases.

The exploration of speech decoding serves three primary purposes. Firstly, since we have discovered the representations of semantic information in the brain regions, can we reconstruct the semantic features of the original speech from the brain signals? Secondly, by examining the decoding results, we can validate the accuracy of our reconstructed source signals. Lastly, recent studies have demonstrated the direct recognition or synthesis of speech from intracranial recordings. However, intracranial electrocorticography is invasive and not user-friendly. Therefore, this study aims to investigate the feasibility of reconstructing speech signals from non-invasive EEG, providing a potentially more accessible and user-friendly approach for speech decoding.

6.1 Accuracy for decoding TAEs from brain signals

Can we decode the TAEs from reconstructed source signals of EEG? To evaluate our decoding accuracy, we employed a leave-one-out cross-validation procedure. In the training process, we used 23 trials for training and reserved one trial for testing, allowing us to decode the TAE from neural signals in each fold. In this study, we measured speech decoding accuracy using the Pearson correlation coefficient between the predicted speech TAEs and the original ones. To establish the chance level, we performed a permutation test where we randomly shuffled the order of neural signals 100 times. We then applied the speech decoding process to the permuted data.

Neural oscillations subserve a broad range of functions in speech processing and language comprehension [110]. In previous research, they have shown that oscillatory at different frequency bands tracks the different speech units, such as the delta band for phrase processing and theta band for syllable processing. In this section, we investigate the decoding accuracy of speech TAEs across various frequency bands. The EEG signals were divided into different frequency bands: 1-4 Hz, 4-8 Hz, 8-12 Hz, 12-30 Hz, and 30-40 Hz. Figure 6.1 illustrates the mean decoding accuracy over 24 trials within each of these frequency bands. To ensure a fair comparison, we transformed the correlation coefficient into a z-value using Fisher's z transformation, satisfying a normal distribution [111]. According to an analysis-of-variance (ANOVA) of the z values revealed the speech decoding accuracy was significantly higher than the chance level ($F = 177, p < 0.001$). This finding indicates that the TAEs can be decoded from the reconstructed source signals.

In previous studies, a stable TAE tracking phenomenon has been observed in brain oscillations within the 1-8 Hz range. In this study, our results indicate a widespread presence of neural oscillations coupled with speech TAEs across different frequency bands. Among these bands, the delta frequency range (1-4 Hz) exhibits the most robust coupling, followed by

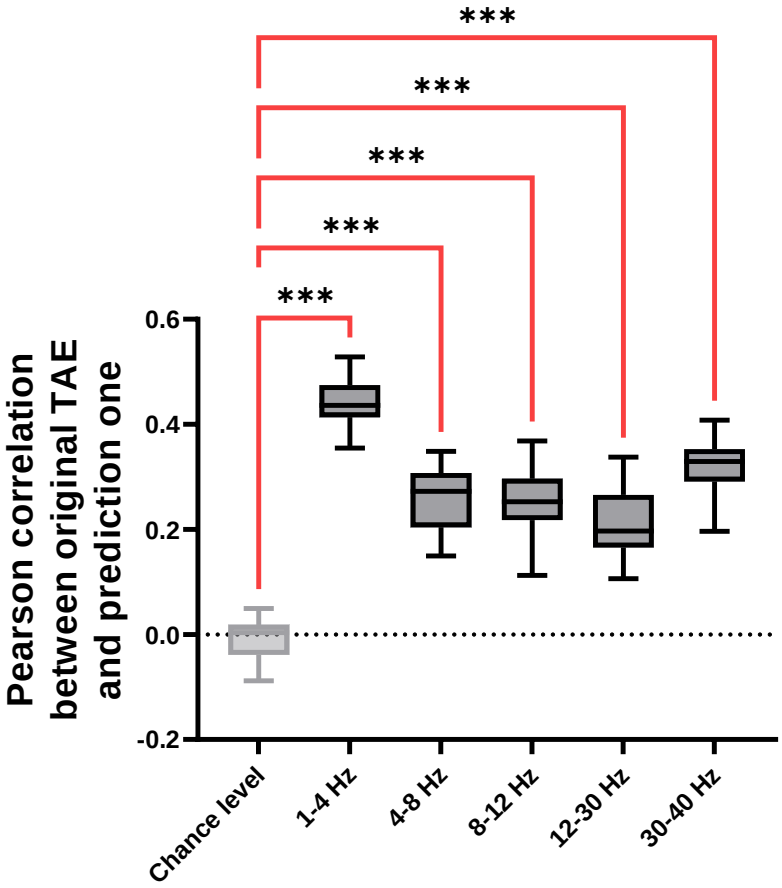

Figure 6.1: TAE decoding accuracies across different frequency ranges of EEG.

the gamma frequency range (30-40 Hz). These findings are consistent with previous results that suggest a stronger relationship between these frequency bands and semantic representation.

6.2 NVS reconstruction results based on LTI model

To assess the amount of speech information that can be obtained from the brain cortex, we employed the NVS to recover speech semantic information from the brain cortex responses. Firstly, we estimated the TAEs of the speech from the brain cortex. The decoding accuracy of the TAEs is depicted in Fig. 6.2A. Subsequently, we utilized the predicted TAEs to reconstruct the speech using NVS. Although our predicted TAEs showed a high correlation with the TAEs of the original speech, it is essential to note that the intelligibility of the NVS in this study was not sufficiently high. This limitation may be attributed to the nonlinearity of the brain, which motivates us to explore the decoding of speech TAEs from brain activity using non-linear convolutional neural networks.

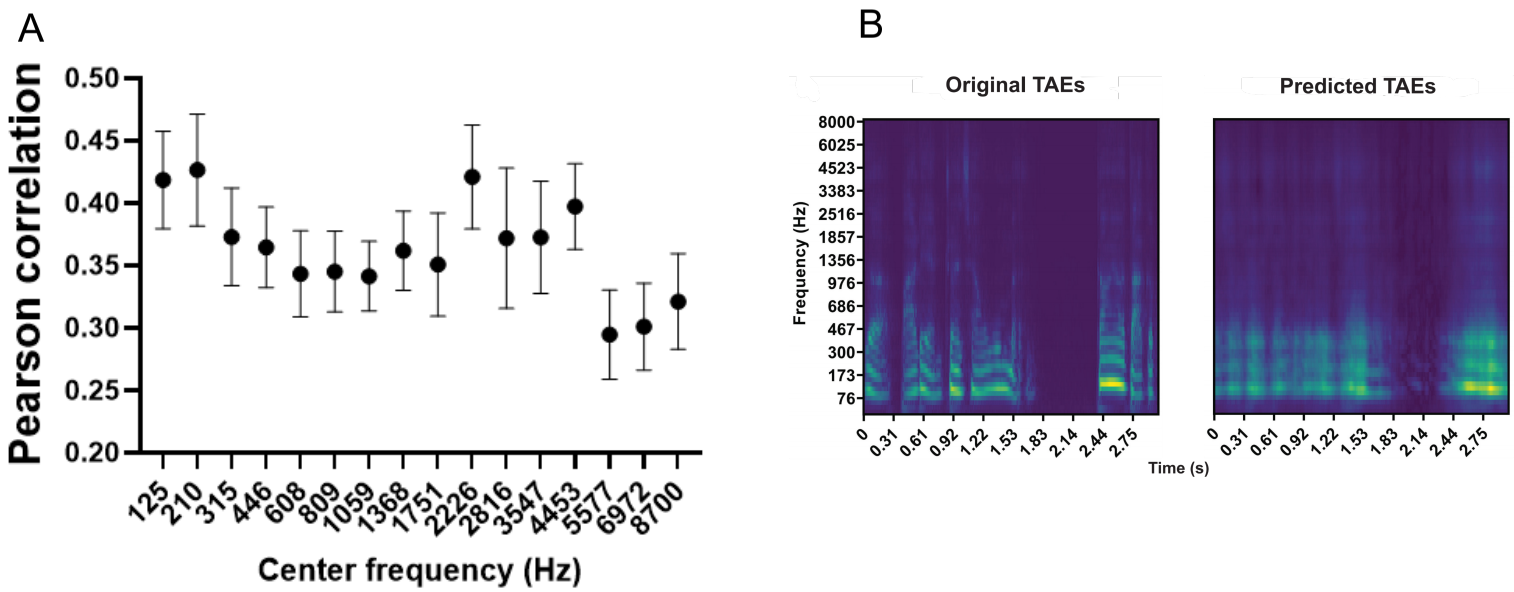

Figure 6.2: Accuracy for decoded TAEs from brain (A) and example of decoded TAEs (B).

6.3 Reconstruct NVS based on convolutional neural networks

In this study, we attempted to reconstruct TAEs using a network architecture based on the Very Large Augmented Auditory Inference (VLAAI) model, which is shown in Fig. 6.3 [112]. The VLAAI network consists of several

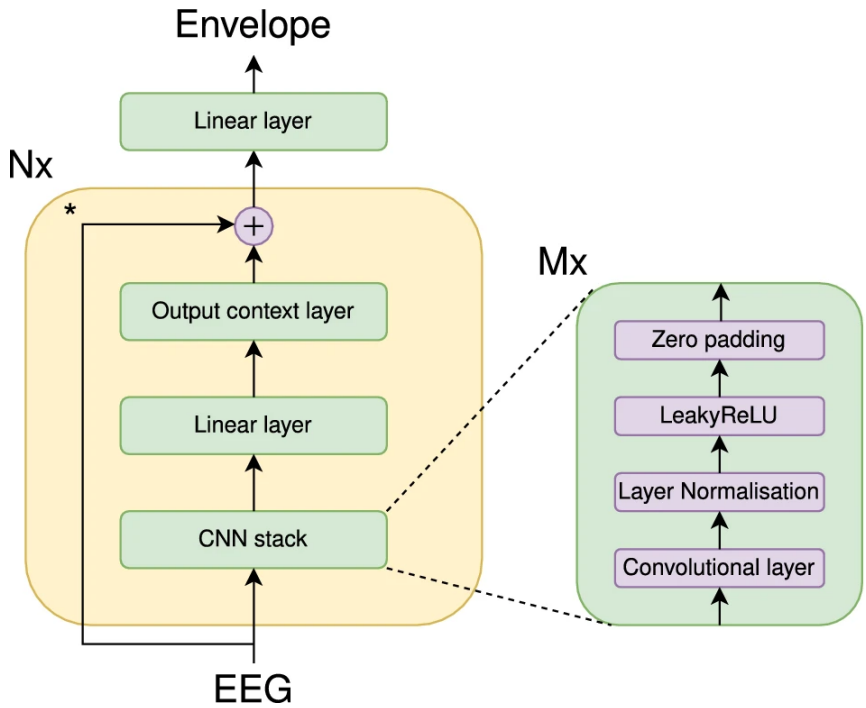

Figure 6.3: Structure of the VLAAI network [112].

blocks, each containing three parts: a convolutional neural network (CNN) stack, a fully connected layer, and an output context layer. The CNN stack consists of M = 5 convolutional layers with varying numbers of filters. Layer normalization, LeakyReLU activation, and zero-padding are applied after each layer. The fully connected layer recombines the output filters of the CNN stack, while the output context layer integrates predictions from previous timesteps to enhance the prediction for the current timestep. This layer utilizes a convolutional operation to transform the previous samples

and the current sample. Skip connections are used in each block except the last, where a linear layer combines the filters of the output context layer into speech TAEs. Similar to the LTI model, we employed a leave-one-out cross-validation procedure to decode the TAEs based on the VLAAI model.

In Fig. 6.4, we compare the decoding accuracy of TAEs using both linear and non-linear VLAAI models. The results of a t-test indicate that there is no significant difference between the two models ($t = 1.556, p = 0.14$).

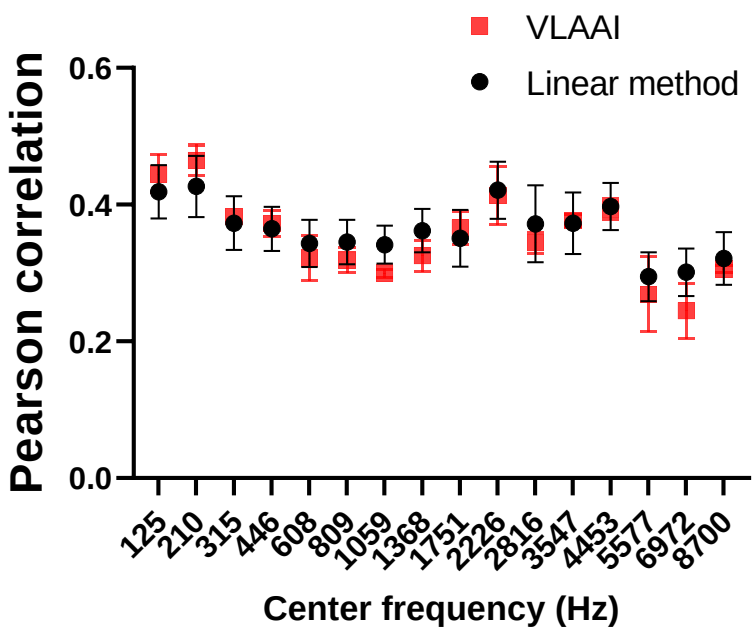

Figure 6.4: Comparison of TAEs decoding accuracy between linear and VLAAI model.

Although we replaced the linear model with a non-linear one, the reconstruction accuracy of the TAE using the non-linear model did not show significant improvement on the test set. Therefore, we shifted our focus to the training set. It's possible that during the training phase, we didn't identify a model that fits the data on the training set well.

In the training process, it is important to note that the non-linear VLAAI model exhibits significantly better fitting results for the original speech TAEs on the training set compared to the linear model. Figure 6.5 displays the

fitting outcomes for both the LTI and VLAAI models, illustrating that the VLAAI model achieves substantially higher levels of fitting accuracy ($t = 30.5, p < 0.0001$). Even though the fitting accuracy of the linear model is not high enough, both the VLAAI and LTI models capture the dynamic cues of TAEs in the training data.

In Fig. 6.6, we show the fitting example of the original speech TAE on the training set using both the linear model and the VLAAI model. It is evident that the non-linear VLAAI model achieves a much closer fit to the original speech TAE compared to the linear model. The non-linear model exhibits a nearly perfect to capture the fine details and dynamics of speech TAE. In contrast, the linear model falls short in accurately capturing the complex patterns of the speech TAE. To ensure that the findings are not a result of overfitting, we reshuffled the EEG segments and their corresponding speech signals for retraining. However, we found that the shuffled EEG data were unable to accurately decode the TAE signals. This indicates that the EEG contains essential information for reconstructing semantic information, but due to the presence of noise and significant inter-individual differences in EEG data, perfect reconstruction of NVS on unseen data is not achievable.

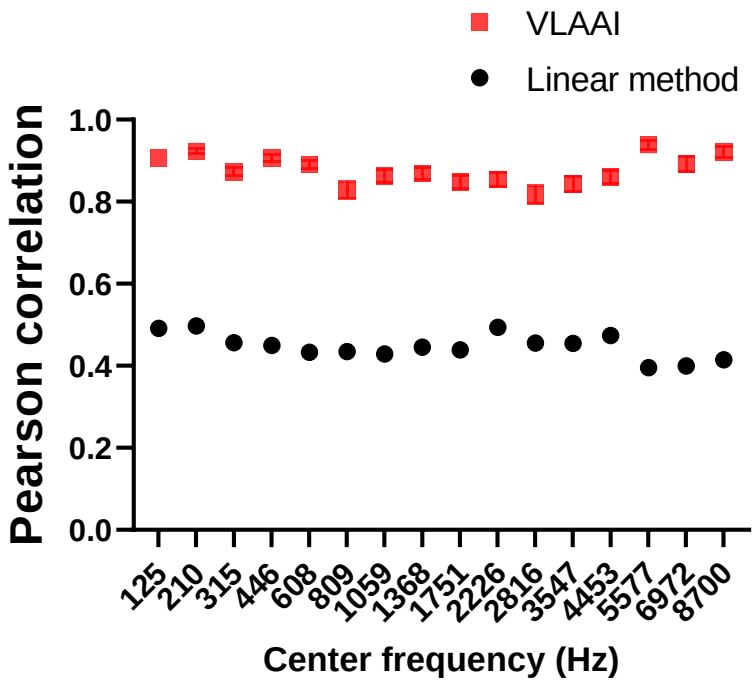

Figure 6.5: Comparison of TAEs decoding accuracy between linear and VLAAI model.

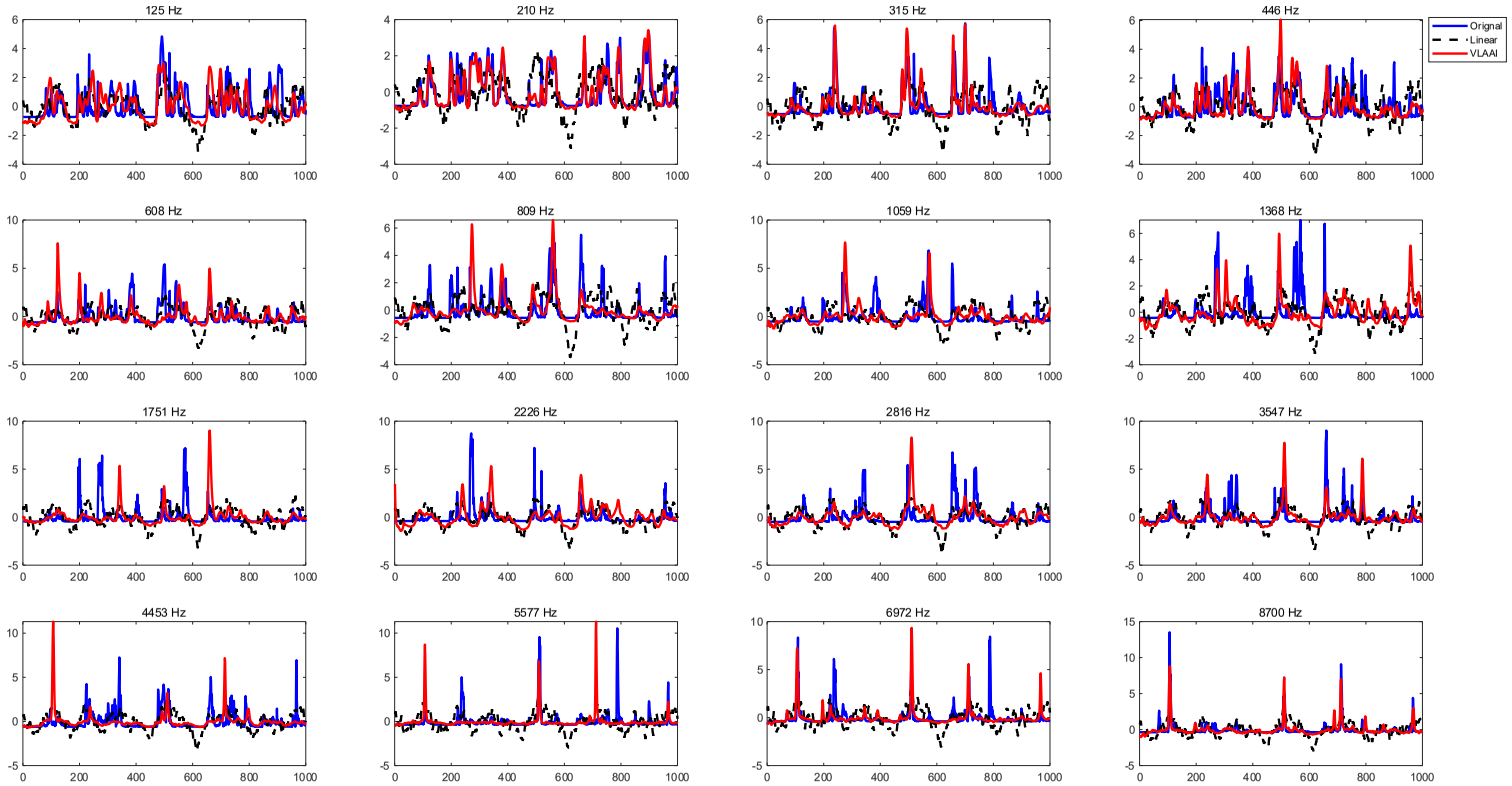

Figure 6.6: Fitting accuracy of TAEs for both the LTI and VLAAI models on the training set.

6.4 Evaluation of intelligibility of reconstructed NVS

To evaluate the intelligibility of the reconstructed NVS using cortical signals, we recruited 10 healthy Mandarin Chinese speakers (mean ± standard deviation age, 28 ± 2.1 years; six males) to assess the intelligibility of the reconstructed NVS. In the NVS perception experiment, four types of NVS were presented to the participants. These included NVS reconstructed using original TAEs, NVS derived from the VLAAI model fitted TAEs, NVS derived from the LTI model fitted TAEs, and NVS from the predicted TAEs in the test set. The reconstructed NVS were randomly presented to the participants. After listening to each speech segment, the participants were asked to rate the intelligibility of the NVS segments. The intelligibility ratings were categorized as very difficult to understand, difficult to understand, moderate, easy to understand, and very easy to understand, with scores ranging from 1 to 5. The final intelligibility scores for different NVS are shown in Table 6.1. Based on the results presented in the table, we observed that while the reconstructed NVS from the test set and the fitted NVS fitting from the linear model were challenging to understand, the fitted NVS from the non-linear VLAAI model exhibited a reasonable level of intelligibility, with a comprehensibility score of approximately 2.9. This finding provides evidence that EEG contains crucial information for reconstructing semantic information.

NVS Type	**Intelligibility Score**
Original TAEs	4.5 (0.72)
VLAAI model fitted TAEs	2.9 (0.87)
LTI model fitted TAEs	1.3 (0.54)
Test set predicted TAEs	1.0 (0.18)

[1] The values in parentheses represent the standard deviations of the accuracy.

Table 6.1: Intelligibility of reconstructed NVS.

6.5 Encoding accuracy of cortex signals from predicted TAEs

To assess the validity of the decoding process results, we utilized the predicted TAEs obtained from the backward decoding process as input to the encoding model. This allowed us to verify whether we could accurately predict the original cortex signals. If we can predict it in with high accuracy, it can show whether our decoding process is reasonable or not.

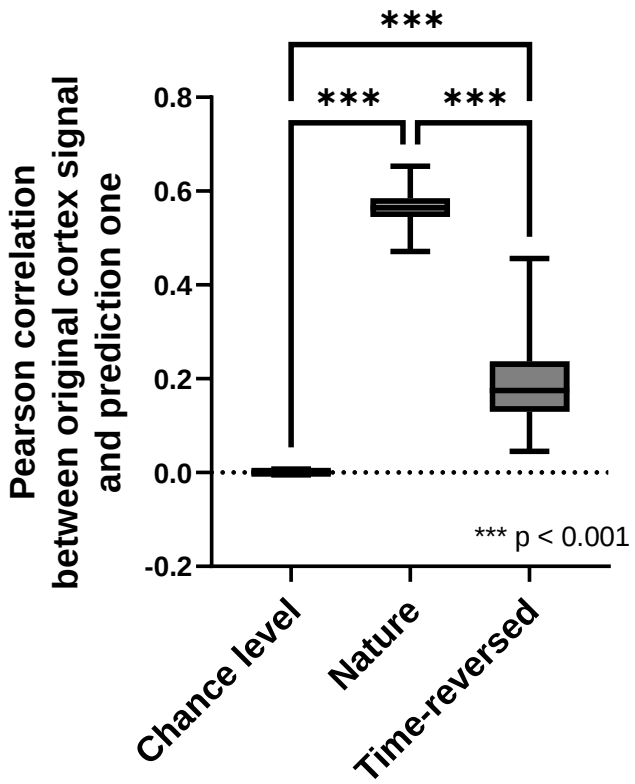

Figure 6.7: Encoding accuracy when predicted TAEs as an input.

Encoding accuracy is defined as the Pearson correlation between the original cortex signals and predicted cortex signals, where the input is the predicted TAEs. The chance level results are obtained through a permutation

test, where the cortex signal and TAEs are not matched. ANOVA analysis reveals that the encoding accuracy of both natural speech and time-reversed speech are significantly higher than the chance level ($F = 113.7, p < 0.001$). Additionally, the encoding accuracy of natural speech is higher than that of time-reversed speech because more speech information is encoded in the cerebral cortex ($t = 219, p < 0.001$).

6.6 Summary

In this chapter, we explored the process of decoding speech from brain signals, aiming to reconstruct the original TAEs from brain cortex responses. Speech decoding has significant implications for neural signal-based speech prosthetics, providing a natural communication channel for individuals with listening impairments.

We assessed the accuracy of decoding speech TAEs from brain signals using the Noise-Vocoded Speech (NVS) technique. The TAEs of the speech were estimated from the brain cortex, and their prediction accuracy was evaluated. We found that the speech TAEs can be decoded from the reconstructed source signals with a significantly higher accuracy than chance level.

Furthermore, we investigated the frequency-specific decoding accuracy of speech TAEs across different frequency bands. Our results revealed a widespread presence of neural oscillations coupled with speech TAEs in various frequency ranges of EEG, with stronger coupling observed in the delta and gamma frequency bands.

To reconstruct NVS from brain cortex responses, we employed the VLAAI network, a non-linear convolutional neural network architecture. The VLAAI model showed superior performance in fitting speech TAEs compared to the linear model in training data. However, it is important to note that perfect reconstruction of NVS on unseen data is challenging due to the presence of noise and inter-individual differences in EEG data.

Overall, our findings highlight the potential of decoding speech from brain signals and provide insights into the neural mechanisms underlying speech

processing. This research contributes to the development of speech prosthetics and offers possibilities for improving communication for individuals with speech-related disabilities.

Chapter 7

General discussion

In Chapters 5 and 6, we conducted separate research on speech encoding and speech decoding. A significant departure from our previous studies was the introduction of spatial information through a hyper-alignment method in the encoding function. In this chapter, we will discuss the benefits and provide some analysis of incorporating spatial information at the cortex level in the encoding/decoding function.

7.1 Scalp level vs. cortex level

In our study, we investigate the benefits of incorporating spatial information into the encoding and decoding functions. By employing the hyper-alignment method, we gain a precise understanding of the encoding and decoding processes at the cortex level.

Figure 7.1 presents the cluster analysis of the encoding function, both at the scalp level and the cortex level, for natural speech and time-reversed speech. As depicted in Figure 7.1, we can observe distinct clustering patterns only when using the cortex level encoding function proposed in our study to differentiate between brain network states associated with natural speech and time-reversed speech. Conversely, these clustering patterns were not evident when examining the scalp level encoding function. Specifically, the left panel of Fig. 7.1 represents the results at the scalp level, while the right panel depicts the results at the cortex level. At the cortex level, clear clusters were observed for natural and time-reversed speech, indicating distinct functional connections. However, at the scalp level, no distinct clusters were observed. To evaluate the performance, F1-scores were

calculated for the actual groupings and the k-means clusters. The average F1-score was 0.5 for the scalp level (Fig. 7.1A) and 0.93 for the cortex level (Fig. 7.1B) across 1000 repetitions. These results highlight the advantages of incorporating spatial information into the encoding and decoding functions, as they capture more robust and discriminative brain network states related to different speech stimuli. It is worth noting that at the scalp level, due to the presence of noise and the mixture of source components, it is challenging to separate natural and time-reversed speech.

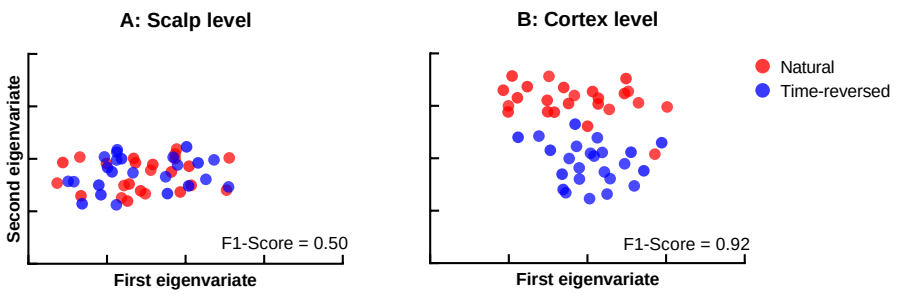

Figure 7.1: K-means clustering of t-SNE embedded distributions obtained at scalp level (A) and cortex level (B).

We also compared the decoding performance between the scalp level and cortex level. Figure 7.2 illustrates the comparisons of decoding accuracies for both levels. To ensure a normal distribution, the correlation coefficient was transformed into a z-value using Fisher's z transformation. Subsequently, a t-test was conducted on the z-values, revealing a significant effect on the decoding accuracies ($t = 21.03, p < 0.0001$). The results of the t-test demonstrate that the decoding accuracy at the cortex level is significantly higher than at the scalp level. This finding indicates that the incorporation of spatial information in the cortex level encoding and decoding functions leads to more reliable and informative representations of the speech information encoded in the brain. Moreover, the decoding accuracy at the cortex level was significantly higher compared to the scalp level, highlighting the enhanced performance and effectiveness of the cortex level decoding approach.

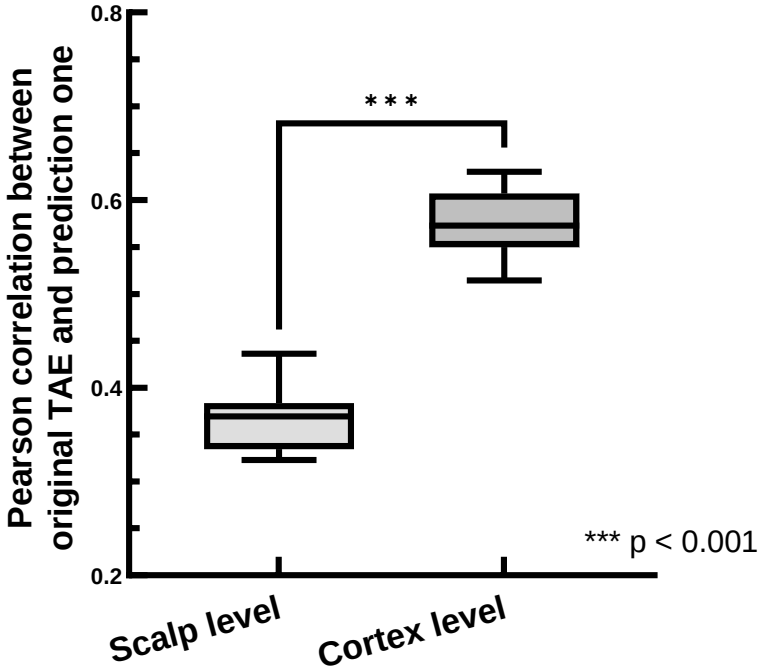

Figure 7.2: Comparison of envelope decoding accuracies between scalp level and cortex level.

7.2 Dynamic brain network analysis during semantic processing

Based on the cortex level encoding function, we constructed a dynamic brain network during semantic processing. MVAR models are capable of analyzing the encoding function in different brain regions and determining causal influences and directed propagation of EEG activity [113]. Consequently, we utilized MVAR models to analyze the encoding function in different brain regions and constructed a dynamic brain network with directional connections. The dynamic brain network is shown in the Fig. 7.3.

Between 0-200 ms, the brain does not form complex networks, and the information flow primarily occurs from the temporal area to the frontal area. This may indicate that the primary auditory areas are transmitting the

collected acoustic information to higher-level brain regions responsible for sound-to-semantic mapping. As time progresses to 200-400 ms, we observe distinct ventral and dorsal pathways in the brain, supporting the traditional dual-stream model observed in fMRI studies. By the time 400-600 ms elapses, we find that the information flow primarily shifts from frontal to temporal areas. This may suggest that higher-level brain regions involved in semantic processing predict upcoming semantic information, aiding the primary auditory cortex in tracking the acoustic information of these speech signals more effectively. When reaching 600-800 ms, a more complex brain network emerges. During this process, there is no clear distinction in the direction of information flow, and the interactions and connections between the temporal and frontal areas become more intricate.

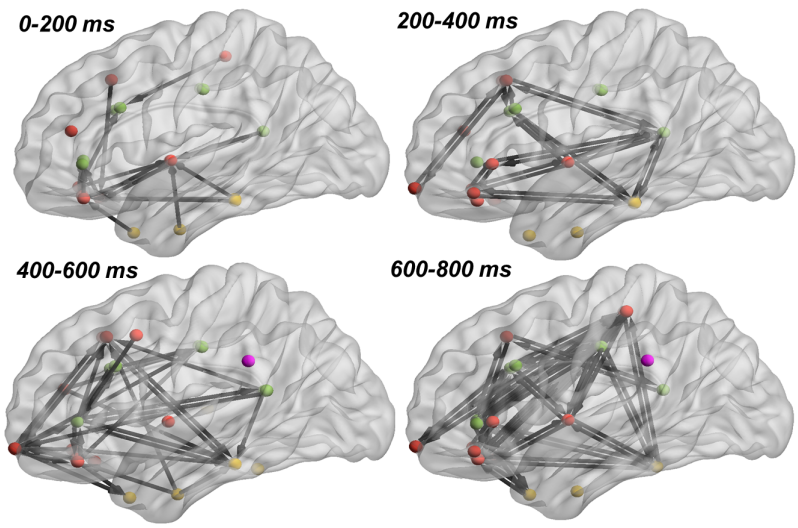

Figure 7.3: Dynamic brain network for semantic processing.

Then, community detection is employed to partition the nodes of the brain network into distinct and non-overlapping subnetworks [114]. Community detection helps us better understand the subnetworks within the brain network that serve different functions. In our study, we identified two main subnetworks, as shown in Fig. 7.4. Subnetwork 1 comprises the primary auditory cortex, inferior frontal cortex, superior temporal cortex, and middle

frontal cortex. According to previous fMRI research, these regions are involved in phrase structure building and lexical selection. Subnetwork 2 primarily consists of Broca area, the middle temporal cortex, fusiform, superior frontal cortex, supramarginal, and postcentral areas. These regions play a key role in sentence-level processing, speech comprehension, and lexical-semantic functions.

Notably, the activity patterns within these two brain subnetworks appear to exhibit desynchronization. In subnetwork 1, the activity during the 100-200 ms and 300-400 ms intervals is positive, whereas in subnetwork 2, the pattern is opposite. Previous studies have also reported similar desynchronization between the temporal and frontal areas [28]. Therefore, this type of desynchronization may be a critical factor in semantic processing. The brain regions of two subnetworks are shown in Table 7.1.

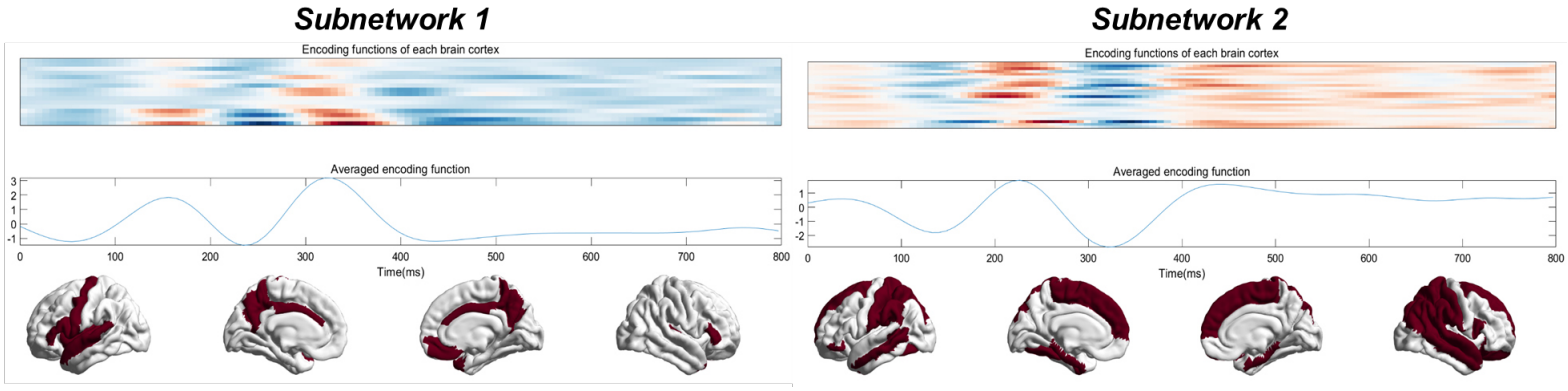

Figure 7.4: Two subnetworks based on community detection.

7.3 Nature speech vs. time-reversed speech

Based on our network analysis, we believe that the desynchronization pattern between the temporal and frontal areas is crucial for the process of language comprehension. Therefore, in natural speech, the activity patterns in the frontal and temporal areas may differ from those in time-reversed speech. This is because time-reversed speech involves minimal language comprehension. Hence, we investigated the frontal and temporal activity patterns for both natural and time-reversed speech. The results are depicted in Figure 7.5.

The analysis reveals that the frontal area and primary auditory cortex coupling was stronger for natural speech compared to time-reversed speech. Specifically, the correlation coefficient was 0.65 for natural speech, while it was 0.23 for time-reversed speech. Notably, the coupling rapidly decreased after 400ms for time-reversed speech. This observation suggests that high-level language areas were not engaged in the auditory processing of time-reversed speech since it was not comprehended by the subjects. Previous studies have also reported the coupling between the auditory cortex and frontal areas, and this coupling tends to increase when speech has higher intelligibility [50].

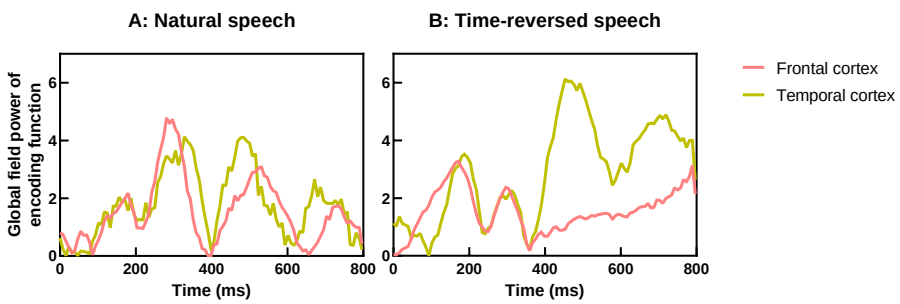

Figure 7.5: Coupling between frontal area (red color) and auditory cortex (blue color) for natural speech (A) and time-reversed speech (B).

Then, semantic representations in the brain cortex are also investigated

in time-reversed speech. To ensure the stability of our results, each time we performed the k-means algorithm, we randomly extracted 20 segments from the brain responses of the distinct stories, with each segment being around 10 seconds. This process was repeated 1,000 times, and the average F1-score of the 1,000 k-means was then computed. For comparison, we also calculated the average F1-score for 1,000 k-means on time-reversed speech. As illustrated in Fig. 7.6A, the results revealed an F1-score of 0.81 for natural speech. This suggests that the brain responses, as anticipated, were grouped into three classes based on the domain of the different stories, affirming the presence of semantic representations in the cortical responses. In contrast, the F1-score was only 0.63 for time-reversed speech in Fig. 7.6B, indicating that since time-reversed speech scarcely contains little linguistic information, we could not distinguish any significant semantic differentiation in its corresponding brain activities.

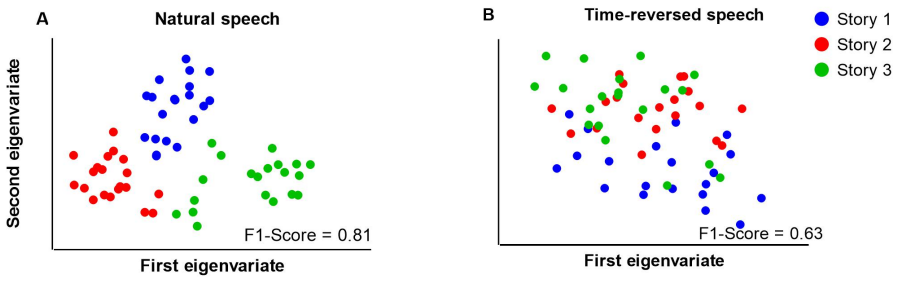

Figure 7.6: K-means clustering of t-SNE embedded distributions for three stories in time-reversed speech.

Next, we compared natural speech and time-reversed speech from a decoding perspective. As natural speech is comprehensible, it encodes more semantic information in brain signals. Consequently, during the decoding process, we can extract and decode more information from cortex responses in natural speech compared to time-reversed speech. To assess this, we compared the accuracy of speech decoding between natural and time-reversed speech.

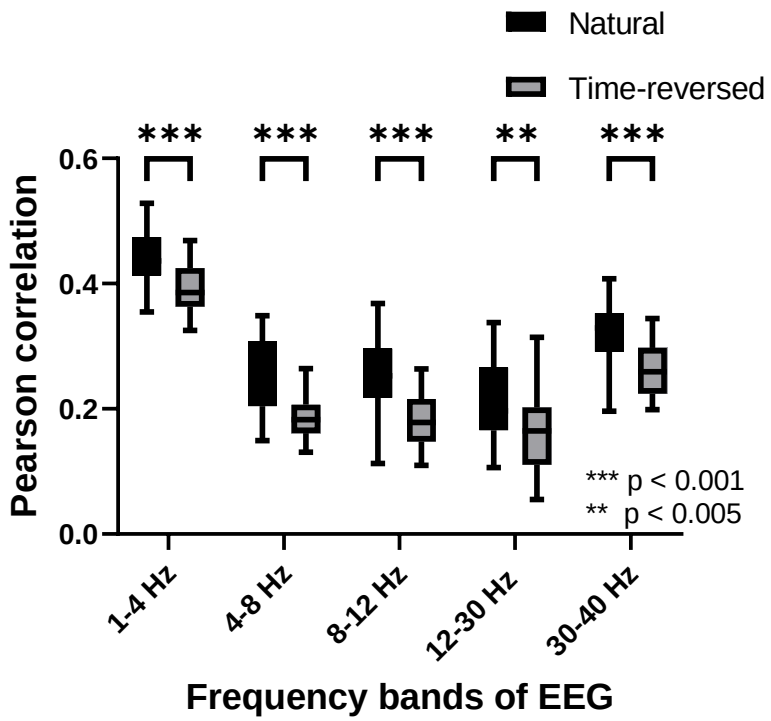

Figure 7.7: Decoding accuracy of TAEs between natural and time-reversed speech.

In Figure 7.7, we can see how accurately the envelope is decoded across different EEG frequency bands. To compare natural speech and time-reversed speech decoding accuracy, Pearson correlation coefficients were transformed into z-values using Fisher's z transformation to ensure a normal distribution. An ANOVA test of the z-values showed significant differences between natural and time-reversed speech ($F = 78.02, p < 0.001$). This indicates that natural speech is decoded more accurately than time-reversed speech, implying that more semantic information is encoded during the encoding stage for natural speech. As a result, during the decoding stage, the TAEs of natural speech can be reconstructed more accurately.

7.4 Summary

In this chapter, we have provided a comprehensive discussion of our research findings and their implications. Here is a summary of the key points covered:

Scalp level vs. cortex level: We introduced the concept of incorporating spatial information using the hyper-alignment method in the speech encoding function. This approach improved the accuracy and reliability of the encoding function by mitigating noise interference. The cortex level encoding function revealed new insights, such as a broader frequency range for brain-speech coupling and enhanced discrimination of brain networks. We also compared the decoding accuracy of speech TAEs between the scalp level and the cortex level, demonstrating improved accuracy by removing noise in the electrode space and achieving more accurate decoding accuracy at the cortex level.

Dynamic brain network for speech comprehension: We constructed a dynamic brain network during semantic processing and investigated the information flow and interactions between different brain regions. The network analysis revealed a shift in information flow from high level brain areas to primary brain areas during the N400 and P600 components associated with language comprehension. This suggests that high-level semantic processing areas facilitate the extraction of acoustic features by the primary auditory cortex through a mechanism of predicting upcoming semantic information.

Overall, we synthesized the research findings from Chapters 5 and 6, highlighting the advantages of estimating encoding/decoding functions at the cortex level, the accuracy and benefits of source reconstruction and the dynamic network dynamics during semantic processing. These findings contribute to our understanding of the neural mechanisms underlying speech encoding and decoding and have implications for future research in the field.

Subnetwork 1	**Subnetwork 2**
caudalanteriorcingulate L	bankssts L
caudalanteriorcingulate R	bankssts R
caudalmiddlefrontal L	entorhinal L
fusiform R	fusiform L
insula R	isthmuscingulate R
lateraloccipital L	lateralorbitofrontal L
medialorbitofrontal R	lateralorbitofrontal R
parsopercularis L	middletemporal L
parstriangularis L	middletemporal R
posteriorcingulate L	paracentral L
posteriorcingulate R	parsopercularis R
precentral L	parsorbitalis L
precuneus L	postcentral L
precuneus R	postcentral R
rostralmiddlefrontal R	rostralanteriorcingulate R
superiortemporal L	superiorfrontal L
temporalpole L	superiorfrontal R
transversetemporal L	superiorparietal L
	superiorparietal R
	superiortemporal R
	supramarginal L
	supramarginal R
	inferiorparietal L
	parsorbitalis R
	parstriangularis R
	transversetemporal R
	entorhinal R
	inferiorparietal R

Table 7.1: Key brain regions for semantic processing.

Chapter 8

Conclusion

8.1 Research conclusion

In this study, we aim to address three questions. Firstly, we investigate the brain regions that play a crucial role in semantic processing. Based on the findings presented in section 5.2, in addition to the well-known language processing regions located in the temporal and frontal cortex, our results indicate a significant involvement of the cingulate cortex in semantic processing. Considering recent fMRI studies [105, 115], we speculate that the activation observed in the cingulate cortex is specifically related to natural speech paradigms. Furthermore, we discovered that a majority of the right hemisphere regions are also activated during natural language comprehension, suggesting a more widespread bilateral brain activity rather than the traditionally emphasized left hemisphere lateralization. These findings shed new light on the neural mechanisms underlying speech comprehension and challenge the conventional understanding of language processing being primarily localized in the left hemisphere. These brain areas are listed in Table 7.1.

The second question focuses on whether we can identify semantic representations within these brain regions. Based on the results presented in section 5.3, we have successfully discovered representational forms that bear similarity to the original semantic information in brain activity. Moreover, we have observed a stronger correlation between neural oscillations in the delta and gamma frequency bands within these brain regions and semantic processing. This suggests that the delta and gamma frequency bands play

a significant role in the neural mechanisms underlying semantic processing. These findings provide further evidence for the involvement of specific neural oscillations in the encoding and processing of semantic information within the identified brain regions. Furthermore, when constructing a dynamic brain network utilizing these regions, we observed an early flow of information from the primary brain areas to the frontal areas, as indicated in section 7.2. However, during later stages such as the N400 phase, there is a noticeable shift in the predominant information flow from high level brain areas to the primary brain areas. This finding may suggest a top-down regulatory role in the language processing process, where higher-level brain regions, once receiving sufficient information, can assist primary brain areas in accelerating information acquisition and integration. These findings provide valuable insights into the mechanisms of semantic processing and the interaction between different brain regions during language comprehension.

The final question explores whether we can decode semantic information from brain signals, as discussed in section 6.3 and 6.4. The results indicate that EEG signals contain crucial information that can be used to reconstruct semantic information. However, it should be noted that the presence of noise in the EEG data, as well as substantial inter-individual differences, make it challenging to achieve perfect reconstruction of semantic information on unseen data. While the EEG signals provide valuable insights into the underlying semantic processes, the decoding accuracy may vary due to individual variations and noise present in the data. Therefore, although the EEG signals carry meaningful information related to semantic processing, it is important to consider the limitations and complexities involved in accurately decoding and reconstructing semantic information from brain signals.

In summary, this study aimed to address three questions. Firstly, it investigated the brain regions crucial for speech comprehension, revealing the involvement of the cingulate cortex in addition to the traditional language processing regions. Secondly, the study successfully identified semantic representations within these brain regions and highlighted the correlation between neural oscillations in the delta and gamma frequency bands and semantic

processing. Lastly, the study explored decoding semantic information from brain signals, emphasizing the challenges posed by noise and inter-individual differences in achieving perfect reconstruction. Overall, these findings provide valuable insights into the neural mechanisms and complexities involved in speech comprehension and semantic processing.

8.2 Research contribution

The research contribution of this study can be summarized as follows:

- Simultaneous investigation of temporal and spatial dimensions: Our research incorporates spatial information by introducing the hyper-alignment method, allowing for the examination of neurological responses from both temporal and spatial dimensions using EEG. Although EEG does not provide the same spatial resolution as fMRI, our research successfully identifies and confirms crucial brain regions involved in speech comprehension. These findings align with previous fMRI studies, enhancing our overall understanding of the roles of different brain regions in the process of understanding spoken language. This advancement in methodology reduces noise interference, enhances accuracy, and provides a more comprehensive view of the brain's language processing activity.
- Our research suggests that the process of comprehending language during natural speech processing involves more brain regions than previously thought. While frontal and temporal cortices have traditionally been considered the main areas responsible for language comprehension, we found that the cingulate cortex also plays an important role. This discovery sheds light on the complex neural network involved in language comprehension. The desynchronization between different subnetworks, primarily within frontal and temporal areas, may be a key mechanism by which the brain processes semantic information.
- Confirmation of the potential for decoding semantic information from the brain: This study demonstrates the potential to decode semantic

information from brain activity by decoding TAEs and combining them with NVS. This significant finding provides a theoretical foundation for future advancements in neuro-interface-based hearing devices. Such devices could utilize non-invasive brain signals to synthesize speech information, thereby enhancing speech perception for individuals with hearing impairments.

In summary, this study contributes to our understanding of speech processing by integrating spatial information, exploring the involvement of brain networks in language comprehension, and demonstrating the potential for decoding semantic information from the brain. These findings have implications for improving speech decoding techniques and advancing the development of neuro-interface-based assistive devices.

8.3 Limitations and Future Directions

While our study has yielded several notable findings, it is important to acknowledge its limitations. Firstly, the brain is not a strictly linear system, and our use of linear models to infer the encoding/decoding function may not fully capture the underlying neural mechanisms of language processing. Future studies should consider more appropriate non-linear models to to represent these processes better.

Secondly, our study primarily focused on decoding TAEs of the speech signal. However, the speech signal also contains additional complex features, such as phonemes and syllables, which were not included in the current study. Future studies should aim to apply our method to decode these additional speech features.

Thirdly, the number of participants in our dataset remains relatively limited. Future studies with larger sample sizes, encompassing other languages such as Japanese and English, would yield more definitive outcomes and further affirm the reliability of our methodology.

Lastly, while our use of cortex level signals allowed us to identify the brain regions most involved in speech comprehension, further research is

needed to gain a deeper understanding of the precise mechanisms underlying these effects. For example, the coupling of brain activity with the speech signal across different frequency bands suggests the existence of a multi-scale temporal coding mechanism. Future studies could focus on elucidating these mechanisms and their functional significance.

8.4 Summary

In this study, we successfully inferred a linear encoding model at the cortex level by simultaneously considering both temporal and spatial information, using EEG. This approach overcomes the limitations of traditional research that cannot capture both spatial and temporal aspects using non-invasive neuroimaging techniques. Based on our inferred cortex level encoding function, our findings challenge traditional language comprehension models by suggesting that a wide range of brain regions, including the right hemisphere, are involved in the processing of natural speech. This cannot be explained solely by language models restricted to specific regions in the left hemisphere. We also identified two distinct and desynchronized subnetworks in the brain, which may serve as the neural basis for language comprehension. This discovery suggests the need to investigate and establish new language comprehension models from the perspective of network collaboration in future research.

Furthermore, through the decoding process, we demonstrated the possibility of reconstructing semantic information from brain activity. This finding provides a theoretical foundation and has the potential to inspire future studies in auditory neuroscience and contribute to the development of more effective speech-brain interfaces.

References

[1] G. Zhang, Y. Si, and J. Dang, "Revealing the dynamic brain connectivity from perception of speech sound to semantic processing by eeg," *Neuroscience*, vol. 415, pp. 70–76, 2019.

[2] C. J. Price, "A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading," *Neuroimage*, vol. 62, no. 2, pp. 816–847, 2012.

[3] N. Joodi and F. Rahmani, "Application of functional magnetic resonance imaging in neurolinguistics: A systematic review," *Frontiers in Biomedical Technologies*, vol. 6, no. 4, pp. 204–216, 2019.

[4] A. M. Beres, "Time is of the essence: A review of electroencephalography (eeg) and event-related brain potentials (erps) in language research," *Applied psychophysiology and biofeedback*, vol. 42, pp. 247–255, 2017.

[5] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant, "Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies," *Cerebral cortex*, vol. 19, pp. 2767–2796, 2009.

[6] W.-Y. Chow and C. Phillips, "No semantic illusions in the "semantic p600" phenomenon: Erp evidence from mandarin chinese," *Brain research*, vol. 1506, pp. 76–93, 2013.

[7] M. Kutas and K. D. Federmeier, "Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp)," *Annual review of psychology*, vol. 62, pp. 621–647, 2011.

[8] J. Brennan, "Naturalistic sentence comprehension in the brain," *Language and Linguistics Compass*, vol. 10, no. 7, pp. 299–313, 2016.

[9] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[10] L. S. Hamilton and A. G. Huth, "The revolution will not be controlled: natural stimuli in speech neuroscience," *Language, cognition and neuroscience*, vol. 35, no. 5, pp. 573–582, 2020.

[11] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, "Correlation-based channel selection and regularized feature optimization for mi-based bci," *Neural Networks*, vol. 118, pp. 262–270, 2019.

[12] L. Faes, G. Nollo *et al.*, "Multivariate frequency domain analysis of causal interactions in physiological time series," *Biomedical Engineering, Trends in Electronics, Communications and Software*, vol. 8, pp. 403–428, 2011.

[13] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[14] J.-F. Demonet, F. Chollet, S. Ramsay, D. Cardebat, J.-L. Nespoulous, R. Wise, A. Rascol, and R. Frackowiak, "The anatomy of phonological and semantic processing in normal subjects," *Brain*, vol. 115, no. 6, pp. 1753–1768, 1992.

[15] J.-F. Démonet, C. Price, R. Wise, and R. Frackowiak, "Differential activation of right and left posterior sylvian regions by semantic and phonological tasks: a positron-emission tomography study in normal human subjects," *Neuroscience letters*, vol. 182, no. 1, pp. 25–28, 1994.

[16] R. Vandenberghe, C. Price, R. Wise, O. Josephs, and R. S. Frackowiak, "Functional anatomy of a common semantic system for words and pictures," *Nature*, vol. 383, no. 6597, pp. 254–256, 1996.

[17] R. R. Benson, D. H. Whalen, M. Richardson, B. Swainson, V. P. Clark, S. Lai, and A. M. Liberman, "Parametrically dissociating speech and

nonspeech perception in the brain using fmri," *Brain and language*, vol. 78, no. 3, pp. 364–396, 2001.

[18] S. D. Newman and D. Twieg, "Differences in auditory processing of words and pseudowords: An fmri study," *Human brain mapping*, vol. 14, no. 1, pp. 39–47, 2001.

[19] A. Vouloumanos, K. A. Kiehl, J. F. Werker, and P. F. Liddle, "Detection of sounds in the auditory stream: event-related fmri evidence for differential activation to speech and nonspeech," *Journal of Cognitive Neuroscience*, vol. 13, no. 7, pp. 994–1005, 2001.

[20] A. J. Newman, R. Pancheva, K. Ozawa, H. J. Neville, and M. T. Ullman, "An event-related fmri study of syntactic and semantic violations," *Journal of psycholinguistic research*, vol. 30, pp. 339–364, 2001.

[21] S. K. Scott, C. C. Blank, S. Rosen, and R. J. Wise, "Identification of a pathway for intelligible speech in the left temporal lobe," *Brain*, vol. 123, no. 12, pp. 2400–2406, 2000.

[22] S. Bookheimer, "Functional mri of language: new approaches to understanding the cortical organization of semantic processing," *Annual review of neuroscience*, vol. 25, no. 1, pp. 151–188, 2002.

[23] G. Hickok and D. Poeppel, "Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language," *Cognition*, vol. 92, no. 1-2, pp. 67–99, 2004.

[24] ——, "The cortical organization of speech processing," *Nature reviews neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[25] C. Whitney, W. Huber, J. Klann, S. Weis, S. Krach, and T. Kircher, "Neural correlates of narrative shifts during auditory story comprehension," *Neuroimage*, vol. 47, no. 1, pp. 360–366, 2009.

[26] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev *et al.*, "The "narratives" fmri dataset for evaluating models of naturalistic language comprehension," *Scientific data*, vol. 8, no. 1, p. 250, 2021.

[27] C. Caucheteux, A. Gramfort, and J.-R. King, "Evidence of a predictive coding hierarchy in the human brain listening to speech," *Nature human behaviour*, vol. 7, no. 3, pp. 430–441, 2023.

[28] S. Dikker, M. F. Assaneo, L. Gwilliams, L. Wang, and A. Kösem, "MEG and Language," Aug. 2019, working paper or preprint.

[29] M. Teplan *et al.*, "Fundamentals of eeg measurement," *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002.

[30] D. Hernández, A. Puupponen, and T. Jantunen, "The contribution of event-related potentials to the understanding of sign language processing and production in the brain: Experimental evidence and future directions," *Frontiers in Communication*, p. 40, 2022.

[31] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[32] G. M. Di Liberto, J. A. O'sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[33] L. Gwilliams, T. Linzen, D. Poeppel, and A. Marantz, "In spoken word recognition, the future predicts the past," *Journal of Neuroscience*, vol. 38, no. 35, pp. 7585–7599, 2018.

[34] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp)," *Annual review of psychology*, vol. 62, p. 621, 2011.

[35] L. Osterhout and P. J. Holcomb, "Event-related brain potentials elicited by syntactic anomaly," *Journal of memory and language*, vol. 31, no. 6, pp. 785–806, 1992.

[36] A. Kielar, L. Panamsky, K. A. Links, and J. A. Meltzer, "Localization of electrophysiological responses to semantic and syntactic anomalies in language comprehension with meg," *NeuroImage*, vol. 105, pp. 507–524, 2015.

[37] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[38] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[39] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 2, pp. 181–191, 2018.

[40] W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen, "The hierarchical cortical organization of human speech processing," *Journal of Neuroscience*, vol. 37, no. 27, pp. 6539–6557, 2017.

[41] P. M. Alday, "M/eeg analysis of naturalistic stories: a review from speech to language processing," *Language, Cognition and Neuroscience*, vol. 34, no. 4, pp. 457–473, 2019.

[42] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[43] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, 2007.

[44] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[45] H. Weissbart, K. D. Kandylaki, and T. Reichenbach, "Cortical tracking of surprisal during continuous speech comprehension," *Journal of cognitive neuroscience*, vol. 32, no. 1, pp. 155–166, 2020.

[46] C. Brodbeck, S. Bhattasali, A. A. C. Heredia, P. Resnik, J. Z. Simon, and E. Lau, "Parallel processing in speech perception with local and global representations of linguistic context," *Elife*, vol. 11, p. e72056, 2022.

[47] M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, and F. P. De Lange, "A hierarchy of linguistic predictions during natural language comprehension," *Proceedings of the National Academy of Sciences*, vol. 119, no. 32, p. e2201968119, 2022.

[48] M. F. Howard and D. Poeppel, "Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension," *Journal of neurophysiology*, vol. 104, no. 5, pp. 2500–2511, 2010.

[49] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, pp. 511–517, 2012.

[50] H. Park, R. A. A. Ince, P. G. Schyns, G. Thut, and J. Gross, "Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners," *Current Biology*, vol. 25, no. 12, pp. 1649–1653, 2015.

[51] L. Wang, P. Hagoort, and O. Jensen, "Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations," *Journal of cognitive neuroscience*, vol. 30, no. 3, pp. 432–447, 2018.

[52] J.-M. Schoffelen, A. Hultén, A. F. Marquand, J. Uddén, and P. Hagoort, "Frequency-specific directed interactions in the human brain network for language," *Proceedings of the National Academy of Sciences*, vol. 114, no. 30, pp. 8083–8088, 2017.

[53] F. Mamashli, S. Khan, J. Obleser, A. D. Friederici, and B. Maess, "Oscillatory dynamics of cortical functional connections in semantic prediction," *Human Brain Mapping*, vol. 40, no. 6, pp. 1856–1866, 2019.

[54] M. X. Cohen, *Analyzing neural time series data: theory and practice.* MIT press, 2014.

[55] S. Homölle and R. Oostenveld, "Using a structured-light 3d scanner to improve eeg source modeling with more accurate electrode positions," *Journal of Neuroscience Methods*, vol. 326, p. 108378, 2019.

[56] G. A. Taberna, M. Marino, M. Ganzetti, and D. Mantini, "Spatial localization of eeg electrodes using 3d scanning," *Journal of neural engineering*, vol. 16, no. 2, p. 026020, 2019.

[57] H. Becker, L. Albera, P. Comon, R. Gribonval, F. Wendling, and I. Merlet, "Brain-source imaging: From sparse to tensor models," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 100–112, 2015.

[58] R. D. Pascual-Marqui *et al.*, "Standardized low-resolution brain electromagnetic tomography (sloreta): technical details," *Methods Find Exp Clin Pharmacol*, vol. 24, no. Suppl D, pp. 5–12, 2002.

[59] L. Koessler, C. Benar, L. Maillard, J.-M. Badier, J. P. Vignal, F. Bartolomei, P. Chauvel, and M. Gavaret, "Source localization of ictal epileptic activity investigated by high resolution eeg and validated by seeg," *Neuroimage*, vol. 51, no. 2, pp. 642–653, 2010.

[60] P. E. Coutin-Churchman, J. Y. Wu, L. L. Chen, K. Shattuck, S. Dewar, and M. R. Nuwer, "Quantification and localization of eeg interictal spike activity in patients with surgically removed epileptogenic foci," *Clinical neurophysiology*, vol. 123, no. 3, pp. 471–485, 2012.

[61] A. Sohrabpour, Y. Lu, P. Kankirawatana, J. Blount, H. Kim, and B. He, "Effect of eeg electrode number on epileptic source localization in pediatric patients," *Clinical Neurophysiology*, vol. 126, no. 3, pp. 472–480, 2015.

[62] T.-H. Eom, J.-H. Shin, Y.-H. Kim, S.-Y. Chung, I.-G. Lee, and J.-M. Kim, "Distributed source localization of interictal spikes in benign childhood epilepsy with centrotemporal spikes: a standardized low-resolution brain electromagnetic tomography (sloreta) study," *Journal of Clinical Neuroscience*, vol. 38, pp. 49–54, 2017.

[63] A. Bluschke, J. Schuster, V. Roessner, and C. Beste, "Neurophysiological mechanisms of interval timing dissociate inattentive and combined adhd subtypes," *Scientific Reports*, vol. 8, no. 1, pp. 1–10, 2018.

[64] W. X. Chmielewski, A. Tiedt, A. Bluschke, G. Dippel, V. Roessner, and C. Beste, "Effects of multisensory stimuli on inhibitory control in adolescent adhd: It is the content of information that matters," *NeuroImage: Clinical*, vol. 19, pp. 527–537, 2018.

[65] S. Asadzadeh, T. Y. Rezaii, S. Beheshti, A. Delpak, and S. Meshgini, "A systematic review of eeg source localization techniques and their applications on diagnosis of brain abnormalities," *Journal of neuroscience methods*, vol. 339, p. 108740, 2020.

[66] O. Etard and T. Reichenbach, "Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise," *Journal of Neuroscience*, vol. 39, no. 29, pp. 5750–5759, 2019.

[67] S. Klamer, A. Elshahabi, H. Lerche, C. Braun, M. Erb, K. Scheffler, and N. K. Focke, "Differences between meg and high-density eeg source

localizations using a distributed source model in comparison to fmri," *Brain topography*, vol. 28, no. 1, pp. 87–94, 2015.

[68] M. Stropahl, A.-K. R. Bauer, S. Debener, and M. G. Bleichner, "Source-modeling auditory processes of eeg data using eeglab and brainstorm," *Frontiers in neuroscience*, vol. 12, p. 309, 2018.

[69] N. Janssen, M. v. d. Meij, P. J. López-Pérez, and H. A. Barber, "Exploring the temporal dynamics of speech production with eeg and group ica," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[70] T. C. Handy, *Event-related potentials: A methods handbook*. MIT press, 2005.

[71] R. Grech, T. Cassar, J. Muscat, K. P. Camilleri, S. G. Fabri, M. Zervakis, P. Xanthopoulos, V. Sakkalis, and B. Vanrumste, "Review on solving the inverse problem in eeg source analysis," *Journal of neuroengineering and rehabilitation*, vol. 5, no. 1, pp. 1–33, 2008.

[72] E. Pirondini, B. Babadi, G. Obregon-Henao, C. Lamus, W. Q. Malik, M. S. Hämäläinen, and P. L. Purdon, "Computationally efficient algorithms for sparse, dynamic solutions to the eeg source localization problem," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1359–1372, 2017.

[73] T. Mannepalli and A. Routray, "Certainty-based reduced sparse solution for dense array eeg source localization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 2, pp. 172–178, 2018.

[74] K. Liu, Z. L. Yu, W. Wu, Z. Gu, J. Zhang, L. Cen, S. Nagarajan, and Y. Li, "Bayesian electromagnetic spatio-temporal imaging of extended sources based on matrix factorization," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2457–2469, 2019.

[75] H. Tanaka, "Group task-related component analysis (gtrca): a multivariate method for inter-trial reproducibility and inter-subject similar-

ity maximization for eeg data analysis," *Scientific reports*, vol. 10, pp. 1–17, 2020.

[76] A. de Cheveigné, G. M. D. Liberto, D. Arzounian, D. D. E. Wong, J. Hjortkjær, S. Fuglsang, and L. C. Parra, "Multiway canonical correlation analysis of brain data," *NeuroImage*, vol. 186, pp. 728–740, 2019.

[77] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[78] F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy, "Brainstorm: a user-friendly application for meg/eeg analysis," *Computational intelligence and neuroscience*, vol. 2011, 2011.

[79] J. Vorwerk, C. Engwer, S. Pursiainen, and C. H. Wolters, "A mixed finite element method to solve the eeg forward problem," *IEEE transactions on medical imaging*, vol. 36, no. 4, pp. 930–941, 2016.

[80] S. Schrader, A. Westhoff, M. C. Piastra, T. Miinalainen, S. Pursiainen, J. Vorwerk, H. Brinck, C. H. Wolters, and C. Engwer, "Duneuro—a software toolbox for forward modeling in bioelectromagnetism," *PloS one*, vol. 16, no. 6, p. e0252431, 2021.

[81] A. Gramfort, T. Papadopoulo, E. Olivi, and M. Clerc, "Openmeeg: opensource software for quasistatic bioelectromagnetics," *Biomedical engineering online*, vol. 9, no. 1, pp. 1–20, 2010.

[82] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman *et al.*, "An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest," *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.

[83] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[84] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 234–242, 2018.

[85] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "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*, vol. 25, no. 5, pp. 402–412, 2016.

[86] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral-temporal representation," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[87] Z. Peng, J. Dang, M. Unoki, and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech," *Neural Networks*, vol. 140, pp. 261–273, 2021.

[88] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.

[89] E. De Boer and P. Kuyper, "Triggered correlation," *IEEE Transactions on Biomedical Engineering*, no. 3, pp. 169–179, 1968.

[90] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE transactions on rehabilitation engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[91] D. Jin, R. Li, and J. Xu, "Multiscale community detection in functional brain networks constructed using dynamic time warping," *IEEE Trans-*

actions on Neural Systems and Rehabilitation Engineering, vol. 28, pp. 52–61, 2019.

[92] Q. Yu, Y. Du, J. Chen, J. Sui, T. Adalē, G. D. Pearlson, and V. D. Calhoun, "Application of graph theory to assess static and dynamic brain connectivity: Approaches for building brain graphs," *Proceedings of the IEEE*, vol. 106, pp. 886–906, 2018.

[93] K. J. Friston, P. Jezzard, and R. Turner, "Analysis of functional mri time-series," *Human brain mapping*, vol. 1, no. 2, pp. 153–171, 1994.

[94] S. M. Smith, "The future of fmri connectivity," *Neuroimage*, vol. 62, pp. 1257–1266, 2012.

[95] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, and M. W. Woolrich, "Functional connectomics from resting-state fmri," *Trends in cognitive sciences*, vol. 17, pp. 666–682, 2013.

[96] W. M. Association, "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects," *JAMA*, vol. 310, no. 20, pp. 2191–2194, 11 2013.

[97] C. Brodbeck, A. Presacco, and J. Z. Simon, "Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension," *NeuroImage*, vol. 172, pp. 162–174, 2018.

[98] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[99] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and clinical neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[100] M. Plechawska-Wojcik, M. Kaczorowska, and D. Zapala, "The artifact subspace reconstruction (asr) for eeg signal correction. a comparative study," in *International Conference on Information Systems Architecture and Technology.* Springer, 2018, pp. 125–135.

[101] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Amica: An adaptive mixture of independent component analyzers with shared components," *Swartz Center for Computatonal Neursoscience, University of California San Diego, Tech. Rep*, 2012.

[102] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, 2019.

[103] L. V. der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, 2008.

[104] Turken, U, and N. F. Dronkers, "The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses," *Frontiers in System Neuroscience*, vol. 5, p. 1, 2011.

[105] R. Leech and J. Smallwood, "The posterior cingulate cortex: Insights from structure and function," *Handbook of clinical neurology*, vol. 166, pp. 73–85, 2019.

[106] N. Ding, L. Melloni, H. Zhang, X. Tian, and D. Poeppel, "Cortical tracking of hierarchical linguistic structures in connected speech," *Nature neuroscience*, vol. 19, no. 1, pp. 158–164, 2016.

[107] A. Kösem, A. Basirat, L. Azizi, and V. van Wassenhove, "High-frequency neural activity predicts word parsing in ambiguous speech streams," *Journal of neurophysiology*, vol. 116, no. 6, pp. 2497–2512, 2016.

[108] L. Meyer, M. J. Henry, P. Gaston, N. Schmuck, and A. D. Friederici, "Linguistic bias modulates interpretation of speech via neural delta-band oscillations," *Cerebral Cortex*, vol. 27, no. 9, pp. 4293–4302, 2017.

[109] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski *et al.*, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Communications biology*, vol. 4, no. 1, pp. 1–10, 2021.

[110] L. Meyer, "The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms," *European Journal of Neuroscience*, vol. 48, no. 7, pp. 2609–2621, 2018.

[111] D. M. Corey, W. P. Dunlap, and M. J. Burke, "Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations," *The Journal of general psychology*, vol. 125, pp. 245–261, 1998.

[112] B. Accou, J. Vanthornhout, H. V. hamme, and T. Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 812, 2023.

[113] S. L. Bressler, A. Kumar, and I. Singer, "Brain synchronization and multivariate autoregressive (mvar) modeling in cognitive neurodynamics," *Frontiers in Systems Neuroscience*, p. 155, 2022.

[114] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, p. 026113, 2004.

[115] F. M. Branzi and M. Lambon Ralph, "The role of the posterior medial network in language comprehension: Dissociating construction of episodic versus semantic representations," *bioRxiv*, pp. 2022–09, 2022.

Publications

Journal paper

[1] Zhou, D., Zhang, G., Dang, J., Unoki, M. and Liu, X. (2022). Detection of brain network communities during natural speech comprehension from functionally aligned EEG sources. Frontiers in Computational Neuroscience, 77.

International conference

[2] Zhou, D., Zhang, G., Dang, J., Wu, S. and Zhang, Z. (2020). Neural Entrainment to Natural Speech Envelope Based on Subject Aligned EEG Signals. Interspeech (pp. 106-110).

[3] Zhou, D., Zhang, G., Dang, J., Wu, S. and Zhang, Z. (2020, December). A Multi-subject Temporal-spatial Hyper-alignment Method for EEG-based Neural Entrainment to Speech. In 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 881-887).

[4] Zhou, D., Unoki, M., Zhang, G. and Dang, J. (2022). Reconstruction of speech spectrogram based on non-invasive EEG signal. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 275-279).

Domestic conference

[5] Zhou, D., Huang, J. and Dang, J. (2018). Effects of Orthographic and

Phonological Information on Text Understanding. The 2018 Spring Meeting of The Acoustical Society of Japan.

[6] Zhou, D., Zhang, G. and Dang, J. (2022). Investigating the neural responses to continuous speech based on reconstructed source signal from EEG. The 2022 Spring Meeting of The Acoustical Society of Japan.

Other publication

[7] Huang, J., Zhou, D. and Dang, J. (2017) Estimation of Speech-planning mechanism based on eye movement. The 2017 Autumn Meeting of The Acoustical Society of Japan.

[8] Huang, J., Zhou, D. and Dang, J. (2017). Investigation of Speech-Planning Mechanism Based on Eye Movement. In International Seminar on Speech Production. Springer, Cham (pp. 175-187).

[9] Zhou, D., Huang, J. and Dang, J. (2018). Investigation of the Comprehension Process during Silent Reading based on Eye Movements. In 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 165-169).

[10] Zhang, Z., Zhang, G., Dang, J., Wu, S., Zhou, D. and Wang, L. (2020). EEG-based Short-time Auditory Attention Detection using Multi-task Deep Learning. Interspeech (pp. 2517-2521).

[11] Yang, K., Zhuang, X., Zhou, D., Wang, L. and Zhang, Z. (2023). Auditory Attention Detection in Real-Life Scenarios Using Common Spatial Patterns from EEG. Interspeech (Accept).