Investigating Neural Mechanisms of Word Learning and Speech Perception

A THESIS
SUBMITTED FOR THE DEGREE OF
Doctor of Philosophy
IN THE
Faculty of Engineering

By

Akshara Soman

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Department of Electrical Engineering
Indian Institute of Science
Bangalore – 560 012 (INDIA)

January 17, 2024

© Akshara Soman

January 17, 2024

All rights reserved

DEDICATED TO

My parents

Acknowledgements

Writing this thesis would not have been possible without the support, encouragement, and guidance of many individuals. I would like to express my heartfelt gratitude to all those who played a significant role in shaping my PhD journey. First and foremost, I am deeply thankful to my PhD advisor, Dr Sriram Ganapathy, for allowing me to work alongside him in the Learning and Extraction of Acoustic Patterns (LEAP) lab. Throughout my PhD journey, Sriram has been an exceptional mentor, providing invaluable guidance, support, and encouragement. I am grateful for his empathy, understanding, and unwavering presence during times of mental distress. His unwavering belief in my capabilities and his continuous motivation pushed me to complete this work.

I extend my heartfelt thanks to my comprehensive examination panel members, Prof. Supratim Ray, Prof. Rajiv Soundararajan, Prof. Muthuvel Arigovidan, and Prof. Chandra Sekhar Seelamantula, for their appreciation of my work and their valuable suggestions. Their insights have greatly contributed to the refinement of my research. I would also like to express my gratitude to all my course instructors who imparted technical knowledge and fostered my critical thinking skills, playing a significant role in my development as a researcher.

I am sincerely grateful to Dr. Neeraj Sharma, Dr. Purvi Agrawal, Dr. Shreyas Ramoji, and Varun bro for the technical discussions we had during the initial stages of my Ph.D. Their different perspectives and interpretations provided invaluable insights for my doctoral research and enhanced my overall research abilities. I would also like to express my gratitude to the project staff in the LEAP lab, including Madhavan, Kinsuk, Venkat, Anshul, Prathibha, Kiran, and Vidhi, who have been associated with me over the years. I would like to express my appreciation to all the current and former members of the LEAP Lab for their love, support, and the memorable moments we shared. I am also grateful for my interactions with national and international colleagues during conferences and workshops, which enriched my research experience.

I would also like to thank everyone who volunteered for my experiments and Axxonet for providing the facility to do EEG recordings. I would also like to acknowledge Prof. Nima Mesgarani for being a mentor during my internship and for providing guidance during my visit to Columbia University.

My experience at IISc would not have been complete without the company and support of Vinila Chechi, Harsha, Asha, Vandana, Kundan, and many others. I am particularly grateful to the Malayali table in the C mess and SIMA for being my IISc family.

Finally, none of this would have been possible without the unwavering support of my parents, Achan, and Amma. They have allowed me the freedom to pursue my interests and have been a constant source of motivation and unconditional love. I dedicate this thesis to them. Achan, thank you for your wisdom and guidance, and Amma, thank you for showing me that hard work can overcome any challenge. Your dedication to your work has been a tremendous inspiration to me. Last but not least, I appreciate my husband, Ojus Mohan, for being my rock throughout this arduous journey. Your presence, unwavering support, and composed nature during my most challenging times have made this endeavour bearable.

Thank you to everyone mentioned above and to all those who have supported me along this rewarding journey.

Abstract

Language learning and speech perception are remarkable feats performed by the human brain, involving complex neural mechanisms that allow us to understand and communicate with one another. Unravelling the mysteries of these mechanisms has far-reaching implications, from theories of human cognition to developing effective language learning strategies and advancing speech technology. By employing a multidisciplinary approach encompassing neural investigations using EEG signals, behavioural analyses, and machine learning perspectives, this thesis seeks to shed light on the underlying processes involved in word learning and speech perception.

The thesis is divided into three parts. The first part examines how imitation-based learning of foreign sounds is captured in the EEG signals. In this listen-and-reproduce setting, subjects were introduced to words from a foreign language (Japanese) and English. The subjects were also asked to articulate the words. The results show that time-frequency features and phase in the EEG signal contain information for language discrimination. Further analysis showed that speech production improved over time, and the frontal brain regions were involved in language learning. These findings suggest the potential of EEG for personalized language exercises and assessing learners' abilities.

The next part of the thesis investigates what changes in neural patterns occur when semantics are introduced and presented in a sentence context. The participants listen to Japanese words in an English sentence, once before understanding the semantics of these words and later with the semantic exposure. We quantify the learning patterns in the EEG signal. Notably, a delayed P600 component emerges for Japanese words, suggesting short-term memory processing, unlike the N400 typically seen for a semantic anomaly in the known language. We have also shown the association of the P600 amplitude with the similarity of newly learned to the known language. The brain regions associated with semantic learning are also identified in this study using the EEG data. These findings demonstrate that there are differences in the underlying cognitive processes involved in rapid and long-term language learning.

In the final part of the thesis, we analyze the neural mechanisms of human speech comprehension using a match-mismatch classification of the continuous speech stimulus and the neural response (EEG). We make three significant contributions on this front - i) Illustrate the role of word boundaries in continuous speech comprehension for the first time, ii) Elicit the encoding of speech data (acoustics) as well as the text data (semantics) in the EEG signal, and, iii) Increased signature of semantic content (text) in the EEG data in acoustically challenging environments of dichotic listening. Previous studies focused on fixed-duration segments without considering the variable length processing of speech in the brain. Our approach involved processing speech and EEG signals with convolutional layers, word boundary-based pooling, and inter-word context through a recurrent layer. We introduced a novel loss function based on Manhattan similarity. The findings have potential applications for understanding speech recognition in noise, brain-computer interfaces, and attention studies.

Overall, this thesis contributes to our understanding of language learning, speech comprehension, and the underlying neural mechanisms. Through the analysis of EEG signals, this work provides valuable insights into the processing of familiar and unfamiliar languages, the effects of semantic dissimilarity, and the role of word boundaries in sentence comprehension. These findings have implications for both human language learning and the development of machine systems aimed at understanding and processing speech.

Publications based on this Thesis

Peer-reviewed Journal Papers

1. A. Soman, P. Ramachandran, and S. Ganapathy, “ ERP Evidences of Rapid Semantic Learning In Foreign Language Word Comprehension,” *Frontiers in Neuroscience*, (2022): p.178.

2. A. Soman, Madhavan C. R., K. Sarkar, and S. Ganapathy, “An EEG Study On The Brain Representations in Language Learning,” *IOP Journal on Biomedical Physics and Engineering Express*, 5(2), (2019): p.25041.

In preparation

1. A. Soman and S. Ganapathy, “ Impact of Semantic Cues on Speech Perception During a Dichotic Listening Task,” To be submitted to *Journal of Neural Engineering*.

2. A. Soman, P. Ramachandran, and S. Ganapathy, “ An EEG dataset exploring semantic learning with audio-visual data,” To be submitted to *Data in Brief* .

Peer-reviewed Conference Papers

1. A. Soman, V. Sinha and S. Ganapathy, “Enhancing the EEG Speech Match Mismatch Tasks With Word Boundaries,” *Proc. Interspeech* (2023).

2. V. Krishnamohan, A. Soman, A. Gupta and S. Ganapathy, “Audiovisual Correspondence Learning in Humans And Machines,” *Proc. Interspeech* (2020).

3. K. Praveen, A. Gupta, A. Soman and S. Ganapathy, “Second Language Transfer Learning in Humans and Machines Using Image Supervision,” *IEEE ASRU* (2019).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Speech Perception and Word Learning

The field of speech perception and word learning is a multidisciplinary area of research that aims to understand how humans process and comprehend spoken language. It combines insights from psychology, neuroscience, linguistics, cognitive science, and computer science to unravel the complex mechanisms involved in perceiving and interpreting speech.

Speech perception refers to the process by which humans extract and interpret linguistic information from auditory signals. It involves decoding the acoustic properties of speech sounds, such as phonemes, prosody, and phonetic cues, to recognize and differentiate words, sentences, and other linguistic units. Prior works have investigated various aspects of speech perception, including speech segmentation, phoneme categorization, temporal processing, and the role of contextual information.

Word learning in a foreign language is a complex cognitive process involving the acquisition and integration of new vocabulary into an individual's existing linguistic knowledge. Neural studies have revealed interesting differences in brain activation when exposed to native and non-native languages [2].

The process begins with the auditory system processing the acoustic signal of the unfamiliar word, analyzing its phonological features like sounds and syllables. The brain then tries to match these features with known linguistic representations and, if literate, considers the word's orthographic form.

As learners become more familiar with the foreign word, neural plasticity comes into play. This is the brain's capacity to reorganize and form new neural connections to accommodate the newly acquired knowledge. Studies in bilingual language acquisition, memory, attention, and perception have demonstrated ongoing neuroplasticity for language learning in the adult brain, which was previously not well understood [3, 4, 5, 6, 7].

Furthermore, context significantly influences word learning, and the brain relies on contextual cues to grasp the meaning and usage of foreign words. The dynamic interaction among various brain regions facilitates gradually incorporating new vocabulary into the learner's language repertoire. This thesis focuses on comprehending how adults acquire and learn words from a foreign language.

Research in the field of speech perception and word learning employs various methodologies to explore the underlying neural and cognitive processes. These methodologies include behavioural experiments, psycholinguistic studies, neuroimaging techniques (such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG)), computational modelling, and machine learning approaches. Researchers can use these methods to investigate the intricate connections between brain mechanisms, linguistic representations, and cognitive processes involved in speech perception and language understanding.

Understanding the neural encoding of speech perception and language learning has important implications for various areas. It can enhance our understanding of language disorders such as dyslexia, aphasia, and developmental language impairments [8, 9, 10, 11]. It can also inform the development of assistive technologies for individuals with speech and language disabilities. In addition to language disorders, this understanding can also be used to improve the design of hearing aids and cochlear implants [12, 13]. By understanding how the brain encodes speech sounds, researchers can develop more effective devices for restoring hearing [14]. Furthermore, studying the neural mechanisms underlying speech perception and word learning can contribute to improving speech technologies like automatic speech recognition systems [15], natural language processing algorithms, and machine translation tools.

Word learning and speech perception are intricate processes that engage multiple neural mechanisms distributed across various brain regions. From the initial perception of speech sounds to the integration of semantic and contextual information, these mechanisms work together to enable the acquisition and understanding of spoken language. Further research in this field will continue to uncover additional details about the neural mechanisms underlying word learning and speech comprehension, contributing to our understanding of human communication.

In recent years, there has been a growing interest in using artificial neural networks to model the neural mechanisms of word learning and speech comprehension [16, 1, 17]. Neural networks are a type of artificial intelligence that can learn to perform tasks by analyzing data. They are effective in modelling a variety of cognitive functions, including language processing. They offer a promising new approach to understanding the neural mechanisms of language processing.

1.2 Relevant Background

1.2.1 Human Brain Anatomy

The brain is a complex organ that controls thought, memory, emotion, touch, motor skills, vision, breathing, temperature, hunger and every process that regulates our body. At a high level, the brain can be divided into the cerebrum, brainstem and cerebellum. Cerebrum is the largest part of the brain and is composed of right and left hemispheres. It performs higher functions like interpreting touch, vision, and hearing, as well as speech, reasoning, emotions, learning, and fine control of movement.

The cerebral hemispheres have distinct fissures, which divide the brain into lobes. Each

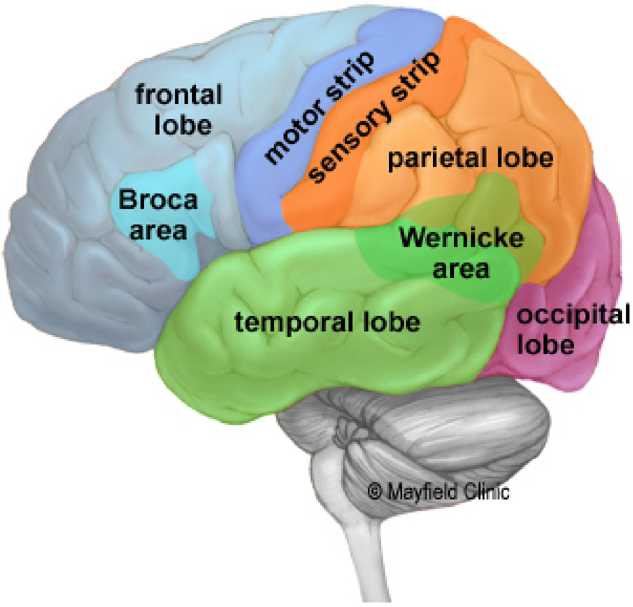


Figure 1.1: Human Brain Anatomy: The cerebrum is divided into four lobes: frontal, parietal, occipital, and temporal. (©Mayfield Clinic)

hemisphere has 4 lobes: frontal, temporal, parietal, and occipital (Figure 1.1). Each lobe may be divided, once again, into areas that serve very specific functions. It’ s important to understand that each lobe of the brain does not function alone. There are very complex relationships between the lobes of the brain and between the right and left hemispheres.

1.2.2 Language Learning and Brain

Broca’s area and Wernicke’s area are the main language-related regions in the brain. Broca’s area is in the left frontal lobe, and Wernicke’s is in the superior temporal lobe, between the primary auditory cortex and the angular gyrus. These areas play a crucial role in regulating the speech process [18, 19]. When recruiting the Broca’s area, there is a significant difference in the usage between one’s native language and a foreign language, which indicates the challenge of speaking a foreign language as fluently as one’s native tongue [20, 21].

The critical period for infants to learn languages effectively is around the age of twelve [22]. Before reaching this age, children can acquire both their native and foreign languages using the same language-related brain areas. However, after the age of twelve, they may employ other areas to learn new languages. Consequently, they need to adopt a new efficient strategy. Children under twelve learn through imitation, displaying greater flexibility, spontaneity, and openness to new experiences, whereas adults tend to study the new language more consciously.

1.2.3 Learning and Memory

In theory, external stimuli can potentially modify the brain's structures and functions to facilitate complex cognitive processes like learning a new language [21]. Learning can trigger the brain's stem cells to generate new cells and create new neural connections. This phenomenon, known as brain plasticity, enables learners to overcome difficulties even after the age of twelve by selecting efficient learning methods.

Memory exists in two main forms: short-term memory and long-term memory [23]. Short-term memory is easily forgettable and lasts only briefly, ranging from a few seconds to a few hours. On the other hand, long-term memory can persist for days, months, or even years. Memories are initially stored as short-term but can be converted into long-term through repetitive and intense stimulation. When using their native language, people predominantly rely on long-term memory, while declarative memory comes into play when speaking a foreign language.

Memory can be broadly categorized into declarative memory and non-declarative memory [24]. Declarative memory is used to recall stored information about past events, while procedural memory does not involve conscious recollection. Declarative memory forms and fades relatively quickly, whereas non-declarative memory is more durable and requires repetition and significant time to establish. When using their native language, individuals mainly rely on non-declarative memory, whereas declarative memory is employed when speaking a foreign language.

1.2.4 Electroencephalography

One of the simplest ways of furthering the understanding of speech perception is by recording neural responses using electroencephalography (EEG). The EEG is a non-invasive neural imaging technique that measures electrical activity in the brain by placing electrodes on the scalp [25]. It has been demonstrated that the EEG signal recorded during a speech listening task contains information about the stimulus [26, 27, 28, 29]. [26] demonstrates that EEG reflects categorical processing of phonemes within continuous speech. Additionally, [28] shows that EEG captures linguistic representations such as semantic dissimilarity. Similarly, [27] reveals that the EEG recorded during the listening state carries language information of the stimulus.

Electrical signals resulting from cortical synaptic activity fluctuate within the range of 10 to 100 milliseconds. EEG and MEG (magnetoencephalography) are widely accessible technologies capable of tracking these rapid dynamic changes due to their adequate temporal resolution. But they have limited spatial resolution compared to other brain imaging methods like computer tomography (CT), positron-emitted tomography (PET) and magnetic resonance imaging (MRI). EEG has excellent temporal resolution compared to most neuroimaging techniques, aside from being comparatively inexpensive.

1.2.5 Event-Related Potential

Event-related potentials (ERPs) refer to changes in voltage observed in the ongoing electroencephalogram (EEG) that are time-locked to specific events, such as the onset of a stimulus or the execution of a manual response [30]. These ERPs are obtained by averaging the EEG data from several epochs, each time-aligned to the event of interest. As more trials are added to the average, the random noise present in the data diminishes, while the consistent signal related to the event gradually emerges from the background noise.

It is crucial to understand that while the ERP waveform at a given moment reflects synaptic activity at that time, it does not exclusively represent neural activity that started precisely at that moment [31]. The resulting positive and negative deflections' amplitude, latency, and topography are used as indicators of the underlying mental processes. ERPs are valuable tools for investigating cognitive processes involved in perception, language, attention, memory, and other mental functions. Important ERP components pertaining to language research are explained in chapter 3.

1.2.6 Machine Learning Methods

We will be utilizing various machine learning algorithms and models to analyze EEG data in the upcoming chapters. The following section will offer a concise background on pertinent techniques.

1.2.6.1 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm used in machine learning for classification and regression tasks. It excels particularly in binary classification problems, where the goal is to categorize elements into two groups.

SVM [32] classifies data by determining the optimal decision boundary that maximally separates different classes. It seeks the best hyperplane to maximize the margin between support vectors, which are the data points closest to the decision boundary.

The distinguishing feature of SVM is its ability to handle both linear and non-linear classification problems. This is achieved through the use of a kernel trick, which implicitly maps input data into a higher-dimensional space. For linear datasets, the kernel can be set as ‘linear’, while for non-linear datasets, options include ‘rbf’ (radial basis function) and ‘polynomial’. This flexibility allows SVM to effectively capture complex relationships within the data.

1.2.6.2 Correlation Analysis

Correlation analysis is a statistical method used to evaluate the strength and direction of the linear relationship between two variables. The goal is to quantify how changes in one variable correspond to changes in another. The result of a correlation analysis is expressed as a correlation coefficient, typically denoted by “r”.

The correlation coefficient “r” is calculated using the following formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where,

- n is the number of data points,
- Σ represents the summation symbol,
- x and y are the individual data points of the two variables.

This formula computes the correlation coefficient by considering the covariance between the two variables and normalizing it by the product of their standard deviations. The coefficient ranges from -1 to 1, where 1 signifies a perfect positive correlation (both variables move in the same direction), -1 represents a perfect negative correlation (variables move in opposite directions), and 0 indicates no linear correlation. However, it's important to note that correlation does not imply causation, meaning that even if two variables are correlated, one does not necessarily cause the other to change.

1.2.6.3 Convolutional Layer

A Convolutional Layer is a fundamental component in Convolutional Neural Networks (CNNs) [33], a class of deep learning models widely used for image and video analysis. The convolutional layer is designed to automatically and adaptively learn hierarchical representations of input data. The key operation in a convolutional layer is convolution, where a small filter (also known as a kernel or receptive field) slides or convolves across the input data. At each position, the filter computes the dot product between its weights and the corresponding region of the input.

Convolutional layers are effective for several reasons. They share parameters, reducing the number of weights compared to fully connected layers, which is particularly advantageous when dealing with high-dimensional inputs. Moreover, the hierarchical structure enables the network to learn complex features by combining simpler ones, enhancing its ability to capture spatial hierarchies and patterns.

1.2.6.4 LSTM

LSTM stands for Long Short-Term Memory, representing a type of recurrent neural network (RNN) commonly employed in deep learning. LSTMs excel at capturing long-term dependencies, making them particularly well-suited for sequence prediction tasks involving text, speech, and time series data. Introduced in 1997 by Hochreiter & Schmidhuber [34], LSTMs have undergone refinement and widespread adoption in the deep learning community.

LSTMs are designed to address the vanishing gradient problem that's present in traditional RNNs. The distinguishing feature of LSTMs is their ability to selectively retain or discard information over long sequences. Each LSTM unit contains a memory cell, input gate, forget gate, and output gate. The memory cell stores information over extended periods, while the gates regulate the flow of information into and out of the cell.

1.3 Thesis Outlook: Neural, Behavioral and Machine Learning Perspectives

1.3.1 Isolated Word Learning

When ones start learning a language, one imitates the sounds they hear without deliberate thinking. As one improves their learning, they rely more on the context to understand what words mean. Further, the mechanisms of first language (L1) and second language (L2) acquisition, as well as their perception, exhibit significant differences [35]. For instance, a child may be exposed to their native language throughout the day, every day, whereas an adult may encounter the foreign language primarily in the classroom setting. The way we learn our first language and a second language is different, and how we hear and understand them is different, too [36]. Certainly, the process of word learning continues throughout a person's life, and the vast majority of their vocabulary is acquired after early childhood [37, 38].

The study conducted by Kuhl et al. [39] describes the development of language during the early years of life and the underlying mechanisms. During the first year of life, infants experience two notable changes in their speech perception skills. Firstly, their ability to distinguish phonetic sounds in a non-native language decreases, and secondly, they show improvement in recognizing phonetic sounds in their native language. However, it remains unclear whether similar changes can be observed in adults when they are learning a new language.

To explore these differences and gain a better understanding of the learning process through imitation, the thesis investigates the evolution of neural representations in individuals who are repeatedly exposed to unfamiliar words. By examining this phenomenon, we aim to address the fundamental question of whether the acquisition of speech sounds is reflected in the recorded EEG responses. Furthermore, our investigation seeks to determine whether EEG has the capability to detect and differentiate language-specific features, thereby providing valuable insights into the neural mechanisms underlying language discrimination. Through this line of inquiry, we aim to understand how imitation contributes to language learning and to advance the understanding of EEG as a potential tool for studying language acquisition processes.

1.3.2 Learning Words in Context

After examining these fundamental questions, the thesis shifts towards exploring how learning patterns change when semantics are introduced. This investigation holds the potential to uncover the neural signatures associated with rapid semantic learning. We aim to examine how neural responses evolve when newly acquired words are integrated into the context of a sentence, shedding light on the cognitive processes involved in sentence-level comprehension. To conduct these experiments, we will utilise words from a foreign language as unfamiliar stimuli, thus providing an opportunity to investigate the neural dynamics underlying the acquisition and utilisation of foreign vocabulary. By delving into these investigations, our aim is to enhance our understanding of how semantic knowledge is integrated into language learning and to uncover the neural mechanisms that support this process.

A significant portion of research on lexical acquisition in children has focused on understanding the cognitive processes involved in explicitly learning names for objects and actions and whether these processes are specific to language [40, 41, 42, 43]. On the other hand, studies on lexical acquisition in adults have predominantly revolved around second language acquisition. Lexical acquisition refers to the process by which individuals acquire and learn words, particularly in the context of language development. This research has emphasized exploring the cognitive and neural similarities (and differences) between word learning in an individual's native language and their second language [44, 45, 46, 47].

While these areas of research have provided valuable insights into how young children and bilingual adults learn words, the process of word learning in adults for their native language likely exhibits similarities and differences compared to that of young children and adult bilinguals. For example, growing evidence from electrophysiological studies suggests that infants may process novel and known word meanings using neural mechanisms similar to those in adults [48, 49, 50, 51]. However, there are important distinctions in how children and adults are exposed to words and how they learn them. Studies on preliterate children often investigate word learning in explicit training contexts, such as learning to name a novel object. In contrast, older children and adults primarily acquire words incidentally, especially during reading [52, 53, 38]. Given the prevalence of adult word learning and its differences compared to child lexical learning, it is not surprising that there is an increasing interest in this topic.

Word representations are intricate and multifaceted. Additionally, the word's meaning needs to be appropriately situated within the context of the mental lexicon's semantic landscape. For example, when learning the name of a new bird, the learner not only acquires specific information about the bird's characteristics (e.g., colour, size, feeding habits) but also links this knowledge to their existing understanding of birds and other creatures [54, 55]. Furthermore, the learner needs to grasp the contextual usage of the word, including how the novel noun interacts with other words like verbs or modifiers to be correctly used in sentences and discourse. Much of this understanding cannot be effectively measured through simple associations of novel words with objects.

The surrounding linguistic context in which a word is initially encountered plays a crucial role in the process of acquiring its meaning. Both children and adults have been shown to successfully fast-map novel word-object associations using recognition and comprehension tasks. However, little is known about the time course of this knowledge acquisition. To delve into this aspect of word learning, the researchers employ an electrophysiological index of word recognition called the N400, as detailed in our study (See chapter 3). This measure is sensitive to aspects of learning and memory utilization that may not always be evident in behavioural measures.

1.3.3 EEG Decoding of Continous Speech Perception

Having gained an understanding of the acquisition of novel sounds and words with their semantics, the thesis delves into investigating the neural correlates associated with continuous speech. In naturalistic speech, the listening cues are integrated in a continuous acoustic stream. To interpret and process this information in real time, the neural encodings representing these cues in terms of spatial, temporal, and rate factors need to be decoded by mechanisms that are not likely continuous [56]. The initial stage of these neural parsing and decoding mechanisms involves discretizing the continuous input signal and its initial neural representation. The idea that perception is "discrete" has been supported and explored in various contexts [57, 58].

The thesis explores the significance of word boundaries in speech perception in the final part. We develop deep learning-based models for stimulus (auditory) - response (EEG) modelling to accomplish this. In contrast to employing multi-trial analysis techniques, we propose utilizing single-trial analysis for continuous speech stimuli. Furthermore, we leverage machine learning models to examine the influence of word boundary information in more complex listening environments and its impact on speech perception. In the course of our investigation, we aim to uncover potential differences in the weighting of input cues, such as acoustics versus semantics, in speech perception under different listening conditions. This endeavour holds the potential to shed light on the hierarchical levels of speech perception and deepen our understanding of how different cues contribute to our comprehension of spoken language.

Taking these factors into account, the primary objectives of this thesis are formulated and outlined in the following section.

1.4 Problem Statement

This thesis aims to investigate the neural mechanisms underlying language processing and learning by analyzing electroencephalography (EEG) signals. The thesis attempts to address the following research questions:

1. What are the neural correlates of word discrimination in adults who are learning a new language?
2. How do the neural responses to known and unknown words change over the course of a word-learning task?
3. What are the effects of semantic dissimilarity on EEG signals during rapid foreign language word learning?
4. What is the role of word boundary information in sentence processing and its impact on EEG-based neural mechanisms of speech comprehension?
5. In the context of a dichotic listening task, what are the relative roles of semantics and acoustic cues in speech perception?

1.5 Analysis Techniques

As mentioned above, this thesis aims to investigate the neural mechanisms underlying language processing and learning by analyzing electroencephalography (EEG) signals. To achieve these goals, a comprehensive research methodology has been designed, combining multi-trial and single-trial analysis paradigms. Correlation analysis, ERP analysis, and machine learning techniques are employed to analyze the collected EEG data. This methodology has been chosen for its appropriateness in addressing the research objectives and providing insights into the neural correlates of language processing and learning.

The multi-trial analysis approach is utilized to examine the neural responses associated with word discrimination in adults learning a new language, as well as to investigate the changes in neural responses to known and unknown words over the course of a word-learning task. This approach is suitable for capturing averaged brain responses across multiple trials, providing reliable and reproducible findings regarding the neural correlates of word discrimination and learning.

To further explore the collected EEG data, correlation analysis is conducted to assess the relationships between different neural measures and behavioural variables. This analysis helps elucidate the associations between brain activity patterns and language processing performance, providing a deeper understanding of the neural mechanisms underlying language learning.

Moreover, ERP analysis is applied to investigate the effects of semantic dissimilarity on EEG signals during rapid foreign language word learning. This analysis allows for the examination of the time-locked neural responses associated with semantic processing and the impact of semantic relatedness on the neural mechanisms of word learning.

The single-trial analysis paradigm is employed specifically for the analysis of continuous speech. This approach allows for examining the temporal dynamics of brain activity at the level of individual trials, enabling the investigation of the importance of word boundary detection in human speech perception. The single-trial analysis provides valuable insights into the dynamic nature of language processing and learning.

Machine learning techniques are utilized to analyze the EEG data and extract meaningful features that can contribute to the classification and prediction of language processing and learning outcomes. These techniques enable the identification of relevant patterns in the data and potentially uncover novel insights into the neural basis of language processing.

By employing this comprehensive research methodology, combining multi-trial analysis, single-trial analysis (for continuous speech) and then analyzing the collected EEG data using correlation analysis, ERP analysis, and machine learning techniques, this thesis aims to address a few research questions regarding the neural mechanisms underlying language processing and learning.

1.6 Key Contributions

The major contributions of this thesis are given below:

1. The thesis provides new insights into the neural representations of known and unknown languages, as well as the evolution of neural responses in the brain for a language learning task. This is done by using electroencephalography (EEG) signals recorded while human subjects listen to English (familiar), Japanese (unfamiliar), and Hindi (native) language stimuli. The results show that time-frequency features and phase contain significant language discriminative information and identify brain regions responsible for language discrimination and learning. The thesis also provides evidence for consistent EEG representations and pattern formation during the language learning task and identifies the neural basis for the improvement of pronunciation over the course of trials for the Japanese language.

2. The thesis investigates the event-related potential (ERP) of EEG signals during rapid language learning in subjects who had no prior exposure to the Japanese language. The results show that semantically matched and mismatched end-words in English sentences elicit different EEG patterns even for newly learned Japanese words, similar to the native language case. The thesis also identifies the presence of a P600 component (delayed and opposite in polarity to those seen in the known language) and its topographical

distribution in the parietal region and left hemisphere. The thesis provides evidence for the absence of the N400 component in this rapid learning task, which suggests the association of N400 with long-term memory processing.

3. The thesis demonstrates the significance of word boundary information in sentence processing by relating EEG to speech input using a network of convolution layers and recurrent layers. The thesis uses word boundary-based average pooling to incorporate inter-word context and improve the match-mismatch (MM) classification accuracy to 93% on a publicly available speech-EEG dataset. The thesis shows that previous efforts achieved an accuracy of only 65-75% for this task, indicating the effectiveness of the proposed approach. The thesis also provides evidence for the importance of considering the discrete processing of speech in the brain for accurate analysis of neural mechanisms of speech comprehension.

4. The thesis extends this proposed MM task approach to more complex listening conditions like dichotic listening. The dichotic speech listening task revealed that EEG signals encode higher-level semantic information more effectively than acoustic envelope information, suggesting that the brain gives more importance to the semantic content of the auditory input under acoustically challenging environments.

5. The proposal of a multi-modal architecture for the MM task in mono-aural and dichotic listening scenarios is a major contribution of this study. The results indicate that the EEG signal jointly encodes the semantic and acoustic content of the stimulus, outperforming individual modalities of text and speech. This highlights the importance of considering multi-modal approaches when analyzing EEG responses to auditory stimuli. The formulation of a novel paradigm, exploration of different listening scenarios, the introduction of a new loss function, and the adoption of a multi-modal architecture all contribute to the field's progress.

Overall, the thesis makes significant contributions to the field of cognitive neuroscience by providing new insights into the neural basis of language learning and sentence processing. The thesis also proposes novel experimental paradigms and machine-learning models that can be used to study these cognitive processes further.

1.7 Key Assumptions and Scope of the Thesis

1. Sample size and diversity: The study was conducted with a limited sample size based on availability and time constraints. The participants mainly consisted of youth, particularly students from our institute campus. We assume that this sample will adequately represent the broader adult community. Appropriate statistical tests were employed for evaluation to establish the significance of the results.

2. Language stimuli: The study discussed in chapter 2 focuses on three languages English (L2), Hindi (L1) and Japanese (unknown) languages and chapter 3 focuses on two languages: English and Japanese. Hindi was chosen as the L1 language because it is the most widely used language in India, spoken by 41% of the population according to the 2011 census [59]. Subjects whose mother tongue is Hindi were recruited for that specific experiment. Japanese was selected as the unknown language due to its dissimilarity from both the L1 and L2 languages used in the study. Consequently, we assume that results based on Japanese stimuli will provide generalizability regarding novel language learning. Japanese does not belong to the same language family as English or Indian languages, presenting a notable contrast. It is not classified into major language families prevalent in India, such as Indo-Aryan or Dravidian. Instead, Japanese belongs to the distinct Japonic language family [60], considered unique and unrelated to other major language families. Our intention was to have subjects with no prior exposure to the novel language before our experiment. Finding subjects not exposed to Japanese was comparatively easier than finding those unexposed to Western languages in an Indian campus community.

3. EEG as a measure of neural activity: EEG provides superior temporal resolution but has limited spatial resolution compared to other neuroimaging techniques like fMRI. As a result, the scalp topography plots do not directly represent specific brain regions; rather, they indicate activations at the corresponding scalp electrode locations.

4. The studies discussed in this thesis focus on healthy adults. Hence, further research is warranted to extend the findings of this thesis to infants or the patient population.

5. The study was conducted in a laboratory setting, which may not fully reflect the real-world experience of learning a new language.

1.8 Organization of the thesis

The main findings of the thesis are organized into five chapters. Each chapter begins with a brief introduction that covers the background, challenges, and relevant literature on the topic addressed in the chapter. The introduction is followed by a description of the stimuli used, the subjects involved, and the data acquisition methods. The next section of each chapter explains the methods used or developed for the work. The results of the study are then presented and discussed in detail. Each chapter concludes with a summary of the key findings. The overall flow of the thesis is illustrated in Figure 1.2.

Chapter 2: Evolution of Neural Responses during Word Learning In Chapter 2, the focus is on the empirical findings obtained from the word learning task. This chapter presents an in-depth analysis of the neural responses recorded during the task. It explores the evolution of these neural responses over time, providing insights into the cognitive processes underlying word learning. The chapter discusses the implications of these findings and their contribution to our understanding of language processing and learning.

Chapter 3: ERP Evidences of Rapid Semantic Learning In Foreign Language Word Acquisition Chapter 3 examines the neural representations in known and unknown

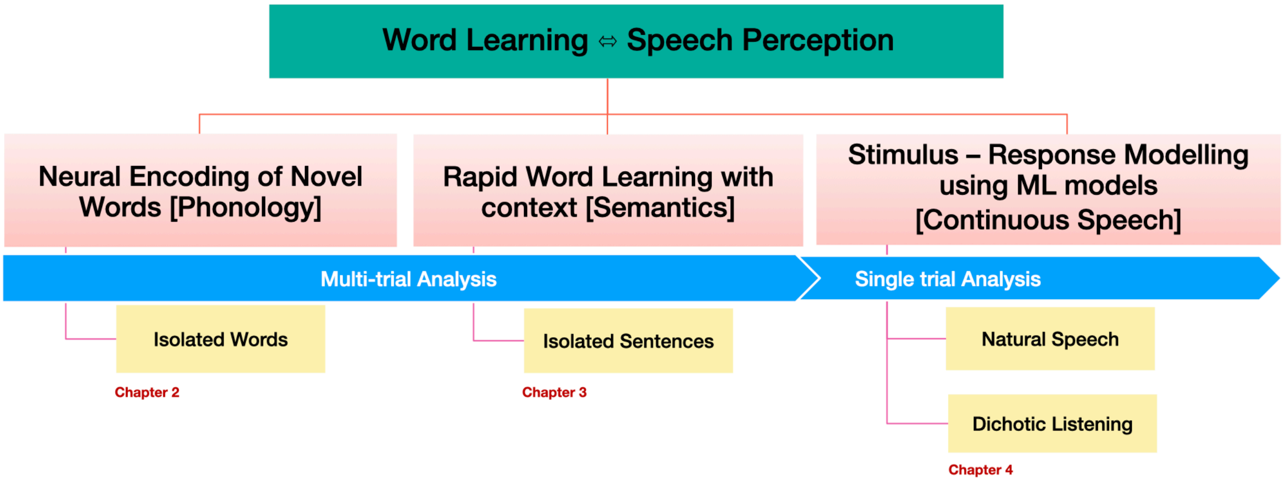

Word Learning ⇔ Speech Perception
Neural Encoding of Novel Words [Phonology]
Rapid Word Learning with context [Semantics]
Stimulus – Response Modelling using ML models [Continuous Speech]
Multi-trial Analysis
Single trial Analysis
Isolated Words
Chapter 2
Isolated Sentences
Chapter 3
Natural Speech
Dichotic Listening
Chapter 4

Figure 1.2: Overview of the thesis.

languages observed during the language learning task. The analysis explores the similarities and differences in EEG patterns between these languages, shedding light on the neural mechanisms underlying language familiarity and acquisition. The chapter draws connections between the findings and the research question of neural representations in known and unknown languages, providing a deeper understanding of language learning processes.

Chapter 4: Impact of Word Boundary Detection in Speech Comprehension Chapter 4 centres around the role of word boundary information in sentence processing and its impact on EEG-based neural mechanisms of speech comprehension. It presents the findings from the analysis of EEG signals, elucidating how word boundaries influence the neural processing of natural speech and speech in dichotic listening conditions. The chapter discusses the implications of these findings and their significance in understanding the cognitive aspects of speech comprehension.

Chapter 5: Conclusion and Future Directions The thesis concludes with chapter 5 summarising the main findings and their implications. This chapter revisits the research questions posed at the beginning and discusses how they have been addressed through empirical investigations. It highlights the thesis's contributions to the language learning field and identifies the proposed research's limitations. Furthermore, chapter 5 suggests potential avenues for future research, indicating the possibilities for expanding upon the current findings.

1.9 Chapter Summary

In this chapter, we have defined the problem statements of interest and given a birds-eye view of where the problem lies in the big picture of auditory neuroscience and speech processing. We also provide a broad overview of the signal processing and machine learning techniques used to understand neural mechanisms. The chapter outlined the thesis contributions and discussed the organization of the rest of the thesis.

Chapter 2

An EEG Study On The Brain Representations in Foreign Word Learning

2.1 Introduction

Speech is the easiest and most effective way of communication used by humans. Humans are inherently capable of distinguishing between sounds from familiar and unfamiliar languages when they listen to them. Prior works have shown that humans can instantaneously differentiate while listening to songs from known and unknown languages [61]. Also, the studies on brain activations showed interesting differences in the areas of the brain that are activated when exposed to native and non-native languages [2].

With the use of function magnetic resonance imaging (fMRI), it was seen that cerebral activations in the brain are more pronounced when presented with a foreign language compared to a known language [62]. Similarly, in speech production, the right frontal areas are more involved when the subject is attending to speak a new language. The activity in the right pre-frontal cortex was also found to be indicative of the language proficiency of the subject

[63].

The difference in response of the human brain to known and unknown stimuli has been of significant interest to facilitate the full understanding of the auditory encoding processes. For example, a stronger P300 peak was observed in the subject's electroencephalogram (EEG) signals when presented with their own names compared to the peak values observed when other stimuli were presented [64]. For infants, the representation in EEG for familiar language and foreign language as well as for familiar and unfamiliar talkers was analyzed in [65], where the delta and theta bands showed important differences.

In the case of learning, several studies have shown that with experience, we gain proficiency in an unknown language and the function and structure of the brain changes during this learning process [66, 67]. Similar to the task of musical training, the experience of learning a new language also includes changes in the brain states. The complexity of speech and language causes challenges in understanding the questions about how, when and where these changes occur in the brain.

With the exemption of a few studies that have attempted to quantify the anatomical changes in the brain during language learning [68, 69, 70], very little is known regarding the changes in the brain during a new language learning in terms of when these changes occur, and how they reflect in the learning. In this chapter, we attempt to quantify some of these questions at the representation level using electroencephalogram (EEG) recordings.

While primary language learning for most adults happens at a very young age, acquiring a new language can happen at any point in the lifetime. For the language learning task, the age of acquisition showed little impact in terms of brain representations when normalized for the proficiency levels [71]. The fundamental question of whether there is knowledge transfer from a known language to a new language is still open-ended. Several studies have shown that known languages play a key role in acquiring new languages. The first language was found to provide an understanding of the grammar [72, 73]. The popular hypothesis for secondary language learning is the establishment of a link between the representations of the new language to the features of the already-known language. Also, continued exposure to a foreign language can help learn the language faster [74, 75].

The task of learning a new language can be quite complicated to analyse. This can be done at multiple levels like phonemic, syllabic, word, or sentence levels. The evaluation of language learning can also be analyzed for multiple tasks like reading tasks, spontaneous speaking etc. This study aims to understand the major differences in brain representations at a word level from a familiar and an unfamiliar language. Additionally, we propose a method to perform trial-level analysis to understand the changes in the representation of words when the subject listens to words from an unfamiliar language.

We record EEG signals from the subjects when the subjects are presented with word segments from a familiar and an unfamiliar language. Along with EEG signals, we also record behavioural data from the subject where the subject reproduces the stimuli presented to him. The key findings from the chapter can be summarized as follows,

- With various feature-level experiments, we identify that the time-frequency representations (spectrogram) of EEG signals carry language discriminative information. These features are also verified for two separate tasks, English versus Japanese and Hindi (native language) versus Japanese.

- The brain regions containing the most language discriminative information are in the frontal cortex and the temporal lobe (aligned with some previous fMRI studies [2]).

- It is seen that the inter-trial variations are more pronounced for the words from unfamiliar language than those from the familiar language in both EEG signals and spoken audio signals. Furthermore, the inter-trial variations in the spoken audio are correlated with those from the listening state EEG representations.

- The EEG signals for the Japanese stimuli are more correlated with the audio signal than those for the English stimuli indicating a higher level of attention to Japanese stimuli.

To the best of our knowledge, this study is one of the first to probe the linguistic differences in EEG level and uncover the language learning process from single-trial EEG analysis.

The rest of this chapter is organized as follows. In Sec. 2.2, we describe the data collection procedure and the pre-processing steps used for EEG data preparation. The feature extraction of EEG signals and the classification between the two languages is described in Sec. 2.3. A similar analysis is done to extract features and to classify the spoken audio signals and this is also described in Sec. 2.3. The evidence of language learning is established in Sec. 2.4. The inter-trial analysis performed on the EEG and the audio signals is described in Sec. 2.4.2. The relationship between the EEG and the audio signals is analyzed and described in Sec. 2.4.2.2. In Sec. 2.5, we discuss the findings from this work and contrast it with previous studies. Finally, a summary of the chapter is also provided in Sec. 2.6.

2.2 Materials and Methods

2.2.1 Subjects

All the participants were Indian nationals with self-reported normal hearing and no history of neurological disorders. In the first experiment, English and Japanese language words were used while in the second subsequent experiment, Hindi and Japanese language words were used. The first experiment had 12 subjects while the second experiment had 5 subjects. In the English/Japanese experiments, six subjects were male (median age of 23.5) and six were female (median age of 24). The subjects who participated in the first experiment (English versus Japanese) were native speakers of south Indian languages, or Hindi. All the subjects in the first experiment setup had intermediate or higher levels of English proficiency. Hence, English can be considered their L2 language, and Japanese is an unknown language.

In the second experiment of Hindi versus Japanese, all the subjects were native speakers of Hindi. A separate set of subjects was recruited for this second experiment. For these subjects,

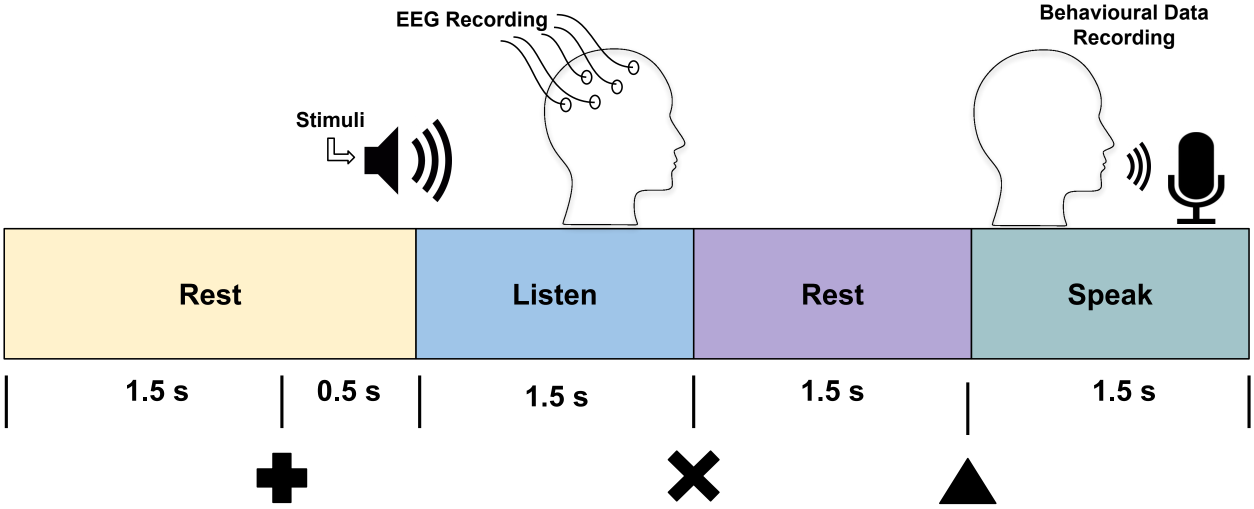


Figure 2.1: Experimental setup used for EEG and behavioral data collection

Hindi is their L1 language, and Japanese is an unknown language. In the Hindi/Japanese experiments, 2 subjects were male (median age of 22) and 3 were female (median age of 23).

2.2.2 Experimental Paradigm

Each block of the recording procedure consisted of five phases as illustrated in Figure 2.1. The first phase was the rest period of 1.5s duration followed by a baseline period of 0.5s. The subjects were instructed to attentively listen to the audio signal played after the baseline period. Then, they were given a rest of 1.5s where they were encouraged to prepare for overt articulation of the stimuli. The last phase is the speaking phase, where the subject was asked to speak the word overtly. The spoken audio was recorded using a microphone placed about one foot from the subject. The subjects were alerted about the change in phase by the display of a visual cue in the centre of the computer screen placed in front of them. The participants were asked to refrain from movement and to maintain visual fixation on the centre of the computer screen in front of them. All subjects provided written informed consent to take part in the experiment. The Institute Human Ethical Committee of the Indian Institute of Science, Bangalore approved all procedures of the experiment.

English		Japanese		Hindi	
Word	**Duration (s) (# units)**	**Word**	**Duration (s) (# units)**	**Word**	**Duration (s) (# units)**
beg	0.50 (3)	南極	0.82 (4)	दरवासा	0.73 (4)
cheek	0.67 (3)	抜き打ち	0.83 (4)	चावल	0.62 (3)
ditch	0.70 (3)	仏教	0.77 (3)	कहानी	0.63 (3)
good	0.50 (3)	弁当	0.72 (3)	धन्यवाद	0.88 (4)
late	0.77 (3)	偶数	0.76 (2)	आसमान	0.80 (4)
luck	0.64 (3)	随筆	0.83 (3)	आदमी	0.61 (4)
mess	0.60 (3)	先生	0.74 (4)	बचपन	0.74 (4)
mop	0.54 (3)	ポケット	0.82 (3)	पुजारी	0.74 (3)
road	0.59 (3)	計画	0.84 (4)	अलमारी	0.72 (4)
search	0.76 (3)	ミュージカル	0.83 (4)	सुप्रभात	0.82 (4)
shall	0.70 (3)	ウィークデイ	0.76 (4)	परिवार	0.69 (4)
walk	0.66 (3)	行政	0.80 (3)	किसान	0.68 (3)

Table 2.1: The list of stimuli used for the experiments, the duration of the words in seconds and the number of speech units. English and Hindi are phonetic languages while Japanese is a syllabic language. The first experiment uses the 12 English and 12 Japanese words while the second experiment uses the 12 Hindi and 12 Japanese words.

2.2.3 Stimuli

In each experiment, the stimuli set contained words from 2 languages. The words were chosen such that they have uniform duration and speech unit variability. In the first experimental setup, the stimuli set includes 12 English words and 12 Japanese words (Table 2.1). The duration of all audio stimuli ranges from 0.5s to 0.82s. In the second experimental setup, the stimuli set includes 12 Hindi words (native language of the subject) and the same 12 Japanese words (Table 2.1)) from the first experiment.

The Japanese was the unfamiliar language for all the subjects who participated in this experiment. Each word was presented 20 times to a subject. In the first experimental setup, all the trials of English and the first 10 trials of Japanese were presented in random order, while the last 10 trials of Japanese were presented in a sequential manner. In the second experimental setup using Hindi and Japanese language words, all the trials were presented in a random order.

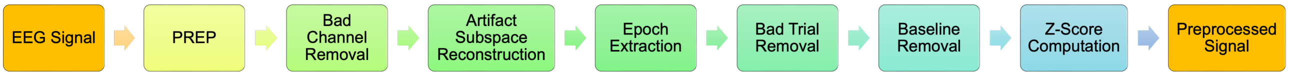


Figure 2.2: Pre-processing pipeline

The stimuli were presented at a comfortable and constant volume from a loudspeaker in front of the subject.

2.2.4 Data Acquisition

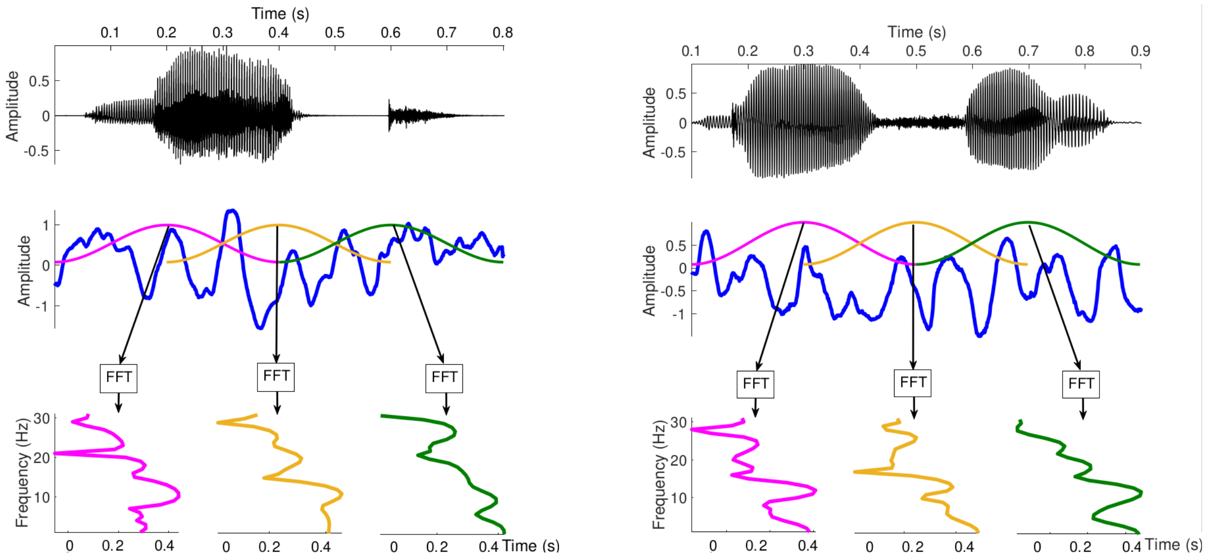
The stimuli presentation and EEG recording was carried out with the Brain Electrical Scan System (BESS) of Axxonet System Technologies, India. The EEG signals were recorded using a BESS F-32 amplifier with 32 passive electrodes (gel-based) mounted on an elastic cap (10/20 enhanced montage). The electrode layout is given in the Appendix for reference. The EEG data were recorded at a sampling rate of 1024 Hz. A separate frontal electrode (Fz) was used as ground and the average of two earlobe electrodes was used as reference. The channel impedances were kept below 10 kOhm throughout the recording. The EEG data were collected in a sound-proof, electrically shielded booth. A pilot recording confirmed that there was minimal line noise distortion or other equipment related artifact. In this study, all the analyses are performed with EEG signals recorded at listening state and with the audio signals used in stimuli as well as the spoken audio (behavioral data) collected from the subjects.

2.2.5 EEG Preprocessing

The EEG signals have a low signal-to-noise (SNR) ratio [76]. Hence, properly designed pre-processing steps are required to enhance SNR and remove unwanted artefacts from data. The pre-processing pipeline used in this chapter is shown in Figure 2.2. As the first step, we filter the EEG data using a 0.1 Hz fourth-order high-pass Butterworth filter to remove the DC drift. Then, the signal is low pass filtered using a 70 Hz fourth-order filter. The 50Hz line noise is suppressed using a notch filter. The channels with high levels of noise in each subject's recording are found using the PREP pipeline [77]. The artifacts such as eye blinks and muscle movements are suppressed using the Artifact Subspace Reconstruction technique [78]. Then, we extract data epochs for various stages of the experiment (such as rest, listening, and speaking) using the EEGLAB [79]. Any epoch with a magnitude of more than 3 standard deviations is removed from future analysis (bad trial removal). The baseline average computed from the 0.5-second-long baseline period is used for baseline subtraction. For each subject, we standardize the neural response of each EEG channel to ensure zero mean and unit variance. The entire preprocessing pipeline was implemented using the EEGLAB toolkit in MATLAB. The subsequent analysis discussed in this chapter was carried out using MATLAB software.

2.3 Language Classification in EEG signals

The language classification approach is used to identify the key features that discriminate the EEG representations of familiar and unfamiliar language. In particular, we try to uncover the best feature and classifier settings for discriminating English and Japanese from EEG signals (and Hindi versus Japanese from the second experimental setup). In these experiments, the chance accuracy is 50%. The training data set consists of 70% of the trials of each stimulus and the rest of the trials form the evaluation set for each subject. The classification is conducted individually for each subject. However, performing inter-subject classification for this task has proven to be challenging. A support vector machine (SVM) with a linear kernel has been used as the classifier to validate the performance of different feature extraction methods. The input data to the SVM classifier is normalized to the range of 0 to 1 along each feature dimension. The SVM-based classification is implemented in MATLAB using the publicly available LIBSVM package [80]. Later experiments comparing different classifier models (LDA, Gaussian, GMM and SVM) reveal that the SVM classifier achieves the best classification performance for the language classification task. The classification performance on channels with the best accuracies is reported in this chapter.



(a) English stimulus

(b) Japanese stimulus

Figure 2.3: Top row: Audio Signal; Middle row (x-axis same as top row): EEG Signal (channel F8) of subject 1 during listening task (Average of 3 trials); Bottom row: (i) Spectrum of windowed EEG signal centered at 0.2 s; (ii) Spectrum of windowed EEG signal centered at 0.4 s; and (iii) Spectrum of windowed EEG signal centered at 0.6 s (window duration is 0.4s).

2.3.1 Feature Extraction

A spectrogram is computed using the Short-Time Fourier Transform (STFT) of a signal. The spectrograms are used extensively in the field of speech processing [81], music, sonar [82], seismology [83], and other areas. The spectrogram has been used to analyze the time-frequency characteristics of EEG rhythms [84] and to detect seizures in epileptic patients [85].

In our spectrogram computation as shown in Figure 2.3, we use a hamming window of duration 400ms and step size of 200ms on the input EEG signal.

2.3.2 Trial Averaging

In order to reduce the effect of noise and background neural activity, the EEG data from each trial is averaged with two other random trials of the same stimulus, either in the temporal domain or in the spectral domain. The number of trials averaged is restricted to 3 as it helps to remove noise and at the same time provides enough number of samples to train the classifier.

Trials were initially split into training and test sets, and averaging was performed only within trials from the same set to prevent data leakage.

The EEG data recorded for a fixed 0.8s duration after the onset of the audio stimulus is used for analysis (the duration of all the audio stimuli ranges from 0.5s to 0.82s). The logarithm of the magnitude of the spectrogram computed on the temporal domain average of EEG trials is termed as *Spec(Avg)* feature. In spectral domain averaging (*Avg(Spec)*), the spectrogram is computed for each trial of the stimulus and then averaged. The logarithm of the average of magnitude spectrum along with the average of the cosine of phase of the spectrograms is used as the feature vector (termed as *Avg(Spec+Phase)*).

2.3.3 Results for Language Classification

2.3.3.1 Effect of Temporal Context

The English and Japanese languages have phonological dissimilarities like the difference in the production of /r/ and /l/ sounds as well as the presence of unique phoneme sounds in English and Japanese [86]. However, it can be hypothesized that the language-specific information may not be evident in shorter segments of speech (phoneme or syllable). The poor performance of language identification at the syllabic level (using a single window of 400ms without context) from neural signals confirms this hypothesis. The language variabilities are more pronounced at the interaction between different sounds which is referred to as co-articulation. Hence, incorporating context aids in language identification.

Context padding involves adding additional frames or time steps to the input spectrogram to provide context for the analysis. Spectrograms are often divided into frames, where each frame represents a short segment of the audio signal. To capture contextual information around each frame, context padding involves adding extra frames before and after the original frame of interest. These additional frames help the model to consider information from neighboring frames, allowing it to capture temporal dependencies and patterns in the data. In our imple-

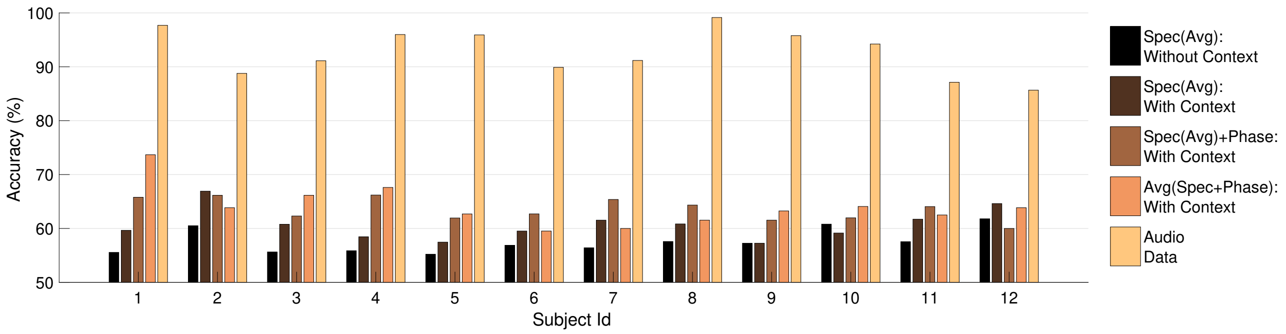


Figure 2.4: Language classification accuracy obtained for the 12 subjects with different feature extraction techniques on EEG data recorded during the listening state. Different feature types are Spec(Avg): Spectrogram of temporal average of trials (Sec. 2.3.2)- with and without context, Spec(Avg)+Phase: Phase information appended to the previous feature (Sec. 2.3.3.2), Avg(Spec+Phase): Average of magnitude and phase of spectrograms of trials. We also compare the performance of language identification from EEG signals to those from the spoken audio data provided by the subjects (Sec. 2.3.4).

mentation, we utilized a context of size 3. This implies including the current frame, along with one frame before and one frame after it, as the feature representation.

Figure 2.4 shows the performance of *Spec(Avg)* features with SVM classifier with and without context padding. The feature extraction with the context of size 3 provides better accuracy than using the features extracted from single window of EEG signal (of duration 0.4s). The features with context that provided the best accuracy in Figure 2.4 are also shown to perform the best in the classification of Hindi versus Japanese shown in Figure 2.5.

2.3.3.2 Effect of Phase Information on Language Recognition

Given the long duration of the spectrogram window, we hypothesize that the phase of the spectrum in the 400ms windows is also a useful feature for classification. We concatenate the cosine of the phase to the magnitude of the spectrogram feature for each frame of the input signal and use it as a feature vector using temporal domain averaging (*Spec(Avg)+Phase*) or using spectral domain averaging (*Avg(Spec+Phase)*). Our experiments indicate that the phase adds meaningful information to the feature regarding the familiarity of the language as shown in Figure 2.4. We can observe that adding the phase information provides better

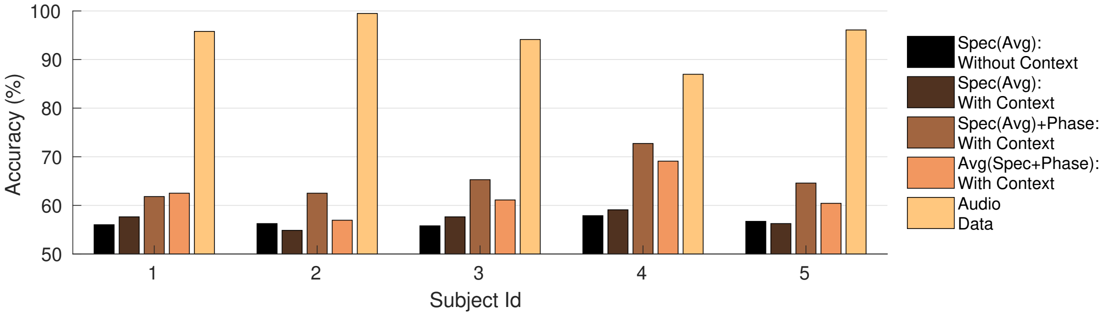


Figure 2.5: Hindi vs Japanese Language classification accuracy obtained for the 5 subjects with different feature extraction techniques on EEG data recorded during the listening state. Different feature types are Spec(Avg): Spectrogram of temporal average of trials (Sec. 2.3.2)- with and without context, Spec(Avg)+Phase: Phase information appended to the previous feature (Sec. 2.3.3.2), Avg(Spec+Phase): Average of magnitude and phase of spectrograms of trials. We also compare the performance of language identification from EEG signals to those from the spoken audio data provided by the subjects (Sec. 2.3.4).

language classification accuracy than using the magnitude of spectrogram alone, for most of the subjects. This observation is also confirmed with the experiments reported on Hindi versus Japanese (second experiment) reported in Figure 2.5.

As seen in Figure 2.4, all subjects achieve language classification accuracy above 59.5% for *Avg(Spec+Phase)* features. Subject 1 attains the highest classification accuracy (73.68%). The average language classification (across subjects) obtained by *Avg(Spec+Phase)* is approximately 64% which is significantly better than the chance level. The t-test conducted at a significance level of 0.05 obtained a p-value less than 10^{-5}. This suggests that significant cues exist in the listening state EEG regarding the language identity of the stimuli. In the Hindi vs Japanese language classification, subject 4 attains the highest classification accuracy (72.73%). The classification performance for the rest of the subjects are also above 60% with phase information added to the feature vector.

2.3.3.3 Performance of Different Classifiers

As shown previously, the spectrogram magnitude information is meaningful along with the phase information.

a. Performance of Different Classifiers					
Model Type	**Discriminative**		**Generative**		
Classifier	SVM	LDA	Gaussian	GMM 2 mix.	GMM 4 mix.
`I. English vs Japanese Classification`					
Average Accuracy (%)	**64.06**	62.79	58.64	60.46	59.99
`II. Hindi vs Japanese Classification`					
Average Accuracy (%)	62.57	52.18	**65.09**	62.19	59.86

b. Language Classification in EEG Spectral Bands						
Spectral Band	δ (0.1-4Hz)	θ (4-8Hz)	α (8-13Hz)	β (13-30Hz)	γ (30-50Hz)	ALL (0.1-50Hz)
`I. English vs Japanese Classification`						
Average Accuracy (%)	62.52	61.54	**64.21**	63.19	63.06	62.83
`II. Hindi vs Japanese Classification`						
Average Accuracy (%)	61.55	**63.71**	61.33	62.44	62.44	62.73

Table 2.2: (a) Performance of different classifiers with Avg(Spec+Phase) features (Spectral Band: 0.1-30Hz). (b) Classification accuracy of SVM classifier with Avg(Spec+Phase) features in different spectral bands.

The performance of different classifiers for the *Avg(Spec+Phase)* features in terms of average accuracy is shown in table 2.2 (a). It is seen that the SVM provides the best performance among them ($p < 10^{-4}$). The Gaussian mixture model (GMM) with two mixtures performs better than a single Gaussian model or a GMM model with 4 mixtures. The input to all four classifiers other than SVM is standardized to zero mean and unit variance along each dimension. For the LDA-based classifier, we use the mean of the two classes of training data as the threshold. The statistical significance of the difference in the performance of classifier models has been evaluated using paired sample t-test with a significance level of 0.05 (with p $< 10^{-4}$). In this statistical test, the SVM classifier is found to be significantly better than the rest. In the second subsequent experiment on classifying Hindi and Japanese words, the Gaussian classifier provides the best performance ($p < 10^{-4}$). The Gaussian classifier, being a simpler classifier, shows better performance in classifying Hindi (L1) versus Japanese while the SVM classifier performs better for the relatively harder task of classifying English (L2) versus Japanese. Also, the data for Hindi-Japanese experiments came from only 5 subjects compared to the data from 12 subjects used in English-Japanese experiments.

2.3.3.4 Language Classification in Different Spectral Bands of EEG

The accuracy of the language identification task varies depending on the different spectral bands of the EEG signal. The analysis indicates that α and β bands capture more language discriminative information as compared to θ and γ bands (Table 2.2 (b)). We obtain the highest classification accuracy of 64.21% in the α band. In the classification experiment involving Hindi versus Japanese, the θ band provides the best performance. This indicates that the language discriminative information is spectrally selective and the dominant language information is present in α and θ bands. The best-performing sub-band rhythms have a statistically significant difference in performance compared to the next best one (with $p < .005$).

2.3.4 Comparison of Language Classification in Spoken Audio and EEG

We also perform the language classification experiment on the behavioural signals (spoken audio) from the subjects. We use the Mel Frequency Cepstral Coefficients (MFCC) [87] as the features for this experiment. The MFCC features with a context size of 53 (800ms) are concatenated and a linear discriminant analysis (LDA) is performed at the word level to reduce the dimension of these features to 23. With these features and the SVM classifier, we obtain an average accuracy of 93% (for both experiments). The comparison of results between audio and EEG shows that, while the spoken audio contains significant information for language classification, the EEG signals at the listening phase can also provide language discriminative cues which are statistically significant.

2.4 Language Learning and EEG

In the rest of the analysis provided, we only use the data collected from the first experiment involving English and Japanese words.

2.4.1 Evidence of Language Learning

In this section, we attempt to establish the evidence for Japanese language learning using behavioural data (spoken audio signals). The aspect of language learning may cover many facets like memory, recall, semantics and pronunciation etc. In this study, we limit the scope of language learning to improvement in pronunciation of the spoken audio. We use an automatic pronunciation scoring setup as well as human expert evaluation for this purpose.

2.4.1.1 Automatic Pronunciation Scoring

The automatic rating of speech pronunciation has been a problem of interest for many analysis tasks as well as for applications like computer-assisted language learning (CALL) [88]. Several methods have been proposed for automatic pronunciation rating based on stress placement in a word [89, 90], learning-to-rank approaches [91] etc. In this study, we use a modified version of log-likelihood based pronunciation scoring with the force-alignment of hidden Markov models(HMM) [92].

A HMM-based speech recognition system is trained using the Corpus of Spontaneous Japanese(CSJ) [93]. A Hybrid HMM-Deep Neural Network (DNN) model is developed using the Kaldi toolkit [94]. For the given Japanese word used in our EEG experiments, the word level HMM is formed by the concatenation of the phoneme HMMs that correspond to the phonetic spelling of the word (obtained from the dictionary of the CSJ corpus). Using the word level HMM (denoted as λ), the likelihood of the speech data $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$ is approximated as [95],

$$P(\mathbf{O}|\lambda) = \sum_{\mathbf{Q}} P(\mathbf{O}, \mathbf{Q}|\lambda) \approx \max_{Q} P(\mathbf{O}, \mathbf{Q}|\lambda) \tag{2.1}$$

where $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_T\}$ denotes the state-sequence of the HMM and T denotes the time duration. The above likelihood can be efficiently solved using the Viterbi algorithm [95]. In this work, the log-likelihood of the behavioural data (spoken audio from the subjects) and the stimuli audio are computed with force alignment and are used as confidence estimates of pronunciation. The main modification of our approach compared to the previous work in [92] is the use of state-of-art speech acoustic modelling using deep neural networks.

2.4.1.2 Pronunciation Scoring by Human Expert

We also evaluate the pronunciations using a human expert[1] based pronunciation rating (for Japanese audio). Given the large number of spoken audio recordings (20 recordings per subject per word), we use a smaller subset of this audio (4 recordings per subject per word from the 1st, 6th, 11th and 16th trials) for evaluation from the human expert in the Japanese language. This was done on a scale of 1-10 (where 1 indicates poor pronunciation and 10 indicates a native speaker's pronunciation). In this evaluation, the human expert was also provided with the stimuli (in a hidden randomized manner similar to the hidden reference in audio quality testing [96]) in order to ensure the effectiveness of the rating. Out of the 12 Japanese words, the 3 words for which the stimuli recording had a pronunciation rating of less than 8 were excluded from further analysis.

2.4.1.3 Improvement of Pronunciation over Trials

In Figure 2.6, we compare the evaluations from the human expert for Japanese language recordings along with the automatic pronunciation scores. For this plot, the logarithm of likelihood scores are normalized and are linearly mapped to the range of $1 - 10$ in order to make the comparison with the human scores. The average rating of all the spoken audio data from the subjects (12 subjects) is plotted for two phases separately - Phase-I ($1 - 10$ trials) and Phase II ($11 - 20$ trials). The stimuli ratings are also recorded for both human experts and automatic

[1] The human expert used in our study was a professional Japanese language tutor. The text used in the stimuli was provided before the pronunciation evaluation.

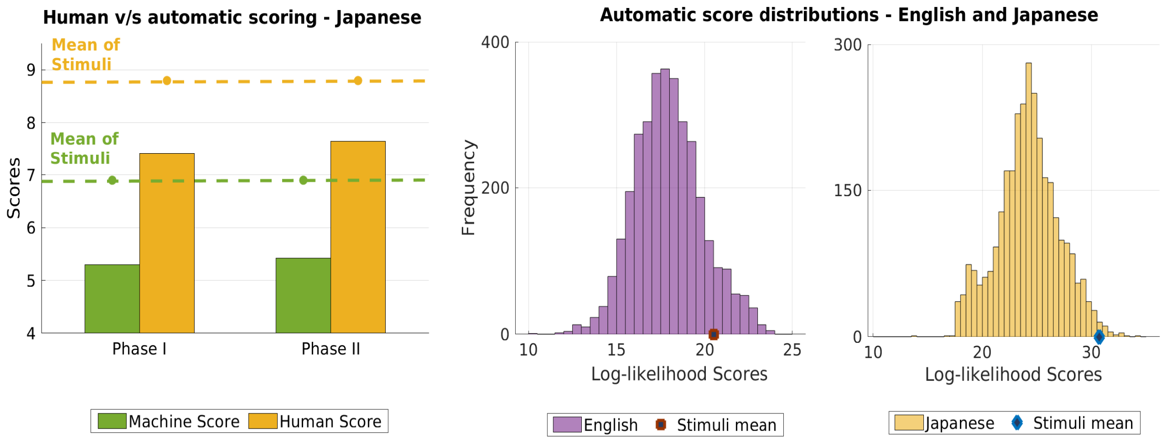


Figure 2.6: The left panel depicts the comparison of human and machine pronunciation scoring for Japanese language audio data. The right panel depicts the histogram of log-likelihood scores (raw machine scores) for English and Japanese spoken audio data. In both cases, mean of the stimuli is also highlighted for reference.

ratings.

As seen in Figure 2.6 (left panel), both the human scoring and the automatic scoring indicate an improvement in the pronunciation of the Japanese words for Phase II over Phase-I. At the subject level, we also find that 10 out of 12 subjects showed an increase in scores (both human expert and automatic method) for Phase-II over Phase-I. Also, using the approach of log-likelihood with forced alignment shows a good match with the human expert-based scoring. We also find the score improvement to be statistically significant for the human expert scoring and the machine scoring (with $p = 0.027, 0.017$ respectively).

In both cases, the mean of the log-likelihood scores for the stimuli is different from the mean of the spoken audio recordings from the subjects. This is expected as the stimuli are clean speech utterances which were recorded in a close-talking microphone setting while the spoken audio recordings from the participants were collected in a far-field microphone setting. However, in the case of English-spoken audio recordings, the mean of the log-likelihood scores for the stimuli is more similar to the rest of the distribution compared to the Japanese language (the percentage of data below the mean value of the stimuli is 70 % in the case of English while is 95 % in the case of Japanese). This difference between the two languages is also statistically significant.

2.4.2 Understanding Language Learning via the EEG

In this section, we use two types of analyses, (i) based on inter-trial distances and (ii) based on distance between audio and EEG envelopes.

2.4.2.1 Inter-trial Distance Analysis

We use the inter-trial distance between EEG signals to quantify the change in representation while listening to the same word over time. The hypothesis here is that in the case of a known language like English, the inter-trial distance is somewhat random (due to the measurement noise in EEG) but small in value throughout. However, in the case of the Japanese, the inter-trial distances may show a pattern of reduction over trials as a consistent representation is formed in the brain.

For testing this hypothesis, the EEG signals recorded during each trial are converted into a log magnitude spectrogram (window length of 100 ms and shifted by 50 ms). The magnitude spectrogram of each channel is converted into a single long vector and a pairwise distance between trials is computed using Euclidean distance between spectrogram vectors. An inter-trial distance matrix of size 20×20 is computed for each channel separately. This is a symmetrical matrix whose elements contain the inter-trial distances between any pair of trials. An example of inter-trial distance in EEG is shown in Figure 2.7.

In order to further analyze the inter-trial distances, the trials are broken down into two phases as before - Phase-I (trials $1 - 10$) and Phase-II (trials $11 - 20$). The mean of the inter-trial distances in Phase-I (denoted as d_1) and Phase-II (denoted as d_2) are calculated. The difference $d_1 - d_2$ is indicative of a change in inter-trial distances over the course of 20 trials. We compare $d_1 - d_2$ averaged over all words and all subjects.

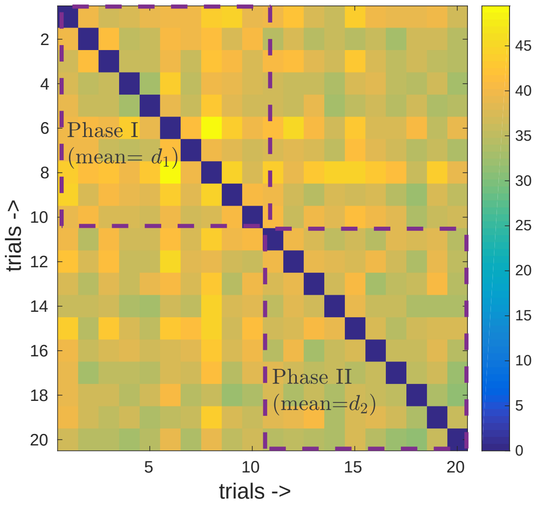


Figure 2.7: An example of distance matrix plotted for a particular word,subject and channel

As hypothesized, the inter-trial distances reduce over time in the case of Japanese but remain more or less uniform in the case of English as seen in (Figure 2.8). The histograms depicting the difference values ($d_1 - d_2$) for all the channels and the stimuli are plotted separately using a Gaussian fit for the Japanese language and English language. In order to confirm the statistical significance, we performed a two-tailed test with the null hypothesis being that the values of $d_1 - d_2$ for both English and Japanese come from the same distribution and the alternative hypothesis being that Japanese measurements of $d_1 - d_2$ come from a different distribution compared to English. The tolerance level alpha was set to 0.05. It is seen the distributions of the difference values for English and Japanese are statistically different.

The brain regions that show the language differences the most are shown in the scalp-plot of the difference in terms of ($d_1 - d_2$) (for English and Japanese separately for each channel averaged over all the subjects) in Figure 2.8. A plot which differentiates the two language level scalp plots is also shown here. The regions that show more changes in English stimuli are in the temporal region while the frontal regions also show this effect in the case of Japanese stimuli. The regions that have higher differences between the two languages are predominately in the frontal brain regions.

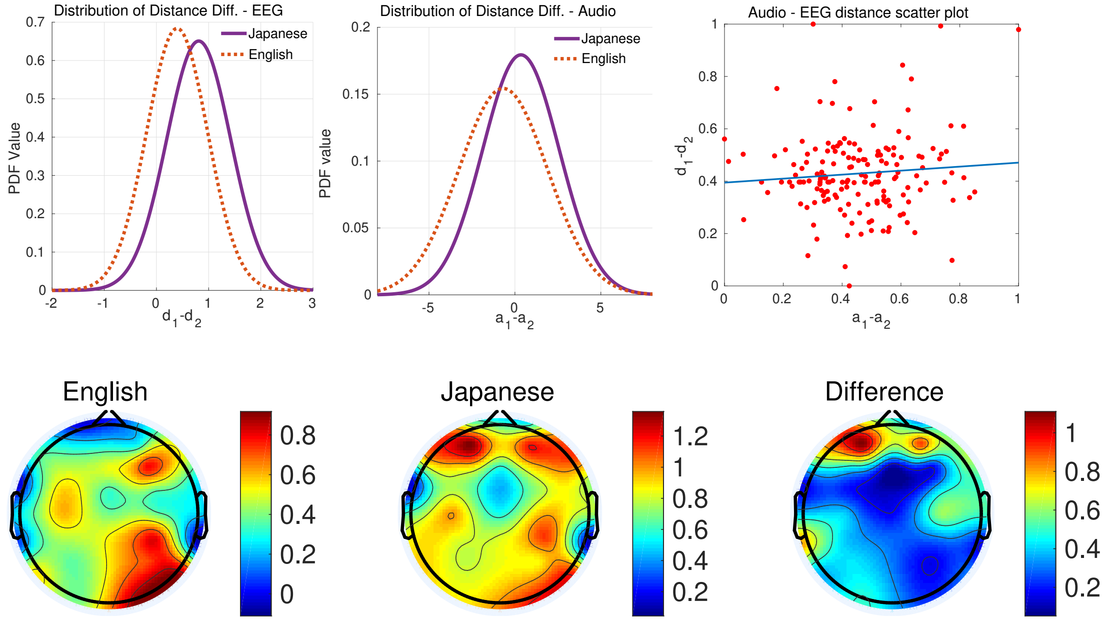


Figure 2.8: (above) Histograms plotted using a Gaussian fit depicting difference between the mean inter-trial distance in the Phase-I and Phase-II ($d_1 - d_2$) for EEG signals.A two-sample t-test is performed between the distribution of English and Japanese in the case of both EEG and audio. It is observed that in both the cases the distributions are statistically significant($\alpha = 0.05$).On the right, correlation between audio and EEG inter-trial distance differences for Japanese trials is shown (EEG data from electrode site FC4 is plotted here). (below) Scalp plots indicating the channels with higher $d_1 - d_2$ difference for English, Japanese and the difference of the two languages.

An extension of this analysis performed for the spoken audio data is done using the audio recorded during the speaking phase of each trial. The silence portion of each recorded audio is removed. Each audio signal is converted into a sequence of MFCC feature vectors. Similar to the analysis done in the previous section, a symmetrical distance matrix of size 20 × 20 is computed for each word. Since the duration of the spoken audio for the same word differs each time, a Euclidean metric-based Dynamic Time Warping (DTW) distance is calculated for the pair-wise trial distance. Similar to the EEG analysis, the trials are divided into Phase I and Phase II. The mean inter-trial distance in Phase-I (denoted as a_1) and Phase II (denoted as a_2) are calculated. The difference a_1- a_2 is computed similarly to EEG and the histogram of the difference in the case of audio for Japanese and English (using the spoken audio data from all subjects) is plotted using a Gaussian fit (shown in the middle of the top panel of Figure 2.8). As seen in the case of EEG, the difference (a_1- a_2) in the mean distance between the two phases is greater in the case of Japanese than in English. The distribution obtained in the case of Japanese has a mean that is significantly larger than zero but not for English. Similar to the case of EEG signals, a two-tailed t-test was performed on the Gaussian fit of English and Japanese (alpha=0.05) and the two distributions were found to be statistically different.

We also analyze the correlation between EEG recorded during the listening state and the spoken audio in terms of the mean inter-trial difference in Phase-I and Phase-II (i.e. the correlation between (d_1- d_2) and (a_1- a_2)). A scatter plot is shown with the difference values for EEG signals along the y-axis and the corresponding difference for audio signals along the x-axis (i.e. (d_1- d_2) versus (a_1- a_2)). An example of the scatter plot difference of (d_1- d_2) versus (a_1- a_2) for the frontal EEG channel (FC4) is shown in Figure 2.8. Each point on the plot indicates a (subject, word) pair. The values along both axes are normalized between 0 and 1. A line of best fit is plotted through the points. The slope of the line (denoted by m) of best fit is positive for most of the channels. Since the scales for the EEG spectrogram and the audio MFCC features are different, the amount of correlation between the listening state EEG and the audio spoken may be unnormalized. Additionally, the mean slope of best-fit lines for Japanese words is found to be higher than in English. These observations indicate that the pattern formation seen in the behavioural data is also correlated with the patterns seen in EEG recordings.

2.4.2.2 Distance between EEG and Audio Envelopes

A direct relationship between the EEG signals recorded during the listening and the audio spoken by the subject during the speaking phase may also present useful insights. Previous studies have attempted to predict the audio envelope using EEG [97] or to perform a correlation analysis between the EEG and audio envelope [98]. In our study, we try to align the EEG and audio envelopes (after down-sampling to the same rate) using dynamic time warping

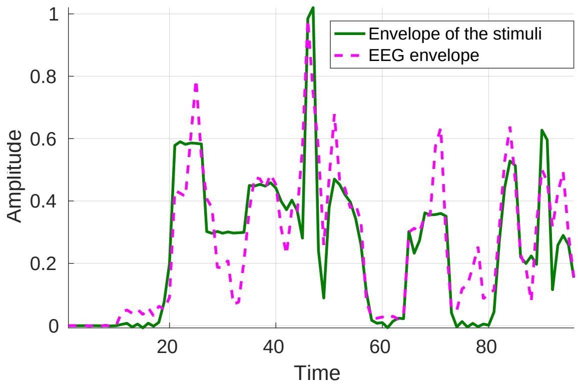


Figure 2.9: Example of aligned EEG signal and spoken audio envelope

technique (DTW) and measure the distance between the two. The distance measure is inversely proportional to the correlation measures used in the past, as smaller distances between audio and EEG envelopes are associated with higher correlations and vice-versa. The choice of a distance measure is to maintain consistency with the previous analysis based on distances.

DTW is a time series analysis technique for comparing and aligning sequences with variable speeds or timing differences, particularly for temporal data with phase shifts or time lags. It calculates an optimal alignment between two time series by warping the time axis, allowing for non-linear distortions [99]. This makes it effective for comparing sequences that might have similar shapes but are temporally out of sync. A sample plot of EEG and audio envelopes that are time aligned using DTW is shown in Figure 2.9.

For the distance computation, the silence portion of the audio is removed and the length of the EEG signals is kept to the stimuli length plus 100ms. Both the signals are converted to their corresponding Hilbert envelopes and the envelopes are down-sampled to 64Hz. The DTW distance between the two aligned envelopes is calculated. It is seen that the mean distance between the two envelopes is greater in the case of Japanese than in English (Figure 2.10).

As a follow-up to the comparison done between the envelope of the EEG signals and the audio spoken, a similar analysis is done between the envelope of the EEG signals and the stimuli presented to the subject. A DTW distance is computed between the envelope of listening EEG

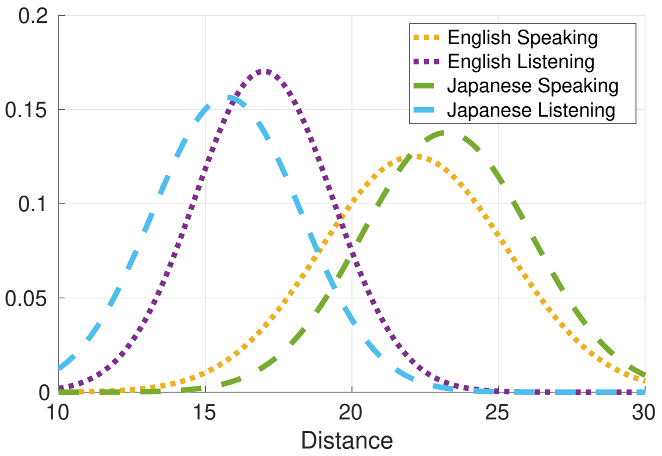


Figure 2.10: Probability Distribution Function of distances between the envelope of listening state EEG and envelope of stimuli presented and the envelope of spoken audio. A two-sample t-test is performed, and it is seen that the distributions of the two languages are statistically different for each state.($\alpha = 0.05$)

and the envelope of stimuli. A histogram of all the distances (between listening stimuli and EEG as well as those between the spoken audio and EEG) for both languages are shown in Figure 2.10. The average distance values between envelopes (for the listening state) are less in the case of Japanese compared to the speaking state. The distance between the envelopes of the EEG signal and the spoken audio is more than the distance between the envelopes of the EEG and the stimuli presented as well.

A two-tail t-test was performed on the distance distributions between the EEG envelopes and audio for English and Japanese. This was done for both distance measures between EEG and stimuli envelope as well as EEG and spoken audio envelope. In both cases, the null hypothesis was that the distributions of English and Japanese are not statistically different, and the alternative hypothesis was that the two distributions are statistically different. The t-test indicated a statistically significant deviation from the null hypothesis in both cases. This supports our claim that the distributions obtained for the relationship between the envelopes of EEG and audio are statistically different for the two languages.

2.5 Discussion

With spectrogram based features and SVM classifier, we achieve an average accuracy of 64% in distinguishing between known and unknown languages in a language classification task. The comparison of outcomes between audio and EEG indicates that, while the spoken audio carries substantial information for language classification, the EEG responses during the listening state can also offer statistically significant cues for discriminating between languages.

Our experiments demonstrate that incorporating the phase imparts valuable information to the feature related to language familiarity, as illustrated in Figure 2.4. It is evident that integrating phase information yields improved accuracy in language classification compared to relying solely on the magnitude of the spectrogram. Section 2.3.3.4 suggests that the information distinguishing languages is selective to specific spectral ranges, with the primary language cues being found in the α and θ bands. As seen in Figure 2.8, the inter-trial distances reduce over time in the case of Japanese but remain more or less uniform in the case of English. The subjects' familiarity with the words from the English language may have resulted in generating invariant EEG responses when presented with these stimuli. In the case of Japanese stimuli, subjects are listening to those words for the first time. Over the trials, subjects form a consistent neural representation of the unfamiliar stimulus. It is evident from the reduction of inter-trial distances of the EEG responses.

The stimulus presentation and listening state EEG recording happen in parallel. Hence, a higher correlation is expected between the two compared to the correlation between the envelopes of the listening state EEG and the spoken audio. This is seen in Figure 2.10. Since Japanese is unfamiliar, the spoken audio is not well aligned with the stimuli. Hence, the distance between spoken audio and EEG envelopes may be higher for Japanese than for English.

All the subjects who participated in our recordings were not exposed to Japanese before but had a good proficiency in English. We hypothesize that due to their unfamiliarity with Japanese, their attention while listening to Japanese stimuli is much more than in English, resulting in a lesser distance between envelopes of EEG and Japanese stimuli compared to English. The absence of semantic processing in Japanese could also explain the reduced distance between the stimuli envelope and the EEG envelope for Japanese. In the speaking state, the subjects tend to reproduce audio that is less correlated with stimuli for the Japanese language than the English language. This may explain the rightward shift of the distribution of distances for Japanese spoken audio in Figure 2.10.

We have additionally performed analysis to find out the channels that capture language learning the most. The channels are identified as the ones that show the maximum difference between Phase I and Phase II. The top five channels are found to be ($O2, AF3, F8, AF4, F7$), located primarily in the frontal region of the brain.

2.6 Chapter Summary

The key findings from this chapter are the following,

- A consistent neural representation is formed when exposed repeatedly to words from an unfamiliar language. This is also consistent with language learning established using pronunciation rating.

- In the listening state, the correlation between audio stimuli and EEG envelope is higher for Japanese than English trials (smaller distance values). The correlation between the EEG envelope of the listening state and the envelope of the spoken audio is less for Japanese than for English.

- The discriminative signatures of the language are encoded in the time-frequency representation of the EEG signals in the range of 0-30Hz, both in magnitude and phase.

In the current setup, unfamiliar words are presented to the subjects without the semantic meaning or the context of the word. In the next chapter (chapter 3), we investigate how the neural responses change when unfamiliar words are provided with semantics. Additionally, longer content is expected to provide future insight into language-level differences compared to word-level analysis. This can be achieved with stimuli containing longer words, phrases and sentences. In the next chapter (chapter 3), the EEG recording experiment introduces a scoring model that rates and gives feedback to the subject depending on how well they pronounce the words during the experiment, based on the pronunciation model introduced in this chapter.

Chapter 3

ERP Evidences of Rapid Semantic Learning In Foreign Language Word Acquisition

3.1 Introduction

Having studied neural encoding in response to repeated exposure to foreign language word sounds without context, our research now investigates the impact of semantics on learning patterns. The association of semantics with speech sounds constitutes the next phase of word learning ([100]), which further develops into sentence formation and syntax/grammar learning. These processes may not be sequential and may be interleaved with each other. This current exploration has the potential to reveal the neural signatures associated with rapid semantic learning. We aim to examine how neural responses evolve as newly acquired words are integrated into sentence contexts, shedding light on the cognitive processes involved in comprehending sentences.

The brain activity related to the perception and cognition of language can be studied through event-related potentials (ERP). The ERPs are computed by averaging the electroencephalogram

(EEG) recordings evoked by the same event. The ERPs triggered by verbal stimuli have been associated with different aspects of language learning ([101]).

In this study, we introduced the semantics of Japanese words to subjects without prior exposure to the Japanese language. Following this language learning task, we performed event-related potential (ERP) analysis using semantically matched and mismatched English sentences where the end words were replaced with their Japanese counterparts. The event-related potential (ERP) of electroencephalography (EEG) signals has been well studied in the case of native language speech comprehension using semantically matched and mismatched end-words. The presence of semantic incongruity in the audio stimulus elicits an N400 component in the ERP waveform. However, it is unclear whether the semantic dissimilarity effects in ERP also appear for foreign language words that were learned in a rapid language learning task. The ERP analysis revealed that, even with a short learning cycle, the semantically matched and mismatched end words elicited different EEG patterns (similar to the native language case). However, the patterns seen for the newly learnt word stimuli showed the presence of the P600 component (delayed and opposite in polarity to those seen in the known language). A topographical analysis revealed that P600 responses were predominantly observed in the parietal region and the left hemisphere. The absence of the N400 component in this rapid learning task can be considered evidence for its association with long-term memory processing. Further, before semantic learning, the ERP waveform for the Japanese end words showed a P300 component owing to the subject's reaction to a novel stimulus. These differences were more pronounced in the centro-parietal scalp electrodes.

In this chapter, we present a study to analyze rapid language learning effects using ERP analysis where the end words of English sentences were replaced with Japanese words. In the first phase of the experiment, the subjects who were proficient in English with no prior exposure to Japanese words were presented with English sentences containing Japanese end words. A subsequent language learning phase introduces the semantics of the Japanese stimuli through its pictorial description. In the final phase, the subjects listen to English sentences again with Japanese end words armed with the knowledge of semantics. The Japanese end words in English sentences may be congruent or incongruent with the sentence context. An ERP analysis of Japanese end words before and after semantic learning illustrates significant changes that highlight the neural processes involved in learning. Further, the difference in ERP of Japanese stimuli in congruent and incongruent conditions (after language learning) elicits delayed and positive ERP components as opposed to the N400 effects observed in the native language under the semantic mismatch condition.

Contributions:

- The study contrasts the ERP effects of semantic incongruity in a proficient language (English) versus a newly acquired language (Japanese).
- This study demonstrates that rapid semantic learning elicited ERP responses in the form of a delayed positive response at 500-700ms from the onset of the end word for the incongruent words.
- The scalp electrodes showing semantic effects from newly learned words of a foreign language (Japanese) are located in the parietal and occipital regions.
- The amplitude of semantic incongruity (ERP component) in the foreign language (Japanese) is more for pure foreign words (hiragana words in Japanese) versus English loan words (katakana words in Japanese).

Relevant Literature

The N400 component was first introduced by [102], where the reading task comprised presenting the participant with a set of sentences that end with a congruent or incongruent word. These semantically incongruent end words in a sentence elicited specific type of ERPs ([103, 104, 105]), known as the N400, a negative-going deflection between 250 and 400 ms after the end word onset. The presence of N400 in semantically incongruent stimuli was also observed in other stimuli presentations ([106]) like reading ([107]) and visual forms ([108]). The N400 was characterized as a reaction to an unexpected or inappropriate, but syntactically correct word at the end of a sentence. The N400 component was not observed for stimuli with syntactical and grammatical errors ([109]). This result is evidence that the N400 wave is more closely related to the semantics than the syntactical processing.

The N400 is not only a response to semantic improbability or anomaly but also as an indicator of the access to semantic information associated with the stimuli ([110]). When a word is congruent in its context, there is little new information to process, and hence, this evokes a lower N400 response than an incongruent word. The amplitude of the N400 is sensitive to a word' s semantic expectancy ([111]) and found to be larger in response to more unexpected stimuli ([112, 113, 114]). The N400 has also been used to show that language comprehension is incremental ([115]), and involves prediction ([116]). Further, semantic information processing happens even without active awareness ([117]). It has also been pointed out that language mechanisms vary across the hemispheres ([118]) and can change over the course of normal aging ([119]).

Even though the N400 has contributed significantly to the understanding of language comprehension, the N400 response is not confined to the language domain alone; hence, it is not a "language component" ([120]). The N400 is not only seen in word comprehension but also for different kinds of pictorial stimuli (eg, comics/cartoons, drawings, pictures of objects, natural scenes), faces, gestures, and environmental sounds. Thus, it can be elicited for any kind of stimulus linked to long-term memory representations ([110]).

3.1.1 N400 as an index of word learning

The N400 has been established in numerous studies to be a useful index of new word learning. In a study by [121], adults were taught the meanings of infrequent and unfamiliar words. The N400 component was seen for unrelated word pairings containing the trained words and not for those involving the unfamiliar words.

[122] investigated context-based learning of novel words using ERPs. The researchers specifically introduced novel word forms in the ending position of meaningful sentences that the participants read during the training phase. In a following relatedness judgement task on word pairs, consisting of a trained novel word (prime) and a real word (target), the study found a reduction in the N400 for targets words that were associated with the prime word compared to the unrelated target-prime pairs.

[123] investigated contextual learning by embedding pseudo-words into meaningful short story contexts. In a subsequent relatedness judgement task, the study showed a reduction in the N400 amplitude for targets corresponding to the novel word.

The N400 has also been demonstrated to be sensitive to new word learning after just one exposure to a novel word in the context of a highly predictive sentence ([124, 125]). The findings of such investigations provide neuro-physiological evidence for understanding the semantic learning of the new words.

3.1.2 Intra-sentential code-switching

The N400, left-lateralized anterior negativity (LAN), and the late positive component (LPC; also referred to as P600) are the three primary ERP components identified in research on intra-sentential code switching.

The LAN is a left-lateralized anterior negativity that occurs in the same time frame as the N400 (300‐500 ms) but with a distinct topographic scalp distribution. [126] observed LAN effects in morphosyntactic processing, as well as in the processing of code-switched sentences. The higher working memory load resulting from integrating morphological signals of the code-switched word into the wider sentence context was interpreted as the switch-related LAN component ([127]).

The LPC (or P600) is a positive-going wave that arises 500‐600 ms after the stimulus and lasts several hundred milliseconds ([128, 129]). It has a wide posterior scalp distribution and is strongest in the centro-parietal areas. [130, 131] and [132] infer that the LPC indexes sentence-level rearrangement or re-analysis. The LPC, according to this view, indicates a sentence-level wrap-up or meaning revision process, which in the instance of intra-sentential code-switching, reflects the sentence-level reorganization of two languages into a cohesive utterance. The LPC has also been linked to the processing of unexpected or unlikely task-relevant events ([133, 134]), as well as the reorganization of stimulus-response mapping ([135]). A switch-related LPC represents bilinguals' perception of a language transition as an unexpected occurrence involving a shift in form rather than meaning ([127]).

3.1.3 P300

P300 is usually elicited using the oddball paradigm, where low-probability target items are interspersed with high-probability non-target (or "standard") items. When captured through electroencephalography (EEG), it manifests as a positive voltage deflection peaking around 300ms [136]. The characteristics of this signal, including its presence, amplitude, spatial distribution, and timing, are commonly employed as measures of cognitive functioning in decision-making processes.

The amplitude of P300 signifies the extent of information processing, along with the allocation of attentional resources to a task and the level of advanced cognitive function [137]. Conversely, an increase in P300 latency indicates suboptimal cognitive performance [138, 139].

In the scientific literature, differentiation is often made in the P3, which is divided according to time of peak amplitude:P3a and P3b. The P3a, or novelty P3, is a positive-going component with peak latency in the range of 250-280 ms. It's topographic distribution shows maximum amplitude over frontal and central electrode sites. P3a has been associated with cognitive tasks of involuntary attention and the processing of novelty ([140]). The P300 (P3b) ERP component is elicited in the process of decision making and usually evoked using the oddball paradigm. It is evoked in connection with a person's reaction to a stimulus and not to the physical traits of a stimulus. P3b is considered an ERP component that reflect cognitive processes involved in stimulus evaluation or categorization [141].

3.2 Materials and Methods

3.2.1 Stimuli

All the speech stimuli used in the study were recorded from speech provided by a single female speaker who was proficient in both English and Japanese languages. The speaker's first language was Tamil, a south-Indian language. The accent of the speaker was Indian English. Given that all the listeners were also of Indian origin, we found that the listeners had no issues understanding the English content. Further, the work employed a single speaker for all the stimuli. This helped to remove the effect of speaker variabilities or speaker switching in the stimuli used. The audio files were recorded in a noise-proof sound booth with a CAD u37 microphone at a sampling rate of 44.1kHz.

The speech stimuli used in the experiment consisted of isolated sentences and isolated words. They were recorded in a soundproof booth. The entire stimuli set used in the experiment is listed in Table 5.1.

The audio and image files used for the experiment are available in this project's GitHub repository[1]. The stimuli set consisted of 90 unique English sentences. The sentences were selected such that the end word was highly predictable from the sentence context. Our stimuli set was taken from the high cloze probable sentences (cloze probability in the range of 67% to 100%) of the Block-Baldwin sentence set ([142]). The audio duration of the sentence varied between 1.4s to 2.4s, with an average of 2.2s. The duration of the end word in these sentences in different stimuli conditions are given in Table 3.1.

The end words of the sentences were either from English or Japanese language and they were

[1] https://github.com/iiscleap/Semantics-EEG-ERPStudy

Stimuli Condition	**Duration of End-word**
	[min. - max.] (in s)
English Congruent	[0.2 - 1.1]
English Incongruent	[0.3 - 1.3]
Japanese Congruent	[0.4 - 1.3]
Japanese Incongruent	[0.4 - 1.3]

Table 3.1: The table shows the range of duration of end word of sentences in different stimuli conditions. Note: We used the same set of words for different conditions for a language. But the duration can vary slightly as it is spoken as part of a sentence in different instances.

Table 3.2: Different conditions of the stimuli sentences used in our experiment with an example.

Stimuli Condition	**Example**
C1 - Eng. Congruent	*The cow gave birth to a* ***calf****.*
C2 - Eng. Incongruent	*The cow gave birth to a* ***book****.*
C3 - Jap. Congruent	*The cow gave birth to a* ***koushi****.*
C4 - Jap. Incongruent	*The cow gave birth to a* ***hon****.*

designed to be either semantically congruent or incongruent with the preceding context in the sentence. All the stimuli conditions are listed in Table 3.2. The first condition (C1) consists of 90 English sentences from the original Block-Baldwin set without any modification. The stimuli for the other three conditions of the experiment were created by replacing the end word with all the preceding words intact. In condition 2 (C2), the end word was replaced by an English word which is unexpected in the sentence context (English incongruent condition); in condition 3 (C3), the end word was replaced by a Japanese word of the congruent semantics (Japanese - congruent condition); and in condition 4 (C4), the end word was replaced by a Japanese word of unexpected meaning (Japanese - incongruent condition). Thus, the experiment had a total of 360 sentences. The sentences were carefully chosen such that the end word can be visualized as an image. Each of the stimuli conditions used the same base set of English sentences. The conditions differed only in terms of the end word of the sentences. Each stimulus was recorded as a full sentence for each condition separately. Each stimulus (EC,EINC, JC (before and after learning), and JINC) condition has 90 sentences for trial averaging to compute the ERP signal.

We choose the Japanese language as the novel language as it does not belong to the same language family as English. 90 Japanese words were used to form sentences in different conditions. At the same time, the Japanese language contains a set of loan words from English termed katakana words. The katakana words are typically English words that have been adapted without translation into the Japanese language ([143]). The katakana words sound similar to their English counterparts (cognate words). By using a mix of non-cognate native Japanese words (referred to as hiragana words) and katakana words, we were able to study the effect of phonetic similarity in learning. The set of Japanese words used in our experiments (Table 5.1) consisted of 38 katakana words and 52 hiragana Japanese words. This unbalanced distribution of katakana and hiragana words is in compliance with the word frequency distribution in the Japanese language ([144]). Thus, for stimuli conditions C3 and C4, the end word can be either a katakana word or a hiragana word.

3.2.2 Experiment Flow

A schematic illustration of the experiment is shown in Figure 3.1. A more detailed illustration is given in the appendix. A short video depicting the experiment flow is available in this project's GitHub repository[1]. The end words were either congruent or incongruent with the context of the sentence. The participants learned the semantics of the unknown Japanese words using image supervision (shown in session 2 of Figure 3.1) and the learning was analyzed when the words were used in sentences both in congruent and incongruent conditions (shown in session S3 of Figure 3.1). In each session, the sentences of different conditions were presented in random order using a loudspeaker.

In session S1, the subject listened to English sentences with congruent or incongruent end words (C1 and C2 from Table 3.2). This session also contained English sentences with Japanese end words. The subject got the first exposure to these Japanese words in this session without the semantic information.

[1] https://github.com/iiscleap/Semantics-EEG-ERPStudy

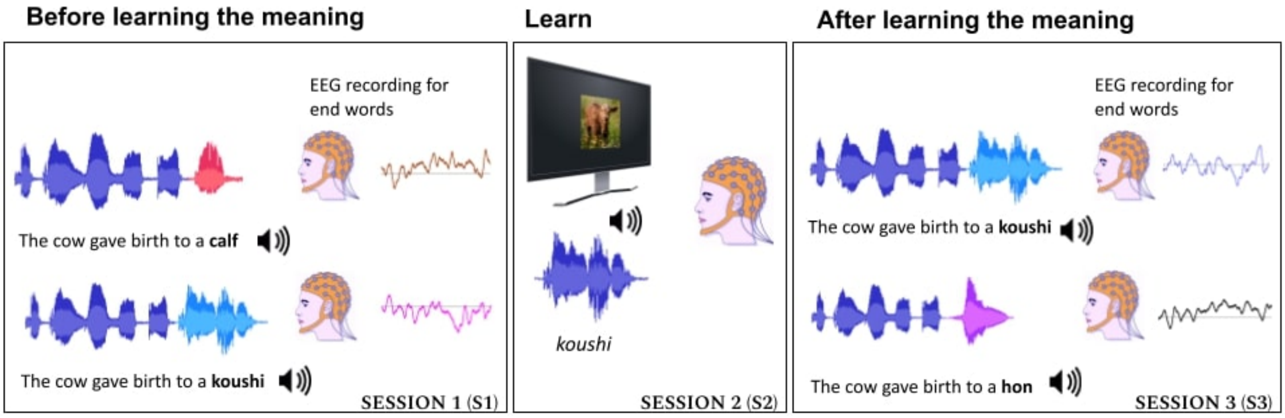

Before learning the meaning
Learn
After learning the meaning
EEG recording for end words
The cow gave birth to a **calf**
The cow gave birth to a **koushi**
SESSION 1 (S1)
koushi
SESSION 2 (S2)
EEG recording for end words
The cow gave birth to a **koushi**
The cow gave birth to a **hon**
SESSION 3 (S3)

Figure 3.1: Experiment pipeline consisted of three sessions. S1 - where the subject listened to acoustics of Japanese words used in English context and English end-words in congruent and incongruent context. S2 - where the semantics of Japanese words were introduced using images. S3 - where the subject listened to Japanese words in English context (both congruent sentences and incongruent sentences).

In the next session (S2), the participants were provided with the semantics of the Japanese words. A block of 5 new Japanese words was considered, and word meanings were conveyed using the respective image. The image form of the word and its audio are presented simultaneously. The 5 words of each block were presented in random order. A retrieval task was also designed to ensure the learning ability of the subject. In the retrieval task, the subject was asked to speak the Japanese word for the image shown on the computer screen. After the subject provided the spoken response, the audio of the correct Japanese word was replayed. Here, the subject has affirmed the learning or corrected their learning if the word recollection was inaccurate. We do not analyze the EEG data from S2 in this study.

In session 3 (S3), the subject listened to the sentence audio played through the loudspeaker. It was an English sentence with a Japanese end word. Here, the end word was either congruent or incongruent to the context of the sentence (C3 and C4 in Table 3.2). The end word was one of the 5 Japanese words learned in the preceding session (S2). Hence, there were a total of 10 sentences (equal number of congruent and incongruent) played in session 3 for the current block. These 10 sentences were presented in random order. After the audio signal was played, a recognition task was carried out to ensure that the subject recollected the meaning of the Japanese end word. In the recognition task, the subject was asked to pick the image corresponding to the Japanese end word. The subjects recorded their choice of image by speaking the corresponding number index on the screen. The subject responses were later evaluated manually to assess their recognition accuracy. The Behavioural results are discussed in Section 3.3.1. Only the EEG responses to the Japanese words, whose meaning was recollected correctly, were used in the subsequent ERP analysis.

In order to avoid the memory load of learning and recalling 90 new Japanese words, S2 and S3 were performed in 18 blocks of 5 words each. Session S3 for a set of 5 words was conducted immediately after the subject was trained on the semantics in session S2. We designed the experiment in this way to reduce the memory load on the subject as we were more interested to analyze the semantic effects of the newly learned words in both congruent and incongruent conditions. A particular Japanese word was presented 5 times to the subject: once in S1 (sentence end word without semantic knowledge), twice in S2 (as isolated words in the learning phase) and twice in S3. In S3, it was used as the sentence end word in congruent and incongruent contexts. A specific sentence-English endword pair was presented only once, while a particular sentence-Japanese endword pair appears twice during the entire experiment. These two occurrences take place in session 1 (before learning the word meaning) and then in session 3 (after learning). For incongruent condition, sentence end word pairing was carried by random shuffling and incongruence was ensured by manual selection. The order of congruent and incongruent conditions in S3 was randomized. Thus, the exposure to new Japanese words was balanced across conditions. The three sessions were recorded in an interleaved fashion in one recording setup for each of the subjects. The experiment design ensured that the subject does not get exposed to katakana words more than hiragana words.

3.2.3 Participants

The participants had self-reported normal hearing and no history of neurological disorders. Twenty-one subjects took part in the experiment. Two subjects were eliminated due to poor EEG data quality and another two were eliminated due to equipment failure. Seventeen adults participated in this study (mean age = 25.7, age span = 22-35, 7 female and 10 male) and they had an intermediate or higher level of English proficiency. This was verified with the Oxford Listening Level Test [1] before the commencement of the experiment. The native language of the subjects was one of the five Indian languages (Malayalam, Tamil, Kannada, Telugu or Hindi). All subjects provided written informed consent to take part in the experiment and received monetary compensation. The Indian Institute of Science Human Ethics Board approved all experiment procedures. The methods were carried out in accordance with the relevant guidelines and regulations.

We have performed the power analysis with an assumed effect size of d=0.5 ([145, 146]) to check the sufficiency of the number of subjects and trials used in the experiments. We performed the power analysis on our experiment using the GPower software ([147]). This analysis revealed that our study is a properly powered experiment(with power value more than 95%). Hence, the effects reported in the study are very likely to be robust and reproducible.

3.2.4 EEG Recording Setup

The EEG signals were recorded employing a BESS F-64 amplifier with 64 passive gel-based Ag/AgCl electrodes placed according to the enhanced 10-20 montage ([27]). It was recorded at a sampling rate of 1024 Hz. An isolated frontal electrode was utilized as ground and the average of two earlobe electrodes was utilized as reference. The channel impedance was kept below 10 kOhm throughout the recording. The EEG recording took place in a sound-proof, electrically isolated room. The software and hardware used during the EEG recording is given

[1] https://www.oxfordonlineenglish.com/english-level-test/listening

in table 3.3.

Table 3.3: Experimental Details

Particulars	Details
Presentation Software	Python based GUI developed by the authors
Monitor	Samsung 27"
Loudspeaker	Dell Ax210 USB Stereo
Microphone	CAD u37

3.2.5 Data Preprocessing

As the initial step, we apply a fourth-order high-pass Butterworth filter with a cutoff frequency of 0.1 Hz to the EEG data to eliminate the DC drift. Subsequently, the signal undergoes a low-pass filtering process using a fourth-order filter with a cutoff frequency of 70 Hz. To address the 50 Hz line noise, we incorporate a notch filter. The PREP pipeline [77] identifies channels with elevated noise levels in each subject's recording. Techniques such as the Artifact Subspace Reconstruction method [78] are employed to mitigate artifacts like eye blinks and muscle movements. Next, the continuous EEG data was band-pass filtered between 1-8Hz. The lower cut-off low-pass filter can be advantageous for our dataset with higher levels of noise [148]. We extract data epochs corresponding to different stimuli and conditions of the experiment utilizing EEGLAB [79]. Any epoch with a magnitude exceeding 3 standard deviations is excluded from future analysis, following the bad trial removal procedure. Baseline subtraction involves computing the average from a 200ms baseline period from the start of each recording block. For each subject, we standardize the neural response of each EEG channel to ensure zero mean and unit variance. The entire preprocessing pipeline is implemented using the EEGLAB toolkit in MATLAB.

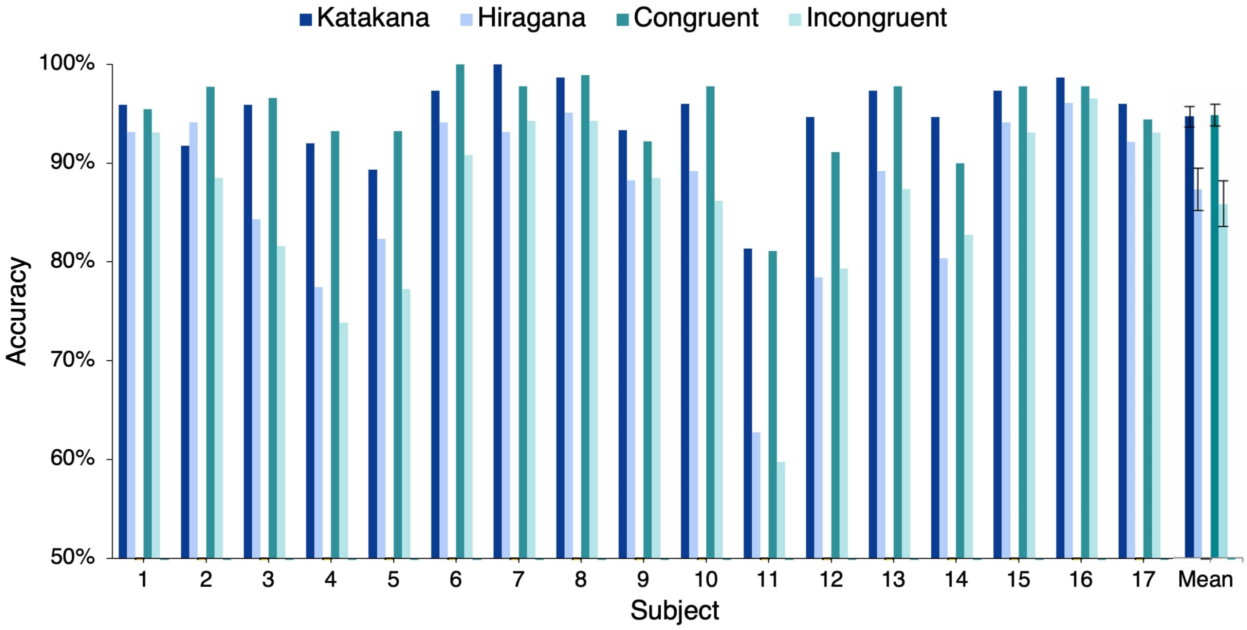


Figure 3.2: Behavioural task performance: This plot shows the percentage of Japanese words correctly associated with the image description by the subjects.

3.3 Results

3.3.1 Behavioural Task

A behavioural task was conducted to ensure that the subject successfully recalled the meaning of the Japanese words while listening to the end word of the sentence stimuli in session S3. From a set of 5 images, the subject was asked to identify the image corresponding to the Japanese end word of the sentence. Figure 3.2 shows the percentage of words whose meaning was correctly identified by the subjects. The solid line shows the overall accuracy obtained by each subject. Eleven out of seventeen subjects correctly recalled more than 90% of the word semantics (chance accuracy was 20%). Thus, the subsequent ERP-based conclusions in S3 had a strong Behavioural basis. As seen in Figure 3.2, the number of correct responses in the Japanese congruent condition was greater than the number of correct responses in the incongruent condition for all the subjects. A right-tailed paired sample t-test showed that the recognition accuracy of congruent end-words was significantly higher than incongruent end-words (congruent: 94.88, incongruent: 85.90, t (16) = 5.92, p = 1.06e−5). This indicated that it was easy to recollect the meaning of a word when it was used in the correct semantic context in a sentence. Figure 3.2 shows that the recognition accuracy of katakana words is better than that of hiragana words for all subjects except one (subject 2). A right-tailed paired sample t-test showed that the recognition accuracy of katakana words was significantly higher than hiragana words (katakana: 94.72, hiragana: 87.31, t (16) = 5.32, p = 3.46e−5). Hence, we conclude that the lexical association of katakana words with English words made it easier to recall katakana words. In the subsequent analysis, only words that were correctly recalled are used.

3.3.2 ERP analysis

The event-related potentials (ERP) are time-locked EEG responses averaged across multiple trials for the same stimulus condition. The ERPs are computed for epochs extending from 100 ms before the end word onset to 800 ms after the end word onset. The time t = 0 in the ERP plots corresponds to the onset of the end word in the stimuli sentences. The difference ERP waves were calculated by subtracting an ERP wave of one condition from the other. All the grand average ERP plots shown in this chapter are ERP responses averaged across 17 subjects. The two-sample t-test was conducted at each time sample to validate the significance of the ERP responses. All difference ERP plots are marked with time regions of significance where the difference value is significantly above zero ($p < 0.05$). This is indicated by horizontal bars at the bottom of the plot.

3.3.2.1 Effect of Incongruity

The ERP response shown by solid line in Figure 3.3 exhibits N400 effect (t (16) = -5.59, p = 2.05e−5) in 300-500ms over centro-parietal and parietal electrodes) for the difference of English congruent response (*C1S1*) from English incongruent response (*C2S1*). This result is aligned

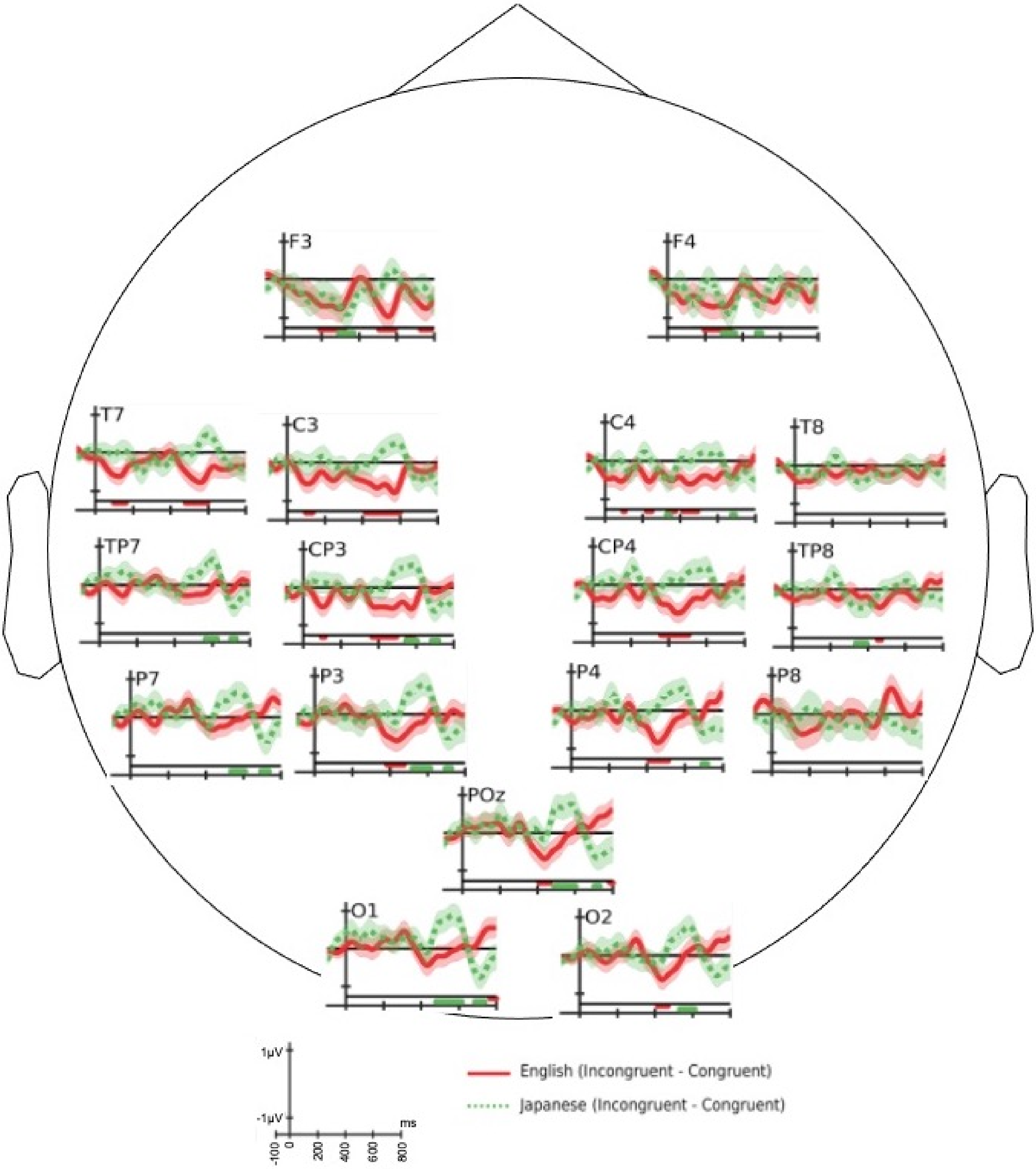


Figure 3.3: The grand average of difference ERP response of congruent end word from incongruent end word: end words in English (solid - red color) and Japanese: session 3 (dotted - green color). The horizontal bars drawn in the bottom of each plot identifies the time regions with significant (two-sample t-test with $p < 0.05$) difference in the ERP response. The bars have the same color of the associated difference waveform.

with the prior research on N400 for auditory tasks ([149]) which is elicited for semantically incongruent stimuli conditions. To the best of our knowledge, the ERP analysis for other conditions that follow is reported for the first time in the literature.

The difference between the grand average ERP response of the Japanese congruent end word from the Japanese incongruent end word is shown in Figure 3.3 as a dotted line. The semantic incongruity of newly learned Japanese end words did not evoke an N400 response relative to the congruent Japanese end words (t (16) = 0.32, p = 0.75 over the time window 300-500ms) over the cluster of centro-parietal and parietal electrode locations.

As observed in Figure 3.3, the English and Japanese end-words had different responses around 400ms and 600ms. Both the English and Japanese difference ERP did not have significant peaks in the early part of the time axis. The difference ERP response for Japanese end words elicited a P600-like component (t (16) = 5.11, p = 5.24e−5 for time window: 500-700ms) over a cluster of centro-parietal and parietal electrode locations. This figure also highlights that ERP effects of semantic incongruity are possible without a long-term learning process. A similar difference in ERP response for English end words in the 500 to 700 ms time window (t (16) = −1.05, p = 0.310) was not observed. In other words, the semantic incongruency in the stimuli did not evoke a P600 response for familiar language.

To confirm these differences statistically, we performed ANOVA on the mean ERP amplitudes across the scalp in two time windows (300-500 ms and 500-700ms) of interest. We considered congruity and scalp region (frontal/central/parietal) as the independent factors. In the N400 time window (300-500 ms), we observed robust effects of congruity for English end-words. We observed similar effects for Japanese end-words in the P600 (500-700ms) window. The ANOVA reveals significant interaction between language and congruity in both the time windows (for 300-500ms window: F(1,68) = 16.13, p = 6e−5 and for 500-700ms window: F(1,68) = 37.85, p = 9e−10).

3.3.2.2 Effect of Word Learning

Figure 3.4 shows the ERP response for the same set of Japanese language words before and after learning the semantics. The Japanese words exposed without semantic knowledge (solid line) evoke early positive peaks betwen 100ms and 300ms. This P300 (P3a) response is possibly an indicator of exposure to novel information. The P300 response disappear in the exposures after the subject learned the meaning of the word. The Japanese end word response before semantic learning (solid line) and Japanese congruent response post semantic learning (dotted line) has a negative deflection before 600ms. Both these responses also have late positive component (LPC) after 600ms, which possibly is an indicator of the recognition of the code-switch. In summary, the ERP for congruent (t (16) = 3.97, p = 5e−4) and incongruent (t (16) = 5.89, p = 1.1e−5) conditions post semantic learning elicit significantly different responses in 500-700ms from word onset.

Figure 3.5 shows the differences in EEG responses to Japanese end words before and after learning their meaning. The difference ERP waveforms of congruent (solid) and incongruent (dotted) conditions do not show any significant difference in the 0-500ms range. Both conditions evoke significant negative peak around 200ms. It is also noted that there is a significant difference between before and after learning in the in-congruent condition than for the congruent condition.

3.3.2.3 Effect of Phonetic Similarity to Known Words

Figure 3.6 shows the difference ERP response for katakana (loan words in Japanese) and hiragana words separately. It shows the difference ERP of congruent condition from incongruent condition. Both hiragana and katakana words show significant P600 response. We performed a paired t-test over a cluster of centro-parietal and parietal electrode locations in the time window 500-700ms to ascertain the statistical significance. The difference ERP of katakana

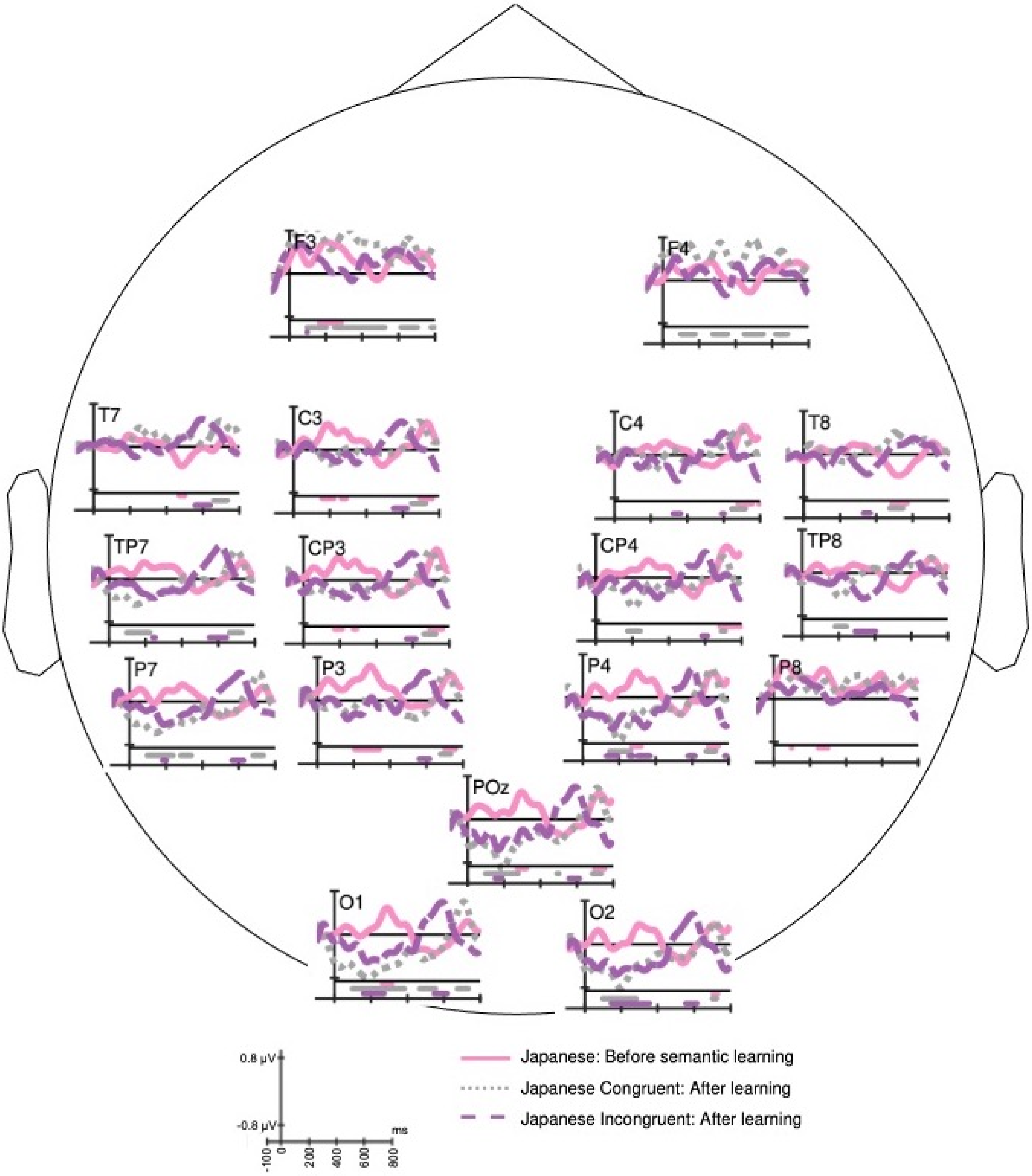


Figure 3.4: The grand average of ERP responses to Japanese end word before learning its meaning (solid), Japanese congruent end word (dotted) and Japanese in-congruent end word (dashed) after learning the meaning. The horizontal bars drawn in the bottom of each plot signify the time regions with significant (t-test with p< 0.05) ERP amplitude from the value of 0. The bars have the same color of the associated waveform.

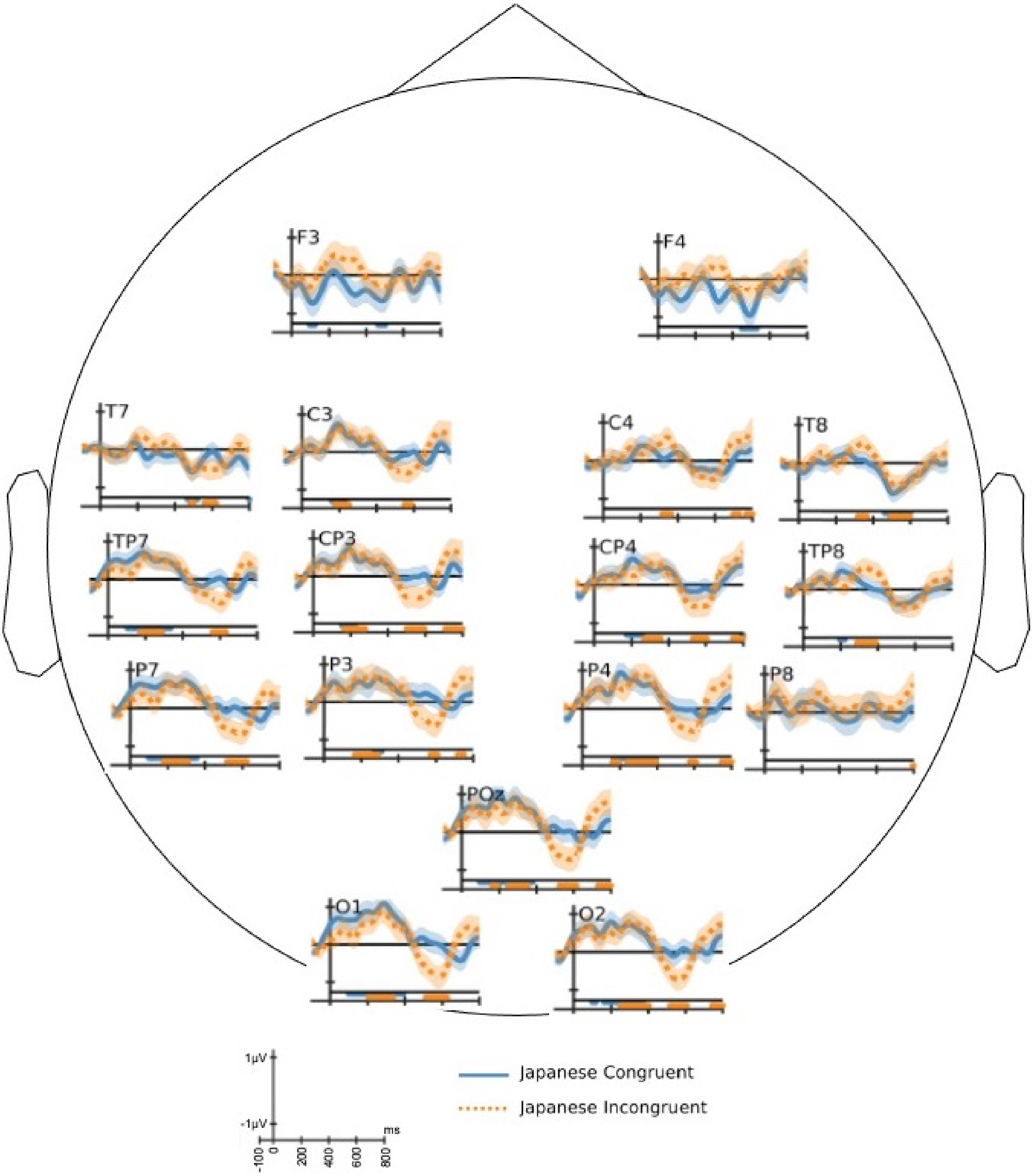


Figure 3.5: Grand average difference ERP response of: Japanese Congruent end word from Japanese end word response before learning its meaning(solid); and Japanese In-congruent end word from Japanese end word response before learning its meaning (dotted). The horizontal bars drawn at the bottom of each plot signify the time regions with significant (two-sample t-test with p< 0.05) difference in the ERP response. The bars have the same colour of the associated difference waveform.

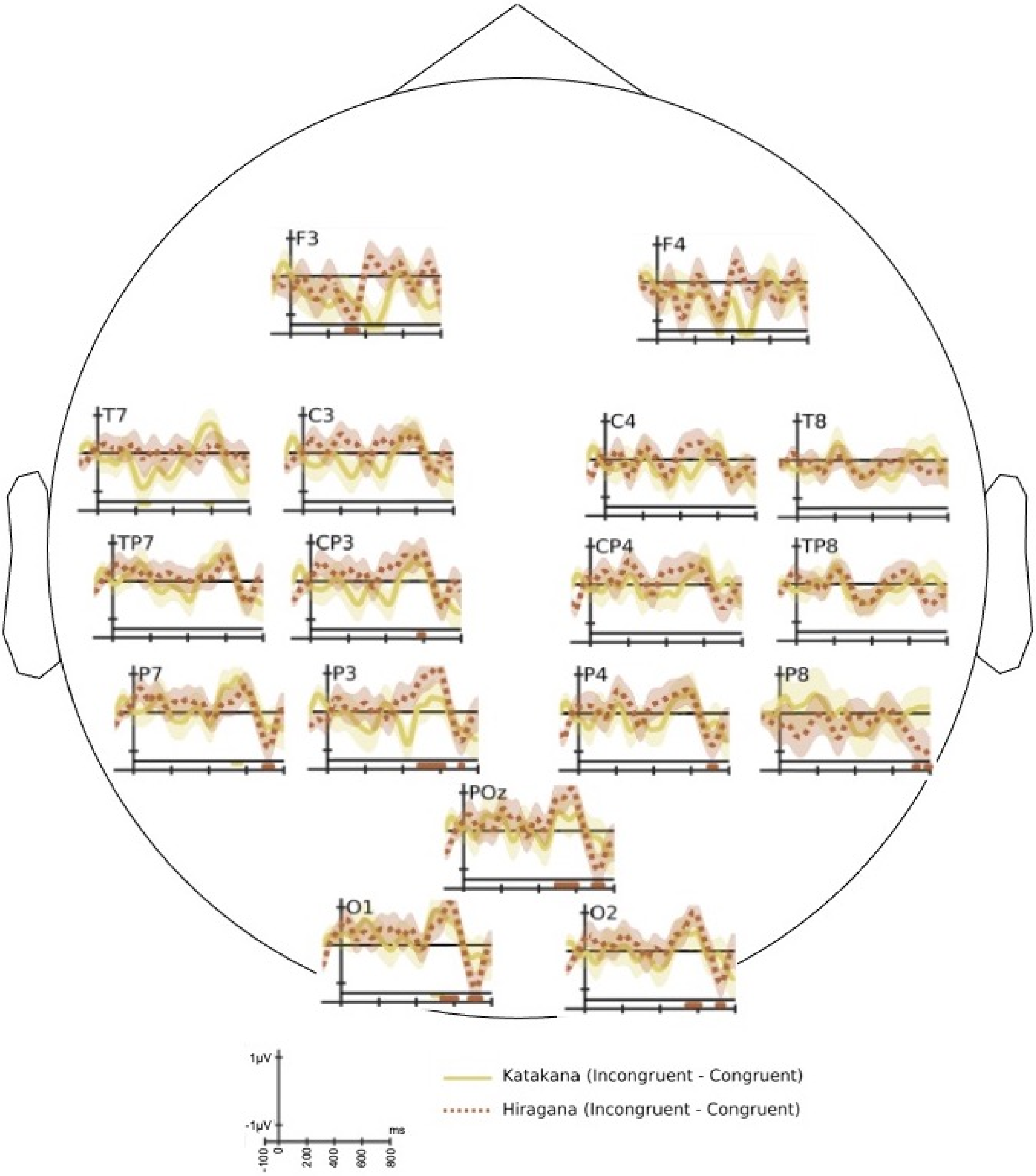


Figure 3.6: The grand average of difference ERP of congruent end word response from incongruent end word response for katakana words (solid) and hiragana words (dotted). The horizontal bars drawn in the bottom of each plot signify the time regions with significant (two-sample t-test with $p < 0.05$) difference in the ERP response. The bars have the same color of the associated difference waveform.

words showed t (16) = 3.39, p = 1.87e−3 and that of hiragana words showed t (16) = 4.18, p = 3.53e−4 in the P600 window. Thus, the hiragana words show larger P600 amplitude than katakana words (t (16) = 3.91, p = 6.24e−4). The statistical significance is more established for hiragana words in the occipital and left-parietal electrode locations.

Figure 3.7 shows that event related potential evoked by katakana and hiragana words before semantic exposure does not show any significant difference. Note that, a statistical significance would be highlighted as black horizontal bar in the figures. The absence of any horizontal bars means that, for all the electrodes considered here, the ERP responses for katakana and hiragana words were not statistically significantly different.

3.3.3 Topographical Analysis

Figure 3.8 shows the topographical distribution of difference of ERP (grand-average) amplitudes in different time windows. The mean value of ERP amplitudes in each time window is plotted here. The top row shows the difference of English congruent end-word responses from English incongruent end-word responses, the middle row shows the difference between Japanese congruent end-word responses from Japanese incongruent end-word responses and the bottom row shows the difference of Japanese end-word responses before learning its meaning from the Japanese end word responses after learning its meaning.

As shown in previous works, the N400 response is significant over the centro-parietal region (see Figure 3.8 top row). Similarly, the Japanese congruent vs incongruent difference is significant over the centro-parietal region in a 450-650ms time window as seen in Figure 3.8 middle row. It is more evident in the left hemisphere than the right hemisphere of the scalp. The last row of Figure 3.8 shows ERP differences between the EEG responses before and after semantic learning in frontal, parietal and occipital regions in the initial time windows after the end word onset. The response over the rear part of the brain is low in magnitude from 350ms onwards, while the response in the frontal part sustains longer. In the frontal electrodes, the ERP after semantic learning is more positive than before semantic learning. This is also more pronounced

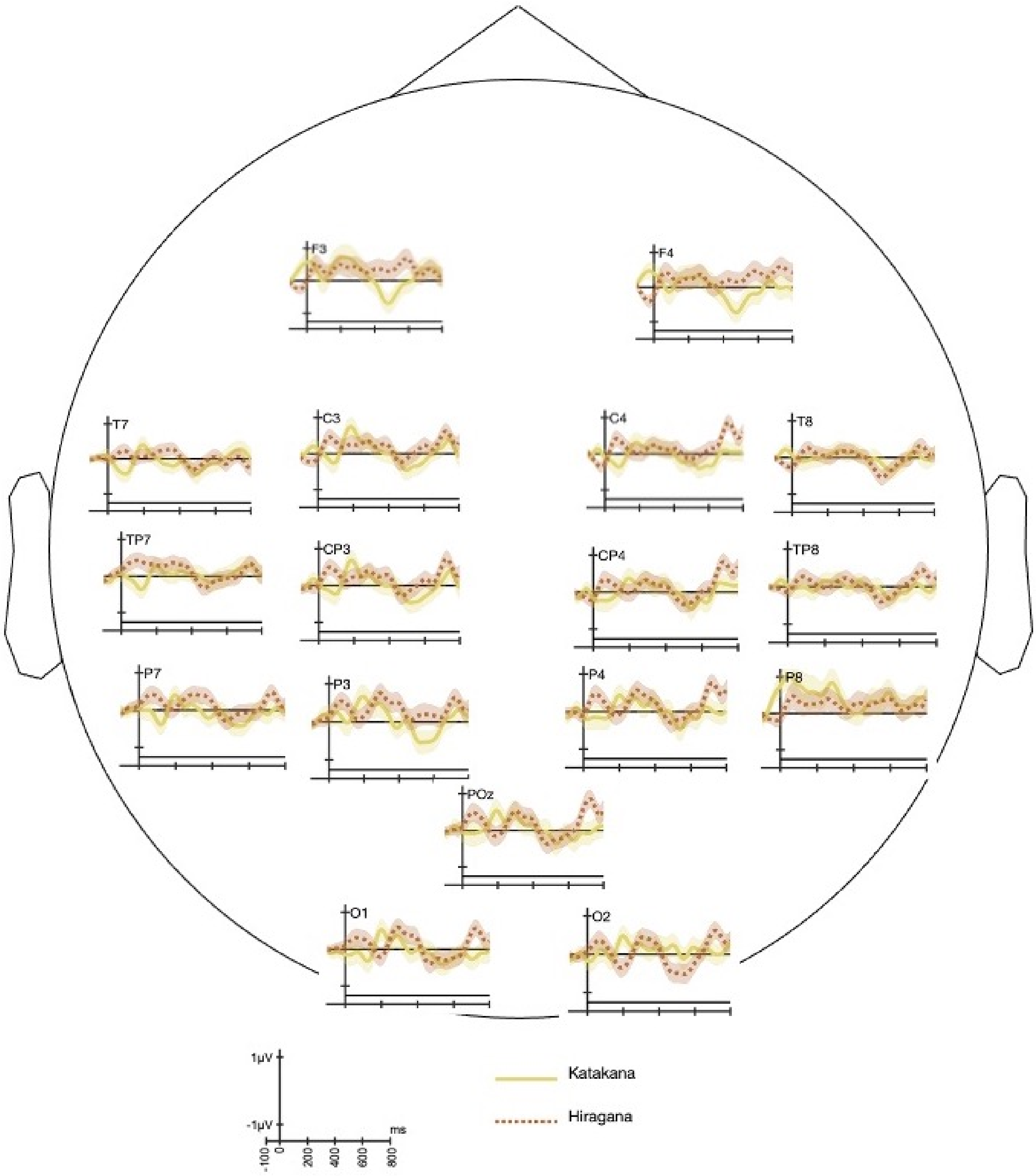


Figure 3.7: ERP plot for Katakana and Hiragana words in session 1 (before semantic exposure) of the experiment. Before knowing the meaning of the word, perception of both types of words did not show any significant difference in event related potentials. Two sample t-tests conducted did not give any time region with significant difference between the responses for two types of words (Note: If there is any significant region, that will be marked with horizontal bars below the black horizontal line at the bottom of each subplot in the figure).

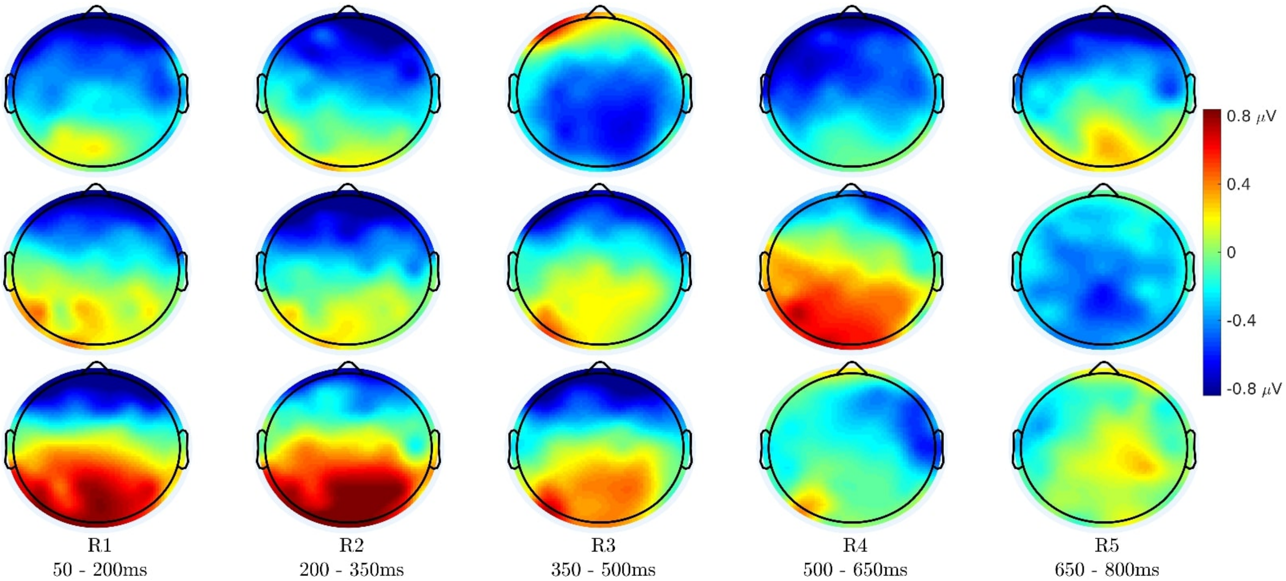


Figure 3.8: Topography distribution of mean difference of ERP (grand-average) amplitudes in different time windows. Top row: Difference of English congruent end word responses from English incongruent end word responses (S1) ; Middle row: Difference of Japanese congruent end word responses from Japanese incongruent end word responses (S3) ; and Bottom row: Difference of Japanese end word responses before learning its meaning from the Japanese end word responses after learning its meaning (both in congruent context).

in the left hemisphere than in the right.

3.3.4 Correlation Plot

Let the ERP waveform for channel c for language l and for condition s be denoted as $x_c^{s,l}(t)$ where channel c ranges from $1-64$, language l corresponds to 1 for English and 2 for Japanese, condition s corresponds to 1 for congruent and 2 for incongruent condition. For different time regions R1-R5 (where R1 ranges from $50-200$ms, R2 from $200-350$ms, R3 from $350-500$ms, R4 from $500-650$ms and R5 from $650-800$ms), the correlation matrix for each language l is computed using,

$$\mathbf{C}_{R_j}^l(i,k) = \sum_{t=s_{R_j}}^{e_{R_j}} x_i^{l,1}(t) x_k^{l,2}(t)$$

where R_j corresponds to regions R1-R5 and s_{R_j} and e_{R_j} denote the start and end time instants of the region. Thus, the matrix of values $C_{R_j}^l$ denotes the cross-correlation between the ERP

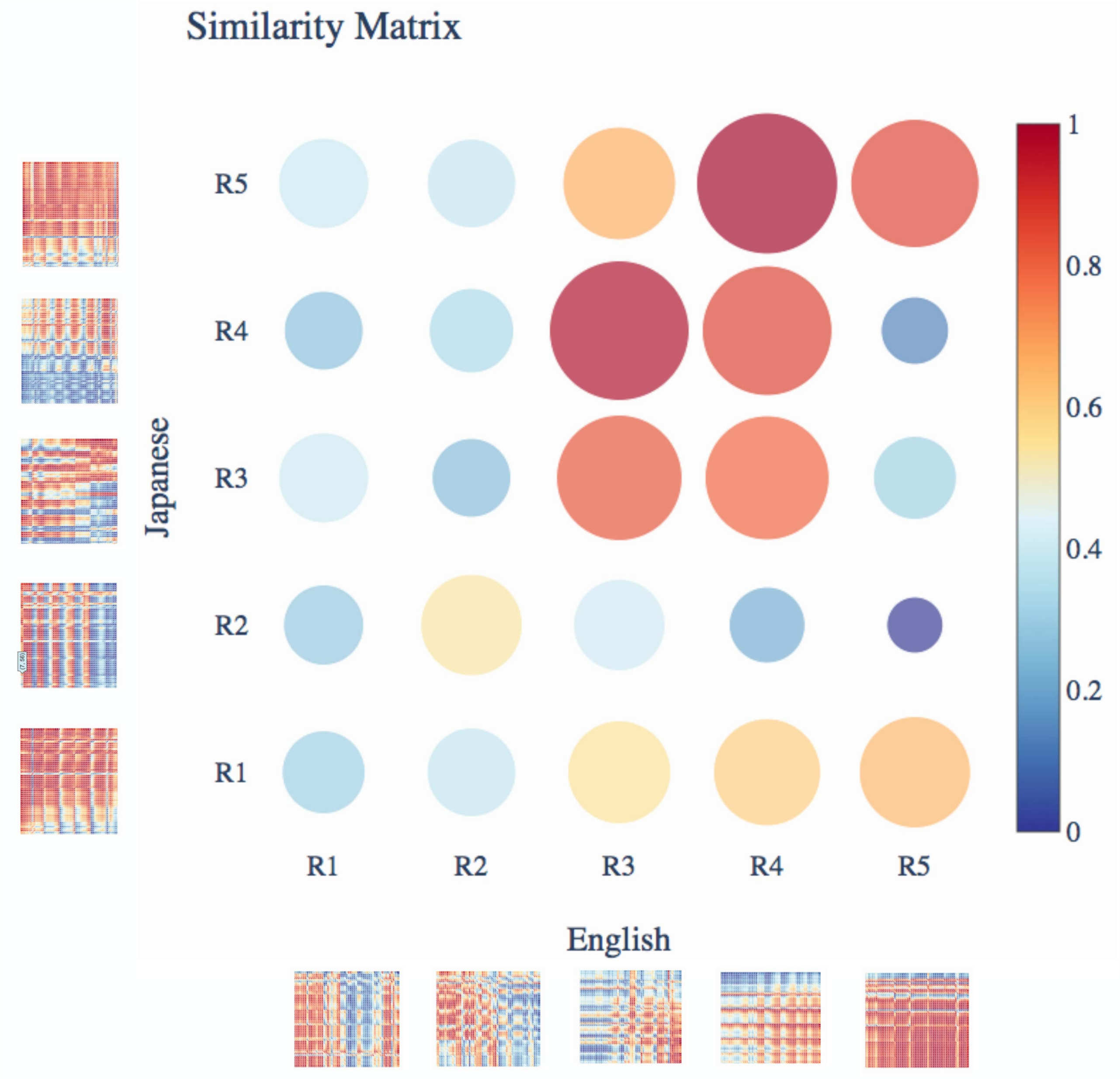


Figure 3.9: Distance matrix between English and Japanese correlation matrices in different time windows. The cross-channel correlation is computed between the congruent and incongruent endword ERP responses across all channels for English (S1) and Japanese (S3). The time regions in the figure are as follows R1: 50-200ms; R2: 200-350ms; R3: 350-500ms; R4: 500-650ms; and R5: 650-800ms. The highest similarity was between English responses at R4 and Japanese responses at R5.

responses for the congruent and incongruent conditions. A high value at location (i, k) for this matrix indicates that for the channel pairs (i, k) the EEG responses to congruent and incongruent conditions are highly correlated (and time-synchronized). Similarly, a low negative value indicates that for the channel pairs (i, k) the EEG responses to congruent and incongruent conditions are negatively correlated (and time-synchronized). And a value for $\mathbf{C}^l_{R_j}(i, k)$ that is close to 0 indicates that the responses to congruent and incongruent stimuli conditions are uncorrelated between channel pairs (i, k).

For each language, these matrices are generated. A distance measure using the Frobenius norm of $\mathbf{C}^1 - \mathbf{C}^2$ is computed for every pair of time regions. This generates the 5×5 distance matrix.

The distance matrix shown in Figure 3.9 is computed between the congruent and incongruent end word ERP responses across all channel pairs. It is computed for different time windows as shown in Figure 3.9 to compare the correlation plot of the ERP waveforms for the two languages. The distance matrix in Figure 3.9 shows that the English difference response in R3 (350-500ms) has high similarity (least distance) with Japanese difference response in R4 (500-650ms). Similarly, we observe a high similarity between the difference response of English in the R4 (500-650ms) window with difference response of Japanese in the R5 (650-800ms) window.

3.3.5 Statistical Analysis of ERP effects

We have used the mean amplitudes extracted from five non-overlapping time windows of 150ms duration between 50 to 800ms from the word onset in repeated-measures ANOVA. The ANOVA used four within-subject factors: language (English/Japanese), congruency (congruent/incongruent), learning (before/after), and scalp region (frontal/central/parietal). The ERP effects suggested group differences in particular topographic regions. The language and congruency had significant interaction in all time windows except at 200-350ms. It should be noted that the interaction is highly significant with a larger F-ratio in the 350-500ms window (F=31.09, $p < 0.01$) and 500-650ms window (F=73.08, $p < 0.01$). The language and scalp region factors had significant interaction in two time windows: at 50-200ms and at 500-650ms.

The factors of learning and scalp region had significant interactions in all time windows except 500-650ms. This implies that the process of learning had topographic selectivity. The highest significance is observed in the 200-350ms window (F=28.49, $p < 0.01$). The congruency and scalp region also had significant interactions in the 50-200ms and 500-650ms windows.

3.4 Discussion

The subjects who participated in the experiment had no prior exposure to Japanese and hence, these Japanese word semantics were accessed from short-term memory in the rapid learning task undertaken during this experiment. Thus, the processing of the newly learned Japanese words was found not to evoke long-term memory regions and the N400 ERP component was not present in the difference waveform of the newly acquired Japanese end words (Figure 3.3). The grand average difference of ERP response of Japanese congruent end word from incongruent end word showed a P600 component (Figure 3.3). This can be attributed to semantic P600 component. The work by [150] showed that the violation of semantic congruity in sentences with strong semantic relationship between its noun and verb can evoke a semantic P600 response. It is also worth noticing that the congruent and incongruent semantic conditions showed different brain responses for familiar language at around 400ms and for newly acquired words at around 600ms. The semantic differences in Japanese words evoked a later response than the English words. This may be due to the reason that the newly learned words required a reanalysis to integrate itself with the sentence. The ERP of incongruent words evoked a significant positive peak while the ERP of congruent words evoked a negative peak around 600ms (Figure 3.4).

Figure 3.6 shows that both katakana and hiragana words evoked P600 component in the difference ERP. The hiragana P600 response has higher amplitude than the katakana response over the parietal and parieto-occipital electrodes. The katakana words are loan words from English, but they are pronounced with the Japanese adaptation. As shown in the Behavioural responses, human subjects find it easier to recall the meaning of katakana words and hence, the involved reanalysis is not as strong as the hiragana words. The observation that hiragana words have higher P600 amplitude than katakana words is similar to the higher amplitude of N400 observed for highly unexpected end word in the known language ([112, 111, 113, 114]). Figure 3.4 and Figure 3.5 show the effects of semantic learning. The Japanese words in the first exposure elicited a significant P300 response owing to the novelty of the stimuli. After semantic learning, the ERP of congruent and incongruent conditions did not have significant differences in the early part of the response. The differences are significant after 500ms from the onset of the end word. This shows that the semantic processing of the newly acquired words may occur with a delay of more than 500ms from the word onset.

The foreign word at the end of the sentence is comprehended as a form change, since it evokes a P600 potential instead of N400 potential. This is comparable to the semantic illusion condition shown by [151]. The code-switching to a newly learned word at the end of the sentence requires re-analysis to integrate it with the prior sentence context. As shown in other code-switching studies like [152, 153], we observe that P600 is elicited for the congruent end word in context. In this work, we show that P600 potential is evoked while perceiving newly acquired word from a foreign language employed in a code-switched manner. Further, we see a difference in the P600 latency in the response for the newly acquired end word used in congruent and incongruent context. When the newly acquired word is used in congruent condition, we see the positive peak appearing at a slightly later time with a comparable amplitude. This difference in the response to the Japanese word used in congruent and incongruent condition, shows the ERP effects of rapid semantic acquisition of foreign language words. This difference in responses for congruent and incongruent cases for the newly acquired word illustrates that the ambiguity is recognized by involving a higher cognitive load. For the newly acquired words, the word used in congruent condition evokes a lesser positive potential around 600ms from the word onset and shows a more positive deflection in 700 to 800ms (peaking around 750ms). This can be observed in Figure 3.4. This implies that the semantic integration of newly learned words happens much later in time compared to the similar process in a proficient language word. The underlying cognitive process may also be bi-phasic: recognition and then integration of the meaning with the sentence context.

The topographic plots show that P600 responses are also stronger over the centro-parietal and parietal regions like the N400 response. But the P600 response has a left hemisphere selectivity. The scalp distribution of difference ERP before and after semantic learning elicit significant responses in the early part of the ERP waveform. It has strong positive response in the frontal part owing to the P300 response. The correlation analysis in Figure 3.9 shows that the highly negative correlation that exists between the EEG responses for congruent and incongruent conditions for English at about 400 ms also appear for Japanese stimuli but at a later time instant of 600ms.

3.5 Chapter Summary

The main contributions from this chapter are the following,

- The investigation of event-related potentials (ERP) in EEG signals during rapid language learning in subjects with no prior exposure to a specific language addresses a significant knowledge gap.

- EEG patterns elicited by semantically matched and mis-matched Japanese end-words in English sentences differ for newly learned Japanese words compared to the English end-words (already proficient).

- A short-term learning task involving new language words triggers a delayed and opposite P600 component compared to the ERP observed for known language words, indicating higher cognitive load during recall of newly learned foreign language words.

- Scalp electrodes show that these semantic activations were predominantly located in the parietal and occipital regions.

- The absence of the N400 component in this rapid learning task suggests its association with long-term memory processing.
- The amplitude of semantic incongruity, as reflected by the P600 ERP component, is higher for pure foreign words in the newly acquired language compared to loan words derived from English, indicating that similarities with known language words facilitate semantic learning.

In summary, the study found that the ERPs for semantically matched and mis-matched Japanese end-words in English sentences are different for newly learned Japanese words. This suggests that recall of newly learned words of a foreign language is more cognitively loaded, and that similarities with known language words will aid in semantic learning. So far, we investigated word learning effects with isolated stimuli, with and without context. In the next chapter, we explore the neural encoding of speech perception with natural continuous stimuli.

Chapter 4

Encoding of Semantic Word Level Information in EEG for Natural and Dichotic Listening

4.1 Introduction

In the preceding chapters, we explored the process of learning words from an unfamiliar language. We analyzed the underlying neural patterns with multi-trial analysis techniques in EEG. This involved experimenting solely with isolated stimuli, such as single words and novel end-words within isolated sentences. However, this chapter advances our investigation to focus on the neural encoding of more naturalistic stimuli. Specifically, we delve into the perception of continuous stimuli in a familiar language. Although we examine the impact of word segmentation on speech perception, the scope of this chapter does not include the actual process of word learning, unlike previous chapters.

The recorded EEG signal during a speech listening task has been shown to contain information about the stimulus [26, 28, 29]. One can investigate how the brain comprehends continuous speech by developing models that relate the speech to the EEG signal using machine learning techniques [154].

The early attempts explored linear models for relating continuous natural speech to EEG responses [155, 156, 157, 158]. They can be categorized into three different types - forward models, backward models, or hybrid models. The forward models predict EEG from speech stimuli, while the backward models reconstruct speech from EEG responses. In many studies, the correlation between the predicted and ground truth signals is considered a measure of neural tracking [98]. However, linear models may be ill-equipped to capture the non-linear nature of the auditory system. Deep neural networks have recently been employed to compare and analyze speech stimuli and EEG responses. Several studies have shown promising results with deep learning models for EEG-speech decoding [1, 16, 159, 160].

In many of the computational approaches, the speech envelope has been the most commonly used feature [155, 156, 158]. It is shown to be synchronized with neural oscillations in the auditory cortex [26, 161]. Other features such as spectrograms [26, 162], phonemes [26], linguistic features [162, 163], and phono-tactics [164] have also been explored with linear forward/backward models. Lesenfants et al. [165] demonstrated that combining phonetic and spectrogram features improves the EEG-based speech reception threshold (SRT) prediction.

While forward/backward models and correlation tasks were previously explored, the match mismatch tasks have been recently investigated as an alternative task [166, 159]. Here, the task is to identify whether a portion of the brain response (EEG) is related to the speech stimulus that evoked it. In the previous studies using the match mismatch task, the auditory stimulus and speech of a fixed duration (5s) are processed through a series of convolutional and recurrent layers [1, 160, 167].

In this work, we argue that the prior works on speech-EEG match mismatch tasks are incomplete without considering the fragmented nature of speech comprehension. While speech and EEG signals are continuous, the neural tracking of speech signals is impacted by the linguistic markers of speech [168]. The most striking evidence comes from models of word surprisal [169] with N400 response evoked for unpredictable words [29, 170]. In the simplest form, we hypothesize that the task of relating continuous speech with EEG must also include word-level segmentation information.

4.2 Tasks Explored

We propose a deep learning model to perform MM classification tasks on variable length inputs using word boundary information. The model consists of convolutive feature encoders for both the speech and EEG inputs. Further, the feature outputs incorporate the word segmentation information, which is obtained by force-aligning the speech with the text data using a speech recognition system, through a word-level pooling operation. The pooled representations are further modelled with recurrent long short-term memory (LSTM) layers to model the inter-word context. The final output from the LSTM network for the speech and EEG streams is used in the match mismatch classification task.

In addition to analysing speech perception during natural speech listening, it is important to analyse how the brain comprehends speech in complex listening conditions like dichotic tasks. Such auditory attention decoding (AAD) experiments have investigated auditory processing and attention mechanisms in the brain [155, 171]. Auditory attention decoding refers to the process of decoding or identifying the specific auditory stimulus or sound that an individual is paying attention to. It involves analysing neural responses or brain activity recorded while the person is engaged in an auditory task, such as listening to multiple sounds simultaneously. One such listening condition is dichotic listening, in which participants are presented with different auditory stimuli simultaneously in each ear. The stimuli are presented simultaneously but with different content (spoken by different speakers) to each ear, creating a situation where the participant must selectively attend to one ear while ignoring or minimizing their attention to the other. This paradigm allows us to study how individuals allocate their attention and process auditory information under conditions of competing or conflicting stimuli.

This capability of the human brain is termed the cocktail party effect, which refers to the remarkable ability of the human brain to selectively focus attention on a specific auditory stimulus while filtering out other competing sounds in a noisy environment, much like being able to hear a single conversation at a crowded party. The primary goal of this study is to examine the cocktail party effect, which involves detecting the specific sound to which a subject is attending. Additionally, we aim to analyse the influence of word boundary information on dichotic listening. By investigating these aspects, we seek to gain insights into how individuals perceive and allocate attention to different auditory stimuli in complex listening environments.

In addition, we explore the relative perception of phonological and semantic characteristics in complex listening environments, such as natural and dichotic listening. We propose utilising a sentence-level model that incorporates word segmentation information [172]. This study proposes a framework for auditory stimulus-response (EEG) modelling using a match-mismatch (MM) task. This framework is designed to capture the neural responses in both monoaural and dichotic listening scenarios. For dichotic listening, we formulated the MM task as a means to perform auditory attention detection (AAD), aiming to identify the stimulus the subject was attending to by analysing the recorded neural response. To conduct this study, we utilised two distinct speech-EEG datasets [28]: one collected during a natural speech listening task and the other during a dichotic listening task.

4.3 Key Contributions

Our approach involves employing a deep learning network [172] for stimulus-response modelling, which consists of two separate sub-networks: one for processing the EEG signals and the other for processing the stimulus. The EEG sub-network comprises convolutional layers, word boundary-based average pooling, and a recurrent layer incorporating inter-word context. The speech sub-network is reconfigured based on the input feature.

This study evaluates the effectiveness of different stimulus features in stimulus-response modelling during complex listening scenarios. In this study, the speech envelope is the acoustic feature used to represent phonological traits. The semantic feature employed is word2vec (w2v), which represents words as vectors in a high-dimensional space based on their contextual usage. We conducted a comparative analysis to evaluate the relative contributions of acoustic and semantic features in both natural and dichotic listening conditions. The results indicate that semantic features exhibit higher accuracy in match-mismatch (MM) classification during dichotic listening, while both features perform similarly in normal listening conditions. Furthermore, our work highlights the significance of word boundary information in auditory attention detection.

The major contributions of this study are:

- Formulating a paradigm for auditory stimulus-response (EEG) modelling through a match-mismatch (MM) task in mono-aural listening as well as in dichotic listening

- In the mono-aural natural speech listening task, the match-mismatch performance of audio signals is enriched by the induction of the word boundaries. Further, the textual information alone provides comparable MM performance as the audio signal.

- In the dichotic speech listening task, the MM performance of text data is significantly higher than that of the audio signal, indicating that EEG signals encode higher-level semantic information than the acoustic envelope information.

- We propose a multi-modal architecture, where both speech and text features are matched with the underlying EEG signal, for the MM task in both tasks. The performance of the multi-modal model improves over the individual modalities of text and speech, indicating that the EEG signal jointly encodes the semantic and acoustic content of the stimulus.

- The research proposes a Manhattan distance-based loss function for the match mismatch task and demonstrates its effectiveness through improved classification performance compared to prior works. A detailed set of ablation experiments investigates the influence of word boundary information on speech EEG matching and auditory attention detection.

4.4 Materials and Methods

4.4.1 Dataset

These scenarios include subjects listening to uninterrupted, natural speech and a situation where the subjects are exposed to the cocktail party effect. In the latter case, the subjects listen to two distinct audio streams simultaneously, each directed to a separate ear. For this discussion, we refer to the first dataset as the natural speech dataset and the second as the dichotic (termed as "cocktail party" in the original dataset) dataset.

In both experiments, the stimuli were presented using Sennheiser HD650 headphones and the Presentation software provided by Neurobehavioral Systems. The participants were instructed to keep their gaze fixed on a crosshair displayed on the screen throughout each trial and to minimise activities such as eye blinking and other motor movements.

4.4.1.1 Natural Speech Dataset

The natural speech (NS) dataset contains electroencephalographic (EEG) data recorded from 19 subjects as they listened to continuous speech. The subjects listened to a professional audiobook narration of a well-known work of fiction read by a single male speaker. The data consists of 20 trials roughly the same length, each containing 180s audio. The trials preserved the chronology of the storyline without repetitions or breaks. The sentence start and end time and the word-level segmentation of the speech recordings are provided. The word segmentation is obtained using a speech recognition-based aligner [173]. The EEG data were acquired using 128-channel BioSemi system at a sampling rate of 512Hz, while the audio data is played at 16kHz. Overall, the speech-EEG data amounted to a duration of 19 hours.

4.4.1.2 Dichotic Dataset

The cocktail party dataset used in this study comprises EEG recordings obtained from 33 subjects. The participants underwent a total of 30 trials, each lasting for 60 seconds. They were presented with two well-known fictional works, where one story was delivered to the left ear and the other to the right ear. Different male speakers read each story. The participants were divided into two groups of 17 and 16 individuals (with one subject excluded), respectively. Each group was instructed to focus their attention solely on the story presented in either the left or right ear throughout all 30 trials. Following each trial, the participants were required to answer multiple-choice questions about both stories, each offering four possible answers. To maintain consistency, the audio streams of each story within a trial were normalized to have the same root mean squared (RMS) intensity. To prevent the unattended story from capturing the participants' attention during silent periods in the attended story, any silent gap exceeding 0.5 seconds was truncated to a duration of 0.5 seconds.

4.4.2 EEG Preprocessing

The EEG preprocessing pipeline utilized in this study for both datasets was based on the CNSP Workshop 2021 guidelines[1] and implemented using the EEGLAB software [174]. The pipeline involved several steps to ensure the quality and reliability of the EEG data.

First, a low-pass Butterworth filter with a cutoff frequency of 32 Hz was applied to the EEG signal. This filter helps attenuate high-frequency noise and artefacts that are unrelated to the neural activity of interest. Next, a high-pass Butterworth filter with a cutoff frequency of 0.5 Hz was employed to remove low-frequency artefacts and eliminate any potential DC offsets or drifts in the signal. Following the filtering steps, the EEG data and any external channels were downsampled to a rate of 64 Hz.

A channel rejection method based on the EEG data's variance was employed to identify

[1] https://cnspworkshop.net/resources.html

and replace bad channels. The channels with excessively high variance, indicating potential artefacts or poor signal quality, were considered bad and replaced using a spline interpolation technique. This interpolation is performed using the remaining channels to preserve the spatial information of the EEG data.

After channel rejection and interpolation, the EEG channels were re-referenced to a specific set of external channels known as mastoids. This re-referencing step helps minimize the effects of common noise sources and improves the interpretability of the EEG data by providing a reference point that is less prone to artefacts and individual differences.

Lastly, to normalize the data and ensure comparability across channels and trials, a z-score transformation was applied. This normalization process computes the z-score independently for each channel and each trial, subtracting the mean and dividing by the standard deviation.

In summary, the EEG preprocessing pipeline involves bandpass filtering (0.5-32Hz), downsampling (to 64Hz), bad channel identification and replacement, re-referencing to mastoids, and z-score normalization.

4.4.3 Acoustic Feature Extraction

4.4.3.1 Speech Envelope

The speech envelope represents the variations in the amplitude of the speech signal over time. It is obtained by extracting the magnitude of the signal's analytic representation using the Hilbert transform. The temporal envelopes of sounds contain critical information for speech perception [175, 176]. It has been shown that the auditory cortex can temporally track the acoustic envelopes [161]. The strength of cortical envelope tracking may indicate the extent of speech perception ([177, 178, 158]). Speech envelope provides valuable information about the overall shape and dynamics of the speech signal, including prosodic features, syllabic structure, and phonemic transitions. Therefore, it is considered an important acoustic feature. However, it should be noted that the speech envelope does not contain significant information regarding the meaning or context of the speech input. This has been used as an acoustic feature for both datasets.

4.4.3.2 Mel Spectrogram

The mel spectrogram of the speech signal is used as a stimulus feature for the natural speech dataset. As the raw audio signals were unavailable in the dichotic dataset, we could not compute mel spectrogram feature for dichotic listening tasks. The mel spectrogram is computed for each sentence. A mel filter bank with 28 filters distributed in the mel-scale ranging from 0-8kHz frequency is used. The input audio is pre-emphasized with a factor of 0.97 before windowing. In order to obtain speech features at a sampling frequency of 64Hz, the spectrogram computation uses a Hamming window function of the width 31.25ms with half overlap.

4.4.4 Semantic Feature Extraction

Semantic vectors for content words were derived using the word2vec algorithm [179]. The text embedding was computed for each sentence. The text transcription of the speech stimulus is provided in the dataset. This study used this text data to obtain features representing the semantics of the stimulus speech signal.

4.4.4.1 Word2vec text embedding

Word2Vec (w2v) is a popular word embedding technique proposed by Mikolov et al. [179]. This algorithm generates a vector representation for each word. This study used pre-trained vectors trained on a subset of the Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases. The phrases were obtained using a basic data-driven approach outlined in [180]. These pre-computed word vectors are available for free download[1]. The fundamental notion of w2v embedding is that words with similar semantics tend to be closer to each other in their vector space representation. Word2vec vectors of constituent words were concatenated together to get sentence representation. This

[1] https://code.google.com/archive/p/word2vec/

representation was fed as input to the models described in the following sections.

4.4.5 Match-mismatch classification task

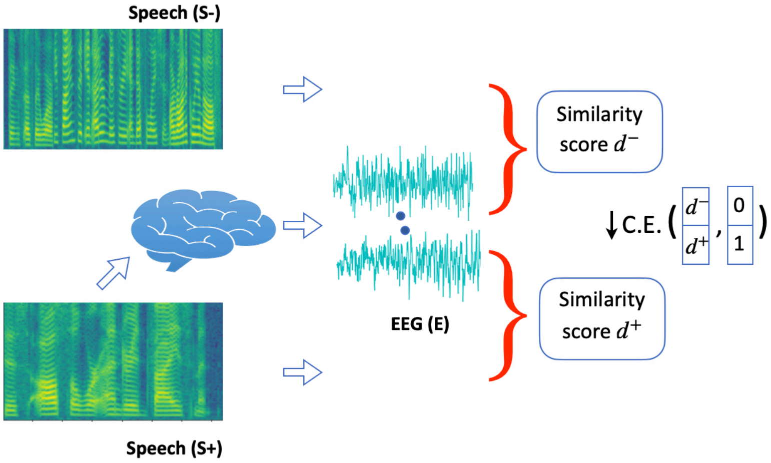

Speech (S-)
Similarity score d^-
$\downarrow$C.E. $\left(\begin{bmatrix} d^- \\ d^+ \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$
EEG (E)
Similarity score d^+
Speech (S+)

Figure 4.1: **Match-mismatch (MM) classification task:** It is a binary classification paradigm associating the EEG and speech segments. The EEG segment (**E**) and corresponding stimulus sentence (**S**+) form the positive pair while the same EEG and another (unrelated) sentence (**S**−) form the negative pair. The similarity score computation is achieved using the model depicted in Figure 4.3. Here, C.E. denotes the cross entropy loss.

The accuracy of a match-mismatch classification task is employed in this study as a measure of the neural tracking of speech. Figure 4.1 illustrates this paradigm in detail. The classification model is contrastively trained to relate the speech segment to its corresponding EEG response. In this study, the segment is chosen to be a sentence. We also compare with prior works [1, 181], which perform this task at the sentence level. The time-synchronized stimulus of the EEG response segment is the matched speech. Another sentence from the same trial of data collection is chosen as the mismatched speech. Selecting mismatched samples from the same trial makes the classification task challenging enough to encourage the model to learn the stimulus-response relationships. This sampling approach also avoids the chances of memorizing the speech features along with its label. A mini-batch contains both matched and mismatched pair of a speech sentence to ensure variability and diversity in data. This ensures that the model learns the underlying similarity patterns between speech and its corresponding EEG responses instead of memorizing the pairs.

4.4.6 Auditory Attention Decoding as a Match-mismatch task

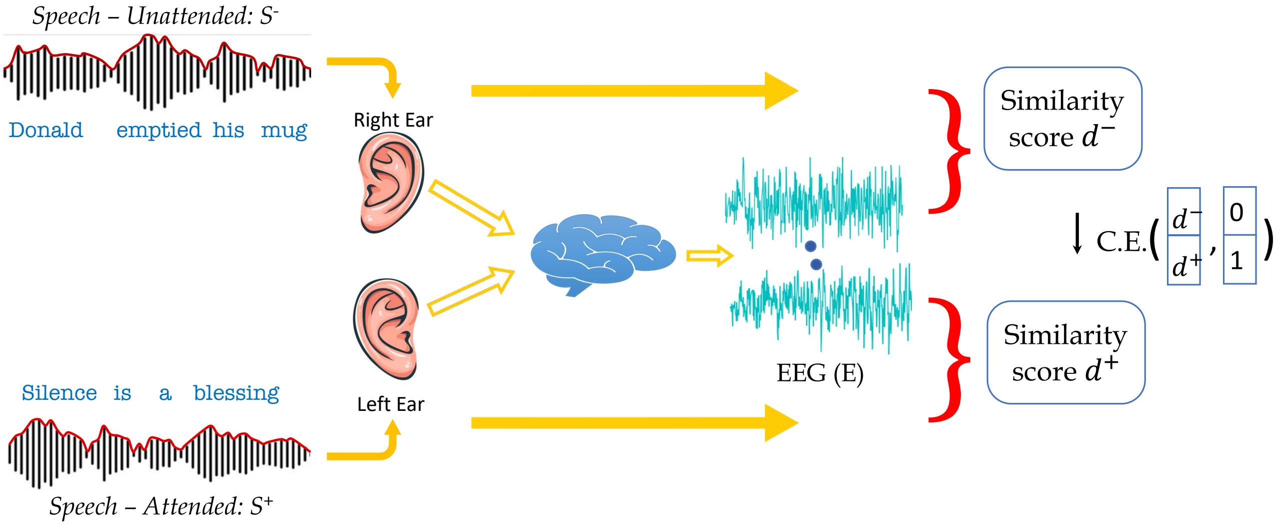


Figure 4.2: **AAD as Match-mismatch classification task:** It is a binary classification paradigm associating the EEG and speech segments. The EEG segment (**E**) and corresponding attended stimulus sentence (**S**+) form the positive pair while the same EEG and corresponding unattended sentence (**S**−) form the negative pair. The similarity score computation is achieved using the model depicted in Figure 4.4. Here, C.E. denotes the cross entropy loss.

In the dichotic listening task, our goal is to determine the specific speech sound on which the subject focused their attention. We approach this by formulating the detection of the attended speech sound as a match-mismatch (MM) task. Figure 4.2 illustrates this paradigm in detail. In this task, the segment of speech that the subject paid attention to is considered the match segment, while the segment of unattended speech played to the other ear is considered the mismatch segment. This task is more challenging than MM tasks on the natural speech dataset. The behavioural experiment that followed the listening experiment showed that the subjects were able to comprehend and understand the sound they attended to while having difficulty answering questions about the unattended speech.

4.4.7 Model architecture

We employed different modelling paradigms to analyze the encoding of acoustic and semantic features in EEG signals.

4.4.7.1 Baseline Model for Natural Speech Listening Task

Recently, Monesi et al. [1] showed that convolutional neural network (CNN) and long short-term memory (LSTM) based architectures outperform linear models for modelling the relationship between EEG and speech. This work employed a match mismatch classification task on fixed duration windows of speech and their corresponding EEG data. The work also demonstrated that mel spectrogram features of the speech stimulus provide the best neural tracking performance compared to other representations like speech envelope, word embedding, voice activity and phoneme identity [181]. They have performed the match mismatch task of 5s duration segments with 90% overlap between successive frames. The prior works [1, 181] use an angular distance between EEG and speech representations, average pooling over time, and a sigmoid operation. The model is trained with binary cross entropy loss [181]. We use this approach as the baseline setup for the proposed framework.

4.4.7.2 Acoustic Encoding

The speech signal representation $\mathbf{S}$ is the mel-spectrogram of dimension $28 \times T$, where T denotes the duration of a speech sentence at 64Hz. Similarly, the EEG data for the same sentence is denoted as $\mathbf{E}$, and it is of dimension $128 \times T$.

Both the speech and the EEG features are processed through a parallel neural pipeline, as depicted in Figure 4.3, without any weight sharing. This sub-network consists of a series of convolutional layers and LSTM layers. The convolutional layers implement 1-D and 2-D convolutions with 1×8 and 16×9 kernel sizes, respectively. The 1-D and 2-D layers have 8 and 16 kernels, respectively. Further, the 2-D CNN layers also introduce a stride of $(1, 3)$ to further down-sample the feature maps.

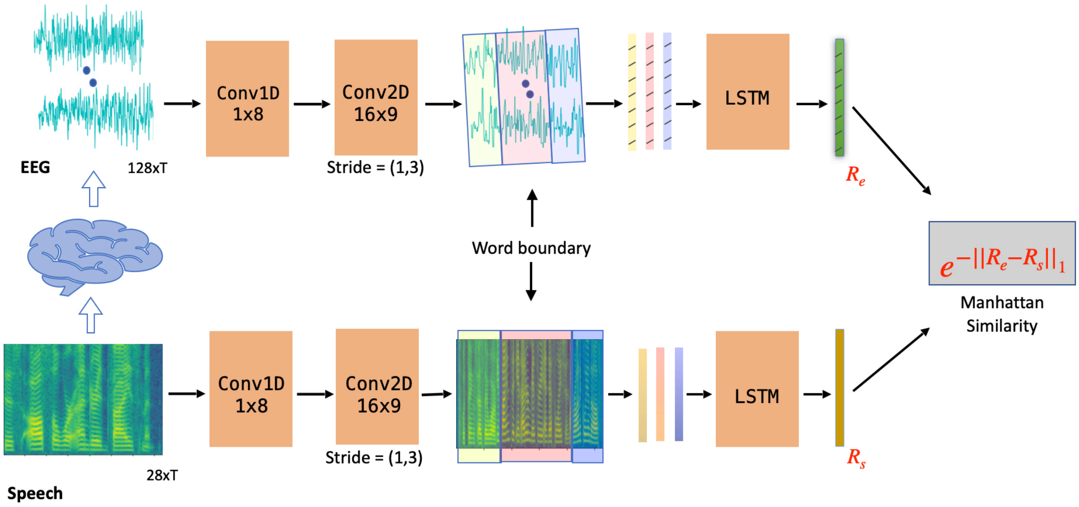


Figure 4.3: Proposed model for match mismatch task on speech EEG data. The model training paradigm is outlined in Figure 4.1.

The word boundary information available in the dataset is converted to the equivalent sampling rate (both EEG and audio representations at $\frac{64}{3}$ Hz). The audio and EEG feature maps are average pooled at the word level using the word boundary information. As a result, for a given sentence, the EEG and speech branches generate vector representations sampled at the word level. An LSTM layer models the inter-word context from these representations. This layer is included in both the stimulus (speech) and response (EEG) pathways. The last hidden state of the LSTM layer, of dimension 32, is used as the embedding for the stimulus/response, denoted as R_s/R_e, respectively.

We propose the Manhattan distance between the stimulus and response embeddings [182]. The similarity score is computed as,

$$d(\mathbf{E}, \mathbf{S}) = \exp(-||R_e - R_s||_1) \tag{4.1}$$

The similarity score for the matched pair $(\mathbf{E}, \mathbf{S}^+)$ and mismatched pair $(\mathbf{E}, \mathbf{S}^-)$ are computed. The model, with a dropout factor of 0.2, is trained using a binary cross-entropy loss, with $[d(\mathbf{E}, \mathbf{S}^+), d(\mathbf{E}, \mathbf{S}^-)]$ mapped to [1, 0] targets.

This speech-EEG network can be used for any task with a speech spectrogram as the input. We modify the speech subnetwork to use the envelope as the stimulus feature, as shown in Figure 4.4.

4.4.7.3 Semantic Encoding

To compare the encoding of text features with speech features, we used a text-EEG match mismatch classification task in a similar fashion. We used the text-EEG model sub-network (bottom row) shown in Figure 4.4 to determine whether the input text sentence matches the given EEG response. The EEG subnetwork resembles the acoustic encoding model, while the text subnetwork consists of a two-layer LSTM. Convolutional layers were not used since the word2vec feature does not contain local information. The similarity scoring component of the network is similar to the acoustic encoding model.

4.4.7.4 Joint Encoding of Acoustics and Semantics

After exploring the effect of acoustic and semantic features individually, we jointly trained the MM model with both features. The subnetworks are combined by computing the sum of losses as shown in Figure 4.4. The combination of the sum of losses ensures that the loss of both text-EEG and acoustics-EEG pairs are reduced individually. The speech network uses envelope as the acoustic input, and word2vec features are employed as semantic features for the text subnetwork.

4.4.7.5 Training and Evaluation Setup

We report the average results of multi-fold cross-validation, with classification accuracy as the metric. The experiments are run with a batch size of 32. The models are trained using Adam optimizer with a learning rate of 0.001 and weight decay parameter of 0.0001. The models are learned with a binary cross-entropy loss.

This study employed two model training scenarios: subject-dependent (SD) and subject-independent (SI). In the SD scenario, all subjects' data were present in both the training and

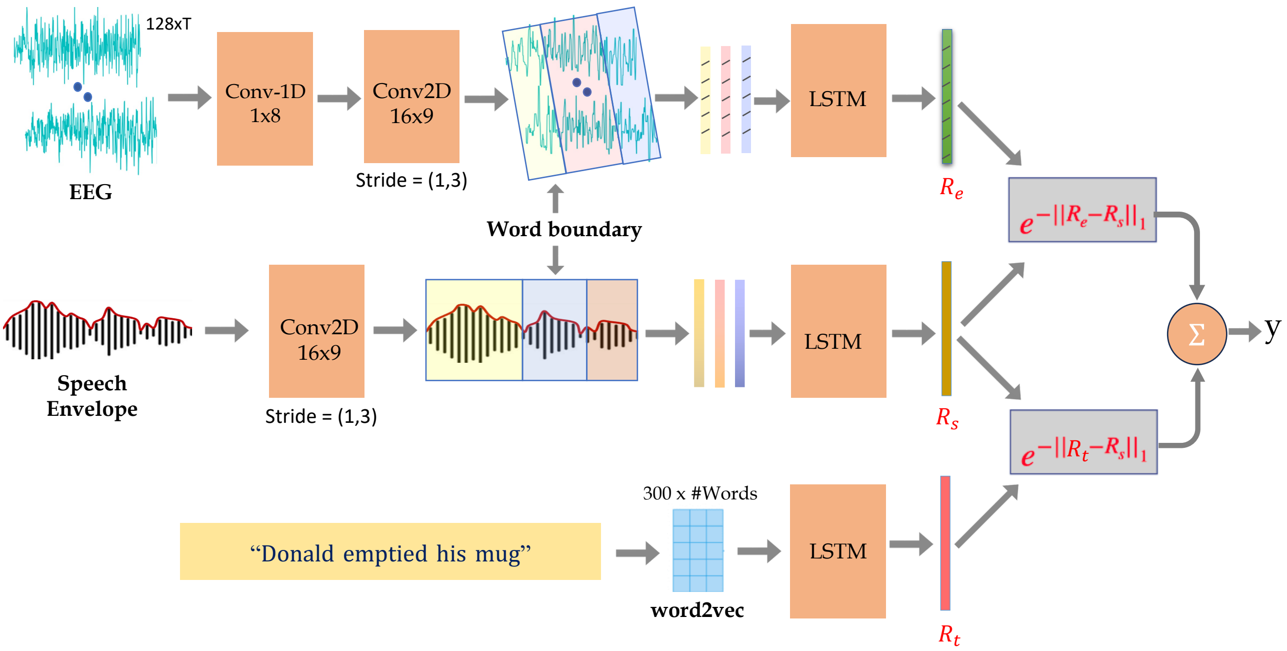


Figure 4.4: Proposed multi-modality model for match mismatch task on speech EEG data. The model training paradigm is outlined in Figure 4.1 for natural speech and in Figure 4.2 for dichotic listening task.

test sets. Conversely, in the SI scenario, the model was evaluated on unseen subjects. For the SI scenario, data from a subset of subjects were kept aside to form the test set, while data from the remaining subjects formed the training set.

In the SD scenario, we adopted the following strategy to create different training and test sets for cross-validation. A subset of trials from all subjects constituted the training set, while the remaining trials of all subjects served as the test data. For each cross-validation fold, we randomly chose 3 trials to form the test data, resulting in 6-fold cross-validation for the natural speech dataset and 10-fold cross-validation for the cocktail party dataset. The difference in the number of cross-validation folds is attributed to the distinct number of stimulus trials in each dataset: the NS dataset contains 20 stimulus trials, while the dichotic listening dataset contains 30 trials.

In the SI scenario, three subjects were randomly selected to form the test set in each fold, and the data from the remaining subjects constituted the training set for that fold. For the natural speech dataset, this resulted in 6-fold cross-validation, while the cocktail party dataset underwent 10-fold cross-validation. The variation in the number of folds is due to the difference in the number of subjects in each dataset: the NS dataset has 19 subjects, and the dichotic dataset has 33 subjects.

The average classification accuracy across cross-validation is reported in the subsequent results section.Please note that Table 4.2 displays the results of a 3-fold cross-validation, chosen as a representative out of the possible 10-fold cross-validation for the NS dataset.

To assess statistical significance, we employed the Wilcoxon signed-rank test [183] to compare the two models in this study. We utilized the predicted probabilities of the two models as samples for comparison, resulting in the number of samples for the significance test being equal to the number of test samples. The same test samples are used to evaluate both models in each fold. The samples from all folds are then aggregated to form the input for the signed-rank test.

4.5 Results

4.5.1 Natural Speech Listening Condition

4.5.1.1 Baseline model on fixed duration segments.

The baseline implementation for comparison is the work reported in Monesi et al. [1]. This architecture is an LSTM model that operates on fixed-duration audio EEG data. All experiments are run for 20 epochs of training. The result of the model with fixed duration frames is given in Table 4.1. In order to increase the amount of training data, we also use 90 % overlap between segments.

4.5.1.2 Baseline model at the sentence level

The baseline model architecture is implemented for fixed-duration segments in training and testing for NS dataset. In order to operate at the sentence level with variable length segments,

Table 4.1: Match mismatch classification accuracy of baseline model [1] for fixed duration sequences under natural speech listening condition. The step size between adjacent frames is 0.5s. The model training was carried out in subject-independent setting with 3-fold cross validation.

Frame Width (sec.)	**Test Accuracy (%)**
1	62.21
3	72.41
5	76.12

Table 4.2: The match-mismatch classification accuracy of speech stimulus and its EEG responses in sentence level for baseline [1] and the proposed model, under natural speech listening condition. Here, a random speech sentence was chosen as the mismatch sample for each EEG sentence. The model training was carried out in subject-independent setting with 3-fold cross validation.

Test Set	**Baseline Model**	**Proposed Model**		
		Cos.	**Euclidean**	**Manhattan**
Fold 1	65.39	88.22	93.49	94.02
Fold 2	65.32	88.73	93.68	94.00
Fold 3	64.98	86.54	93.72	93.91
Average	65.23	87.83	93.63	**93.97**

we have modified the dot product operation as element-wise multiplication followed by an average pooling. This score is passed through the sigmoid function, and the model is learned on sentence-level audio-EEG pairs. For the mismatch condition, a random speech spectrogram is paired with the EEG to generate the score. These results are reported in Table 4.2.

4.5.1.3 Proposed model with sentence level processing

The results with the proposed model are also reported in Table 4.2. We compare three different similarity scoring approaches, i) Angular (Cosine) similarity, ii) Negative L2 distance (Euclidean) and iii) proposed Manhattan similarity (Eq. 4.1). As seen in the results, the Euclidean and Manhattan similarity improves over the cosine similarity. The proposed EEG-speech match-mismatch classifier model achieves an average accuracy of 93.97%, which is statistically significantly higher than the baseline model's sentence-level performance (Wilcoxon

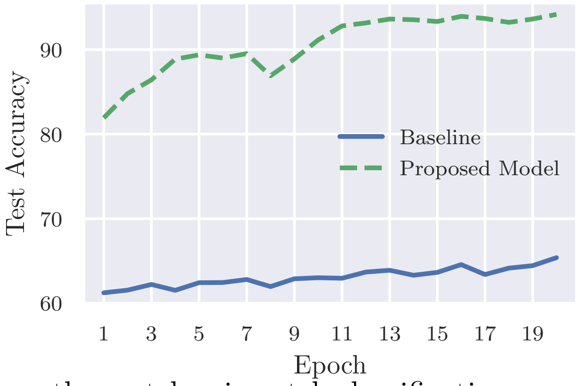


Figure 4.5: This figure shows the match-mismatch classification accuracy of the proposed model for test fold-1 as a function of the training epoch for the baseline model and the proposed approach under natural speech listening condition. The model training was carried out in subject-independent setting.

Table 4.3: Impact of mismatch sample selection strategy on classification accuracy under natural speech listening condition. The reported result is the average accuracy for 3-fold cross-validation with subject-independent training configuration.

Mismatch Selection Strategy	**Test Accuracy (%)**
Random Sentence	93.97
Next sentence	91.56

signed-rank test, $p < 1e - 4$). The epoch-wise accuracy for test fold-1 is also illustrated in Figure 4.5.

4.5.1.4 Mismatch sample selection for sentence processing

Previous match-mismatch EEG-speech studies [159, 1] dealt with fixed-duration speech and EEG segments. Cheveigne et al. [159] used an unrelated random segment as a mismatched sample, while studies like [1, 160, 181] employ a neighbouring segment as the mismatched sample. The sampling of the mismatched segments from the same trial ensures that the distribution of the matched and mismatched segments is similar. We explore a similar strategy for sentence-level analysis by selecting the neighbouring sentence in the same trial as the mismatched sample. Table 4.3 shows how the mismatch selection strategy affects the classification accuracy. The average accuracy has a slight degradation when the next sentence is used as the

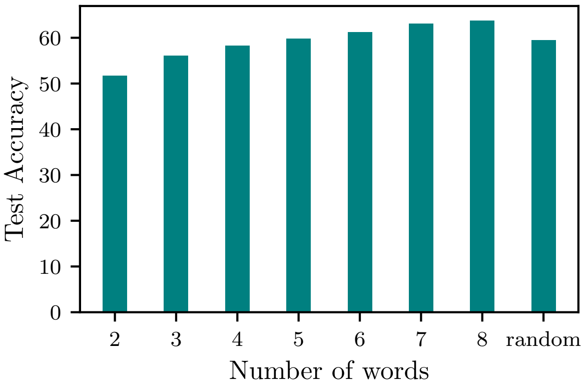


Figure 4.6: Average match-mismatch classification accuracy of the proposed model for random word boundaries,under NS listening condition. The reported result is the average accuracy of 3-fold cross-validation for SI training.

mismatch sample.

4.5.1.5 Importance of accurate word boundaries

We conducted several ablation tests to understand the impact of the word boundary information. The model is fed with random word boundaries in the first set of experiments. Each sentence is assumed to contain a fixed number of words and their boundaries are chosen at random. The results are reported in Figure 4.6. The accuracy improves gradually when the number of word boundaries is increased, even though they are random. The accuracy of the experiment using 8 words in a sentence is 64%, which is significantly lower than the model's performance with accurate boundary information (Wilcoxon signed-rank test, $p < 0.0001$). The final experiment shown in Figure 4.6 assumes a random number of words in each sentence with random boundaries, and it provided an accuracy of 60%.

In the second set of experiments, we provide accurate word boundary information but skip the word boundary information at every n-th word. These results are reported in Table 4.4. For example, Skip-3 in this table corresponds to removing the word boundary inputs at every 3-rd entry. The pooling is done with the rest of the available word boundaries for these experiments. As seen in Table 4.4, the results with a higher value of n (of skip-n experiments), approach the

Table 4.4: Accuracy (%) in match mismatch task for varying levels of word-boundary information ,under NS listening condition. The reported average result is the average accuracy of 3-fold cross-validation for SI training. Here, $Skip - n$ denotes removing every nth word boundary information in the model.

Test set	Skip-2	Skip-3	Skip-4	Skip-5
Fold 1	82.45	88.96	90.43	90.64
Fold 2	81.86	88.77	90.32	90.28
Fold 3	82.60	88.79	90.30	90.01
Average	82.30	88.84	90.35	90.31

Table 4.5: Average match-mismatch classification accuracy with speech envelope as stimulus feature for different evaluation configurations in two listening conditions. For the natural listening condition, a 6-fold cross-validation was performed, while the dichotic listening experiment undertook 10-fold cross-validation based on the data available in each dataset.

Listening Condition	**Subject Dependent**	**Subject Independent**
Natural	93.63	94.09
Dichotic	62.12	45.99

setting without any removal (accuracy of 93.97%). It is also noteworthy that, even with the Skip-2 setting (word boundary information available for every alternate word), the performance is 82.3%, significantly better than the baseline model. This study also demonstrates that accurate word boundary information significantly impacts the match mismatch classification, which further illustrates that the EEG signal encodes the word level tracking of speech.

4.5.2 Dichotic Listening Condition

4.5.2.1 Subject Dependence

Compared to the match-mismatch detection of natural speech under normal listening conditions, the task of auditory attention detection (AAD) using match-mismatch classification proves to be more challenging. In our case, the subject-independent setting did not yield satisfactory performance as shown in Table 4.5. Therefore, we opted for a subject-dependent configuration for the AAD task. All the further experiments mentioned in this chapter are

Table 4.6: Match-Mismatch classification accuracy of speech stimulus and EEG responses at the sentence level for various Listening Conditions in a subject-dependent training scenario. Average results for different feature modalities are reported in this table.

Listening Condition	**Speech Envelope**	**Text word2vec**	**Multi modality**
Natural	93.63	93.24	93.38
Dichotic	62.12	83.06	**84.60**

performed with subject-dependent configuration for both tasks unless otherwise stated.

4.5.2.2 Combined Impact of Acoustic and Semantic Cues

Table 4.6 presents the results of the match-mismatch classification accuracy between speech stimuli and their corresponding EEG responses, focusing on different listening conditions. We compared different stimulus feature cases in subject-dependent training scenarios.

In the dichotic listening condition, the unattended speech played to the other ear was used as the mismatch sample. The accuracy values reported in this table are the average accuracy obtained from multi-fold cross-validation. Specifically, the natural speech condition underwent a 6-fold cross-validation, while the dichotic listening condition employed a 10-fold cross-validation approach.

For natural speech conditions, both semantic and acoustic features result in similar accuracy (Wilcoxon signed rank test, $p = 0.03125$). While for dichotic listening, word2vec features result in significant improvement in accuracy (Wilcoxon signed rank test, $p = 0.00195$).

4.5.2.3 Comprehension Scores - A measure of attention

During the dichotic listening task, participants were required to answer multiple-choice questions after each trial to assess their comprehension and, consequently, their attentiveness to the played stories. They were asked questions about both stories. The scores obtained were normalized to a range of 0 to 1. Therefore, we made a decision to include only trials in which the participant achieved a comprehension score higher than 0.5, considering them as attended and using them for matched cases. In the case of a story that was intended to be unattended, if its comprehension score exceeded 0.5, it indicated that the participant was paying attention to the story. Consequently, that trial was excluded from our analysis. Figure 4.7 shows the percentage reduction in the number of trials after removing unattended trials. It shows that there is only an average reduction of about 6.7% in the number of trials across all classes.

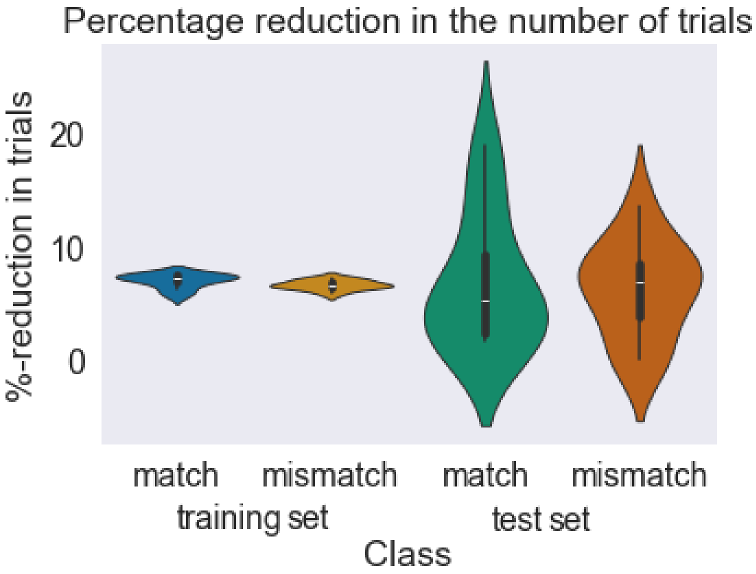


Figure 4.7: This violin plot shows the percentage reduction in the number of trials after the removal of unattended trials.

In Table 4.7, the first row shows the baseline performance, when none of the trials were disregarded based on the behavioural score. Subsequently, two experiments were conducted. In the first experiment, trials were eliminated from both the training and test datasets using the behavioural score. In the second experiment, the behavioural score was employed to exclude unattended trials solely from the training data, while all trials in the test dataset were evaluated regardless of the subject's attentiveness.

The data presented in Table 4.7 demonstrates that filtering based on attention significantly affects the performance of the model when using the envelope as the stimulus feature (Wilcoxon signed rank test, $p = 0.03710$). However, this notable improvement is not observed when using

Table 4.7: Match-Mismatch classification accuracy of speech stimulus and EEG responses at the sentence level for the dichotic listening condition in a subject-dependent training scenario utilizing the comprehension score of subjects. We have Comprehension question scores for attended and unattended stories.If the comprehension score of a trial is less than 0.5 for the attended story, ignore that trial. Similarly, if the unattended score of a trial is greater than 0.5, ignore that trial from the mismatch data.

Comprehension Score based Trial-Filtering	**Speech Envelope**	**Text word2vec**	**Multi modality**
No Filtering	62.12	83.06	84.60
Both train and test data	66.48	83.86	84.61
Only on train data	66.68	84.27	83.19

the word2vec text feature or in the case of multi-modal feature input (Wilcoxon signed rank test, $p = 1.0$). The enhanced performance of the former case indicates the robustness of the word2vec features to minor variations in the test data. The model incorporating semantic features exhibits greater generalizability.

4.5.2.4 Importance of word boundary on Auditory Attention Detection

Word-level segmentation holds significant importance in speech perception during normal listening conditions[172]. To investigate the role of word segmentation in auditory attention detection, we conducted a comparison of match-mismatch classification performance using the proposed model with and without word boundary information in the dichotic listening condition.

As seen in the table, Table 4.8, word level segmentation plays an important role in auditory attention detection. Word boundary information has more impact on the semantic feature than the envelope feature.

4.5.2.5 Importance of accurate word boundaries

We conducted several ablation tests to understand the impact of the word boundary information. The model is fed with random word boundaries in the first set of experiments. Each

Table 4.8: Role of word boundary information in auditory attention detection. The table shows the match-mismatch classification accuracy of speech stimulus and its EEG responses at sentence level for dichotic listening condition in the subject-dependent training scenario. The model was trained with and without word boundary information for comparison. Here, an unattended speech sentence was chosen as the mismatch sample for each EEG sentence. (Multimodality: sum of losses)

Stimulus Feature	**Without word-boundary**	**With word-boundary**
Envelope	56.90	62.12
word2vec	64.06	83.06
Multi-modal	62.35	**84.60**

Table 4.9: Average match-mismatch classification accuracy for auditory attention detection with random word boundaries.

Number of words	**2**	**3**	**4**	**5**
Envelope	60.00	59.82	60.11	61.02
word2vec	80.36	79.57	80.64	80.15
Multi-modal	78.14	77.81	76.74	76.80

sentence is assumed to contain a fixed number of words, and their boundaries are chosen at random. The results are reported in Table 4.9. The accuracy improves only slightly when the number of word boundaries is increased, even though they are random.

In the second set of experiments, we provide accurate word boundary information but skip the word boundary information at every n-th word. These results are reported in Table 4.10. For example, Skip-3 in this table corresponds to removing the word boundary inputs at every 3-rd entry. The pooling is done with the rest of the available word boundaries for these experiments.

4.6 Discussion

In this study, we have attempted to validate the hypothesis that speech perception in the brain is segmented at the word level. For this task, we developed a deep neural network model

Table 4.10: Accuracy (%) in auditory attention detection task for varying levels of word-boundary information. Here, $Skip-n$ denotes removing every nth word boundary information in the model.

Stimulus Feature	**Skip-2**	**Skip-3**	**Skip-4**	**Skip-5**
Envelope	63.06	62.37	62.00	62.37
word2vec	83.97	84.12	82.97	83.19
Multi-modal	84	**85.33**	83.81	84.52

consisting of convolutional encoders, word-level aggregators and recurrent layers. A novel loss function for this task based on Manhattan similarity was also proposed.

The proposed model validated the hypothesis by improving the accuracy of match-mismatch classification of speech and EEG responses at the sentence level. The incorporation of word boundary information yields statistically significant improvements compared to the baseline model, demonstrating the importance of this information in the neural tracking of speech.

In the second study, we conducted a comparison of EEG-stimulus models between natural speech and dichotic listening conditions. We performed these experiments in a subject-dependent manner, as attempts to achieve subject-independent configuration for AAD detection with this dataset were not successful.

Our findings suggest that modelling EEG responses and stimuli yield superior performance during natural speech perception. Furthermore, we observed that the relative importance of textual and acoustic content is approximately similar in the natural listening condition. However, during dichotic listening, the human brain exhibits a prioritization of perceiving the semantics over the acoustic characteristics of the attended speech. Word boundary information emerges as a crucial component for speech perception during the dichotic listening task, although not all boundaries are as critical as in the case of natural speech.

4.6.1 Natural Speech: Role of Speech Envelope vs Semantic Feature

Table 4.6 displays the outcomes of the match-mismatch classification task conducted using two distinct stimulus features: speech envelope and word2vec features. The classification was done at the sentence level using data recorded under normal listening conditions. In this task, a random sentence was selected as the mismatch sample for each EEG sentence. The results do not show a major difference in performance between the two features (Wilcoxon signed rank test with $\alpha = 0.01$, $p = 0.03125$). Therefore, it can be concluded that both acoustic and semantic features hold comparable importance in speech perception during normal listening conditions.

4.6.2 Dichotic Listening: Role of Speech Envelope vs Semantic Feature

To assess the relative impact of acoustics and semantics in speech perception under more complex listening conditions, we conducted a match-mismatch classification for auditory attention detection at the sentence level. For this experiment, unattended speech was used as the mismatch sample. The model training followed a subject-dependent approach described in Section 4.4.7.5.

The results, presented in Table 4.6, demonstrate that the stimulus-response model using semantic features as input outperformed the model utilizing speech envelope features (Wilcoxon signed rank test, $p = 0.00195$). This indicates that semantic features are better entrained in the EEG response recorded during the dichotic listening experiment.

The findings suggest that semantics hold higher significance than acoustic traits for speech perception under complex listening conditions.

4.7 Chapter Summary

The key findings of this chapter are as follows:

- We validated the segmented nature of human speech perception. Incorporating word boundary information improves stimulus-response modelling accuracy for sentence-level processing.

- The proposed neural network can incorporate multiple modalities of stimulus data (speech and text) to correlate with the neural response (EEG).
- The information encoded in EEG is also related to the underlying semantic content (textual content) of the stimulus.
- The match-mismatch performance is enhanced by the combined use of low-level speech features along with textual features during different listening conditions.
- The human brain tends to prioritise assimilating semantic information over the acoustics while listening to speech in dichotic condition.

The results of these experiments can have various applications. They can help improve our understanding of attention disorders, such as attention deficit hyperactivity disorder (ADHD) or hearing impairments. Furthermore, they may contribute to the development of brain-computer interfaces (BCIs) that can decode and interpret a person's attention state in real time, leading to advancements in areas such as auditory prosthetics or assistive communication devices.

Chapter 5

Summary and Future Perspectives

5.1 Key contributions of the thesis

5.1.1 Isolated Word Learning

The study focused on the neural correlates of word discrimination in adults learning a new language. Repeated exposure to words from an unfamiliar language led to the formation of a consistent neural representation, aligning with language learning as indicated by pronunciation rating. The correlation between audio stimuli and the EEG envelope was stronger for Japanese than English trials, suggesting differential auditory information processing during word discrimination. These findings suggest that as individuals learn a new language, their neural systems develop specific patterns of activation that contribute to the accurate discrimination of words. It implies that neural plasticity plays a crucial role in language acquisition, facilitating the formation of neural representations that align with the process of learning and discriminating words.

5.1.2 Learning Words in Context

The effects of semantic incongruity were investigated in a proficient language (English) versus a newly acquired language (Japanese). Rapid semantic learning in the newly acquired language elicited delayed positive responses in event-related potentials (ERPs) at 500-700ms from the onset of the end-word for incongruent words. The delayed positive ERP responses observed for incongruent words in the newly acquired language indicate the engagement of semantic processing mechanisms during language learning. It suggests that the brain rapidly adapts to new semantic information and establishes distinct neural signatures, facilitating the process of acquiring and differentiating meanings in a new language. Scalp electrodes showing these semantic effects were predominantly located in the parietal and occipital regions. These specific brain areas are engaged in the integration of semantic information during language acquisition.

5.1.3 EEG Decoding of Continuous Speech

A match-mismatch classification model incorporating word boundary information was introduced for speech EEG matching tasks. The model utilized a loss function based on the Manhattan distance, resulting in improved classification performance compared to previous approaches. Experimental illustrations highlighted the model's effectiveness in capturing and utilizing word boundary information. Ablation experiments provided insights into the specific impact of word boundary information on speech EEG matching tasks, emphasizing its relevance in neural processing. This finding highlights the importance of temporal cues and linguistic segmentation in understanding the neural mechanisms underlying speech comprehension.

A paradigm for auditory stimulus-response modelling was developed in monoaural and dichotic listening contexts. Including word boundaries enhanced the match-mismatch performance of audio signals during monoaural natural speech listening. In this listening condition, textual information alone produced comparable results to audio signals, emphasizing the significance of linguistic cues. In the dichotic speech listening task, EEG signals exhibited a higher match-mismatch performance for text data than the audio signal, indicating a prominent encoding of higher-level semantic information over acoustic envelope information. A multimodal architecture for the match-mismatch task demonstrated improved performance over individual modalities, suggesting that EEG signals encode semantic and acoustic content jointly. This finding supports the hypothesis that semantic processing plays a crucial role in speech perception and indicates the significance of considering both linguistic and acoustic factors in understanding the neural mechanisms underlying language comprehension.

Overall, these findings have broad implications for our understanding of the neural processes involved in word discrimination, language learning, semantic processing, and speech perception. They provide valuable insights into the plasticity of the brain during language acquisition, the role of different neural regions in processing semantic information, the integration of auditory and visual cues in audio-visual correspondence learning, the importance of word boundary information in speech processing, and the contribution of semantic content to neural responses during speech perception.

5.2 Limitations

While this research has provided valuable insights into the neural processes underlying word discrimination, language learning, and speech perception, it is important to acknowledge the limitations of this work. These limitations should be considered when interpreting the findings and assessing their validity and generalizability.

The word-learning experiments primarily focused on a specific language-learning context and language pairs (e.g., English and Japanese). The findings may be influenced by the unique characteristics of these languages and the specific learning experiences of the participants. It is important to recognize that different languages and language pairs may exhibit variations in neural processing and learning outcomes. Therefore, caution should be exercised when extrapolating the results to other languages or language learning scenarios.

Using EEG signals as the primary measure of neural activity has limitations in spatial resolution compared to other neuroimaging techniques. Combining EEG with other imaging modalities, such as fMRI, could provide a more comprehensive understanding.

5.2.1 Isolated Word Learning

Firstly, the sample size used in this study may have implications for generalisability. A larger and more diverse sample size would enhance the external validity of the findings and allow for a better understanding of how these processes unfold across different individuals and demographic groups.

Learning involves memory and medium to long-term analysis as well. This was not performed in this study. The isolated word study discussed in chapter 2 did not involve the semantics of the novel words. Especially for adults, learning is mostly associated with semantics. This study did not consider the multi-modal aspects of learning.

Additionally, we did not evaluate the effects of repetitive suppression. In the past, studies using MEG signals have shown that there are two major effects seen in the brain when the same words are presented repeatedly. In Repetitive Enhancement (RE), the frontal regions in the brain get activated when the same word from an unknown language is presented to the subject multiple times [184], after which the activations drop, leading to Repetitive Suppression (RS). The RS is also observed when a word familiar to the subject is presented. These studies indicate that activations are seen till new brain connections are formed, after which the intensity of the activations drops.

5.2.2 Learning Words in Context

We introduced the meaning of novel words with visual modality, but the analysis was done with a code-switching setup. This work did not study the effects of visual vs auditory introduction of semantics. We did not analyse the medium to long-term effects of learning. The ERP effects shown in chapter 3 were averaged over all subjects and words. We did not perform a single-trial analysis on this data.

5.2.3 EEG Decoding of Continuous Speech

The EEG analysis (chapter 4) was primarily done with a match-mismatch task. This may not reveal all the nuances of EEG encoding. Verifying the results with other correlation/decoding models might further establish the claims.

5.3 Future Directions

Based on this study's findings, several potential future research areas can further enhance our understanding of the neural processes underlying word discrimination, language learning, semantic processing, and speech perception. We identify some possible future extensions:

1. Longitudinal Studies: Conduct longitudinal studies to track the neural changes over time during language acquisition. Examining how neural representations evolve as individuals progress in their language-learning experience can provide valuable insights into the dynamic nature of language processing and the neural plasticity involved.

2. Cross-Linguistic Comparisons: Expand the investigation to include a broader range of languages and compare the neural processes involved in word discrimination and language learning across different linguistic contexts. This can help identify language-specific effects and generalize the findings to a diverse range of language learners.

3. Neural Connectivity: Investigate the functional connectivity patterns between different brain regions involved in word discrimination, language learning, and semantic processing using fMRI. Analyzing the interactions and communication between brain regions can provide a more comprehensive understanding of the neural networks underlying these processes.

4. Individual Differences: Explore individual differences in language learning and their relation to neural processes. Investigate factors such as age, aptitude, and prior language

experience to determine how these variables influence the neural correlates of word discrimination, language learning, and semantic processing. Especially the sensitivity of P600 magnitude for various variables like language proficiency can be investigated.

5. Patient population: Conducting the repetitions of sounds study (Chapter 2) and the ERP study on short-term learning with context (Chapter 3) among patients with language disorders can offer valuable insights into the existence or absence of neural markers previously identified in a healthy adult population. These findings can be used as a metric for identifying individuals with language disorders.

6. Multi-modal Approaches: Further explore the integration of multiple modalities, such as incorporating visual or gestural cues along with auditory stimuli, to investigate the role of multimodal information in language acquisition and speech perception. This can help unravel how different sensory inputs contribute to neural processing and facilitate language learning.

7. Naturalistic Language Learning: Conduct studies in more ecologically valid settings to investigate word discrimination and language learning in real-life language learning contexts. Observing neural processes during naturalistic language learning situations can provide insights into how the brain adapts to authentic language input and facilitates comprehension.

8. Big data-based models: It would be advantageous to train the stimulus-response models introduced in chapter 4 using larger speech-EEG datasets and enhanced machine learning architectures.

5.4 Impact of this Thesis

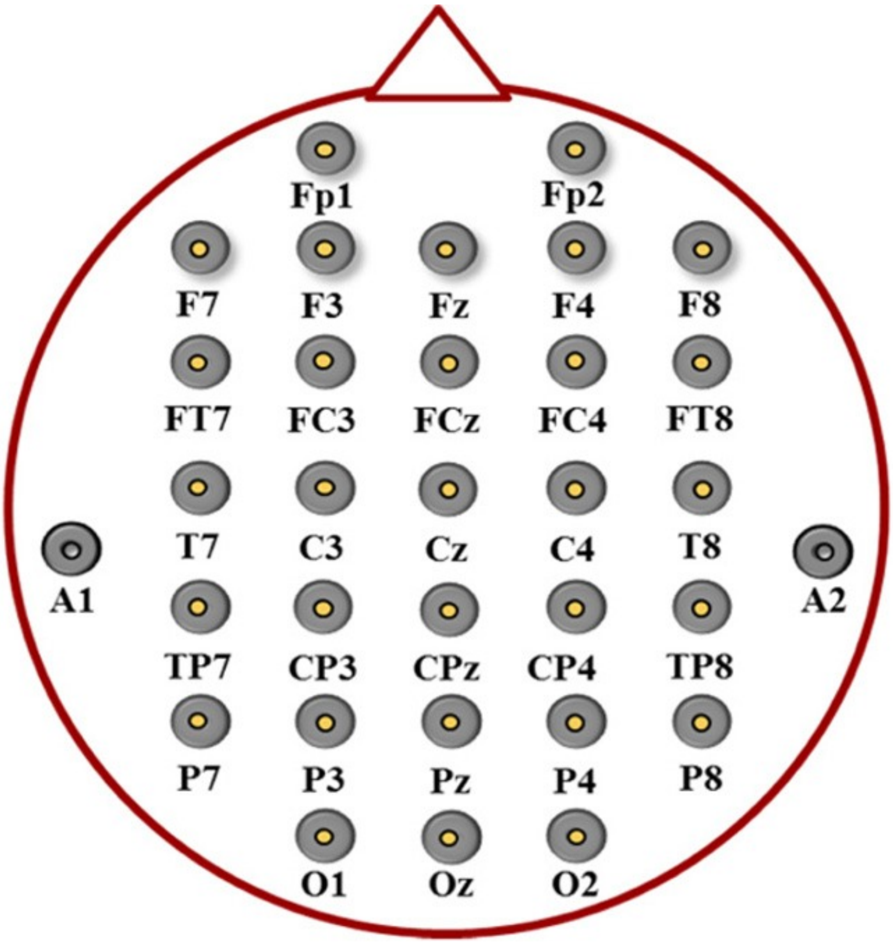
The potential impact of this research extends beyond theoretical advancements. The insights gained from this study can inform language education and intervention strategies, aiding individuals in language learning. Educators can design effective teaching methods tailored to individual learners' needs by recognising the neural mechanisms underlying word discrimination and language acquisition. Furthermore, the findings can contribute to developing neurofeedback-based interventions and assistive technologies for individuals with language-related difficulties.

Insights from this thesis can also contribute to technology development. Findings from chapter 2 can pave the way for the development of Computer Assisted Language Learning (CALL) technologies. One example is the pronunciation scoring technique discussed. The insights gained from chapter 4, understanding the underlying auditory attention mechanism of humans, can contribute to the development of assistive technologies for individuals with speech or language impairments.

Investigating language learning and speech perception at the neural level opens doors to further exploration and understanding of human communication's intricate processes. It challenges us to consider the complex interplay between language, cognition, and the brain. By continuing to delve into these areas, we can unlock new insights and pave the way for future advancements in language technologies.

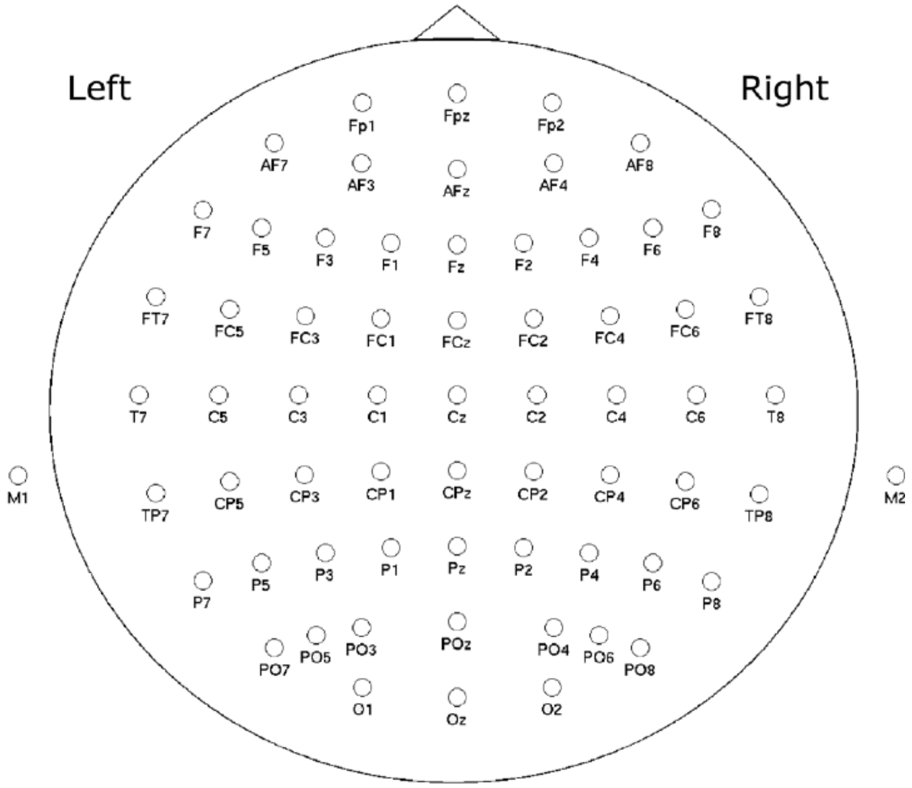
Appendix

Electrode Layout with 10-20 enhanced montage



Electrode placements of 32 channels according to the international 10‑20 system. This is the electrode layout of the EEG cap used for data recording in Chapter 2.

64 channel Electrode Layout



Electrode placements of 64 channel electrode cap used for data recording of Chapter 3.

Stimuli Set of Semantic Learning Experiment

List of highly predictable English sentences and end words in different conditions used in the experiment discussed in chapter 3.

Sentences and End words				
Sentence Prefix	English congruent end word	English incongruent	Japanese congruent	Japanese incongruent
She made the bed with clean	bedsheets	swan	mokkori	maggu
My T.V has a fifty-inch	screen	van	gamen	komugi
The beer drinkers raised their	mugs	hen	maggu	mokkori
Harry could see the blooming	flowers	rope	hanabana	senro
The bread was made from whole	wheat	garage	komugi	hon
I made a phone call from a	booth	bowl	būsu	chizu
A termite looks like an	ant	goalpost	ari	kēki
For your birthday I baked a	cake	map	kēki	ami
Kevin went to the library to read	books	calf	hon	kouzui
The fruit was shipped in wooden	box	tablets	hako	hachinosu
We're lost so let's look at the	map	spoon	mappu	ari
Household goods are moved in a	van	beehive	kasha	naifu
The honey bees swarmed round the	beehive	thorns	hachinosu	besuto
I cut my finger with a	knife	screen	naifu	nedoko
The candle flame melted the	wax	clock	rou	toge
This key won't fit in the	lock	pool	jou	suwan
The baby slept in the	crib	lock	nedoko	rou
A rose bush has prickly	thorns	pie	toge	koromo
Ruth poured the water down the	sink	ant	nagashi	genkotsu
The cop wore a bullet-proof	vest	jar	besuto	taiko
After his bath he wore a	robe	bomb	koromo	shippu
The soup was served in a	bowl	cot	bouru	kesshouten
They marched to the beat of the	drum	cloth	doramu	nuno
The sailor cleaned the deck of the	ship	broom	senpaku	nawa

Continuation of Table 5.1

Sentence Prefix	English congruent end word	English incongruent	Japanese congruent	Japanese incongruent
They played a game of cat and	mouse	truck	nezumi	torakku
We shipped the furniture by	truck	booth	torakku	supūn
He tossed the drowning man a	rope	crib	rōpu	suwan
Stir your coffee with a	spoon	mouse	supūn	kuraun
At breakfast he drank apple	juice	gown	jūsu	bouru
He hit me with a clenched	fist	police	genkotsu	oin
The king wore a golden	crown	drum	kuraun	houki
The duck swam with the white	swan	ant	suwan	shako
Let's decide by tossing a	coin	crown	koin	būsu
The girl swept the floor with a	broom	truck	houki	tokei
We heard the ticking of the	clock	mugs	tokei	yakuzai
The doctor prescribed the	tablets	sling	yakuzai	medaru
Unlock the door and turn the	knob	flowers	nobu	sankakukin
Her entry should win a	medal	lock	medaru	torappu
The mouse was caught in the	trap	ship	torappu	nedoko
The house was robbed by a	thief	mop	dorobou	benchi
Wash the floor with a	mop	coin	moppu	tento
Harry slept on the folding	cot	medal	nedoko	mendori
The man is sitting on the	bench	mold	benchi	kouzui
The heavy rains caused a	flood	bench	kouzui	kaeru
The chicks followed the mother	hen	mat	mendori	ushi
We camped out in our	tent	juice	tento	hitsuji
The pond was full of croaking	frogs	wheels	kaeru	hato
The shepherd watched his flock of	sheep	fist	hitsuji	mune
The swimmer dove into the	pool	cap	pu-ru	gaun
The cigarette smoke filled his	lungs	bread	mune	ryourin
The bride wore a white	gown	wheat	koromo	koushi
We swam at the beach during high	tide	trap	shio	satsu

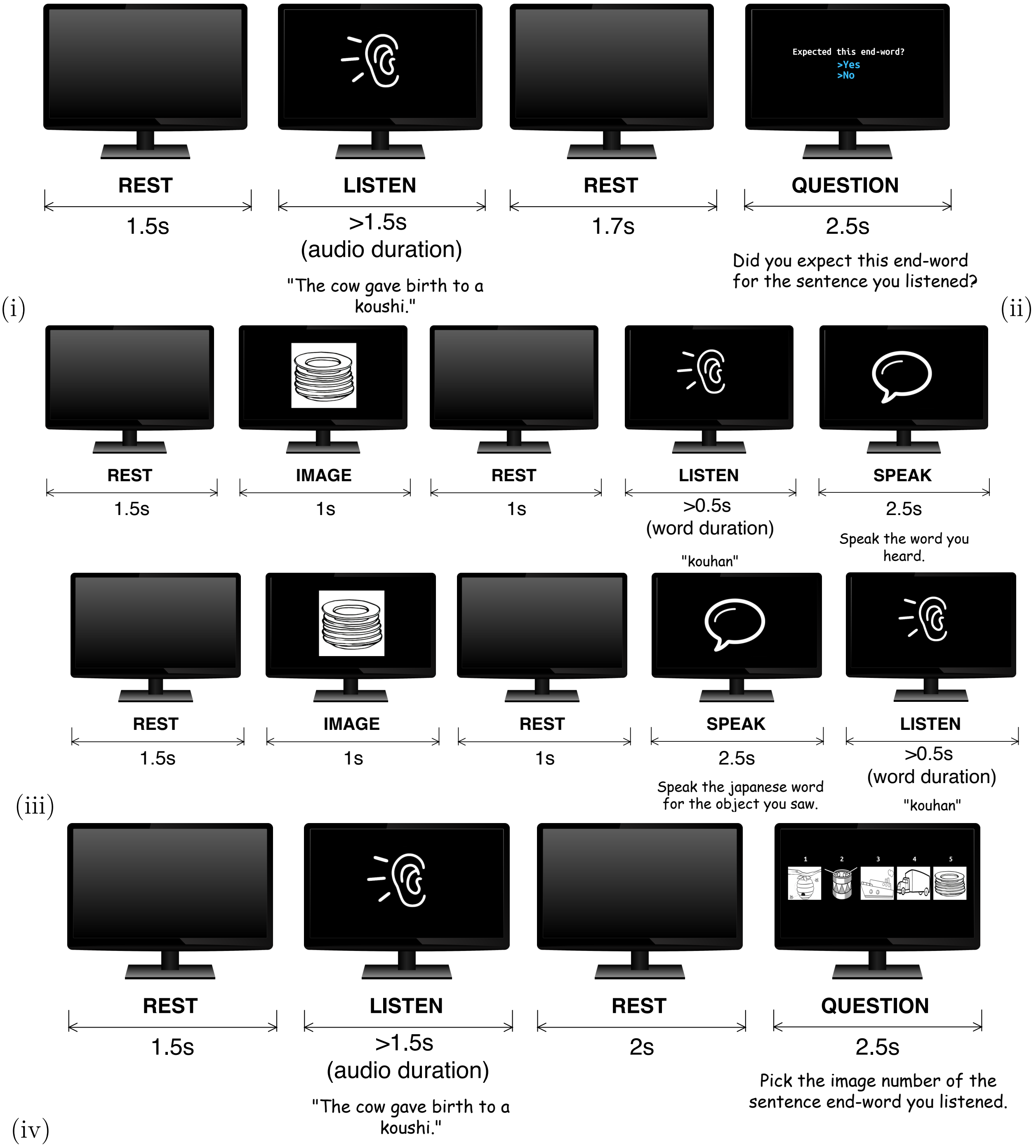
Continuation of Table 5.1

Sentence Prefix	English congruent end word	English incongruent	Japanese congruent	Japanese incongruent
A bicycle has two	wheels	bedsheets	ryourin	kyappu
The cow gave birth to a	calf	tent	koushi	pai
Paul was arrested by the	police	track	satsu	matto
On a sunny day she wore a	cap	dove	kyappu	jūsu
For dessert he had apple	pie	knob	pai	dorobou
Please wipe your feet on the	mat	ox	matto	nobu
When she got out of the car she closed the	Door	stars	doa	sakana
He mailed the letter without a	Stamp	milk	kitte	miruku
In the shower he washed his face with	Soap	plates	sekken	kyoukai
After every meal it’ s good to brush your	Teeth	gifts	hanarabi	isu
He brought his bait to the lake to catch	Fish	pants	sakana	cha
Joan fed her baby some warm	Milk	eyes	miruku	taiyou
Every Sunday the family goes to	Church	bank	kyoukai	shita
The man happily sat down in the comfortable	Chair	socks	isu	ushi
He liked lemon and sugar in his	Tea	shark	cha	shuzu
The player’ s cap protected him from the	Sun	nest	taiyou	yubiwa
While eating Steve accidentally bit his	Tongue	fire	shita	kagi
The farmer spend the morning milking his	Cow	yarn	ushi	hikouki
Susan put on the socks and	Shoes	bag	kutsu	jumoku
Bob proposed and gave her a diamond	Ring	kite	yubiwa	hoshiboshi
Carolyn couldn’ t start her car without the right	Keys	door	kagi	
Tim joined the Airforce as he always wanted to fly an	Aeroplane	stamp	hikouki	kaban
To learn about their ancestors they drew a family	Tree	fish	jumoku	sokkusu
In the night sky it is easier to see all the	Stars	church	hoshiboshi	omeme
It was windy enough to fly a	Kite	tea	kaito	ginkou

Continuation of Table 5.1				
Sentence Prefix	English congruent end word	English incongruent	Japanese congruent	Japanese incongruent
The thief ran by and snatched the lady’ s	Bag	tongue	kaban	doa
Derek’ s feet were cold, so he put on some	Socks	chair	sokkusu	kouhan
Without her sunglasses the sun hurt Annie’s	Eyes	cow	omeme	zubon
He deposited his new paycheck at the	Bank	ring	ginkou	sekken
On her birthday she excitedly opened the	Gifts	shoes	purezento	nesuto
After dinner the maid collected the family’ s	Plates	sun	kouhan	kaji
Sid needed a belt to hold up his	Pants	fish	zubon	neiru
The birds lay the eggs in the	Nest	teeth	nesuto	purezentsu
Dan gathered more wood for the	Fire	aeroplane	kaji	kitte
We had a candle-light dinner in a lakeview	restaurant	camera	restoran	pasupoto
Mary missed her company bus, so she took a	taxi	passport	takushi	restoran
She clicked the picture with a	camera	taxi	kamera	takushi
Travelling to foreign land needs a visa and	passport	restaurant	pasupoto	kamera

Illustration of different phases of the semantic learning experiment

The different stages of the experiment of chapter 3 is shown pictorially.(i) Listening Session, (ii) Learning Session, (iii) Recall Session, and (iv) Test Session. These stages are in serial order for a particular word. But as a whole, stages for different words are interspersed and randomized. (Sources of Images used in the experiment: ©*pixabay.com, unsplash.com, pexels.com*).


Expected this end-word?
>Yes
>No
REST
1.5s
LISTEN
>1.5s
(audio duration)
"The cow gave birth to a koushi."
REST
1.7s
QUESTION
2.5s
Did you expect this end-word for the sentence you listened?
(i)
(ii)
REST
1.5s
IMAGE
1s
REST
1s
LISTEN
>0.5s
(word duration)
"kouhan"
SPEAK
2.5s
Speak the word you heard.
REST
1.5s
IMAGE
1s
REST
1s
SPEAK
2.5s
Speak the japanese word for the object you saw.
LISTEN
>0.5s
(word duration)
"kouhan"
(iii)
1
2
3
4
5
REST
1.5s
LISTEN
>1.5s
(audio duration)
"The cow gave birth to a koushi."
REST
2s
QUESTION
2.5s
Pick the image number of the sentence end-word you listened.
(iv)

Bibliography

[1] Mohammad Jalilpour Monesi, Bernd Accou, Jair Montoya-Martinez, Tom Francart, and Hugo Van Hamme. An LSTM based architecture to relate speech stimulus to EEG. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 941–945. IEEE, 2020. xix, 3, 78, 86, 88, 92, 93, 94

[2] Coralie Pallier, Stanislas Dehaene, J.B Poline, D Lebihan, A.M Argenti, Emmanuel Dupoux, and J. Mehler. Brain imaging of language plasticity in adopted adults: can a second language replace the first? *Cerebral cortex*, 13(2):155–61, 2003. 1, 21, 23

[3] Jubin Abutalebi, Stefano F Cappa, and Daniela Perani. What can functional neuroimaging tell us about the bilingual brain. *Handbook of bilingualism: Psycholinguistic approaches*, pages 497–515, 2005. 2

[4] Albert Costa and Núria Sebastián-Gallés. How does the bilingual experience sculpt the brain? *Nature reviews neuroscience*, 15(5):336–345, 2014. 2

[5] Peter Indefrey. A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language learning*, 56:279–304, 2006. 2

[6] Rosa M Manchón. The psycholinguistics of second language writing. In *The Routledge handbook of second language acquisition and psycholinguistics*, pages 400–412. Routledge, 2022. 2

[7] Janet G Van Hell and Natasha Tokowicz. Event-related brain potentials and second language learning: Syntactic processing in late l2 learners at different l2 proficiency levels. *Second Language Research*, 26(1):43–74, 2010. 2

[8] Saloni Krishnan, Kate E Watkins, and Dorothy VM Bishop. Neurobiological basis of language learning difficulties. *Trends in cognitive sciences*, 20(9):701–714, 2016. 2

[9] Kamakshi V Gopal, Erin C Schafer, Lauren Mathews, Rajesh Nandy, Derek Beaudoin, Laura Schadt, Ashley Brown, Bryce Phillips, and Joshua Caldwell. Effects of auditory training on electrophysiological measures in individuals with autism spectrum disorder. *Journal of the American Academy of Audiology*, 31(02):096–104, 2020. 2

[10] Tilde Van Hirtum, Pol Ghesquiere, and Jan Wouters. A bridge over troubled listening: Improving speech-in-noise perception by children with dyslexia. *Journal of the Association for Research in Otolaryngology*, 22:465–480, 2021. 2

[11] R Holly Fitch and Paula Tallal. Neural mechanisms of language-based learning impairments: insights from human populations and animal models. *Behavioral and Cognitive Neuroscience Reviews*, 2(3):155–178, 2003. 2

[12] Jessica Defenderfer, Samuel Forbes, Sobanawartiny Wijeakumar, Mark Hedrick, Patrick Plyler, and Aaron T Buss. Frontotemporal activation differs between perception of simulated cochlear implant speech and speech in background noise: An image-based fnirs study. *Neuroimage*, 240:118385, 2021. 2

[13] Cristen Olds, Luca Pollonini, Homer Abaya, Jannine Larky, Megan Loy, Heather Bortfeld, Michael S Beauchamp, and John S Oghalai. Cortical activation patterns correlate with speech understanding after cochlear implantation. *Ear and hearing*, 37(3):e160, 2016. 2

[14] Robert P Carlyon and Tobias Goehring. Cochlear implant research and development in

the twenty-first century: a critical update. *Journal of the Association for Research in Otolaryngology*, 22(5):481–508, 2021. 3

[15] Byeongwook Lee and Kwang-Hyun Cho. Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. *Scientific reports*, 6(1):37647, 2016. 3

[16] Jaswanth Reddy Katthi and Sriram Ganapathy. Deep correlation analysis for audio-eeg decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2742–2753, 2021. 3, 78

[17] Tobias de Taillez, Birger Kollmeier, and Bernd T Meyer. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *European Journal of Neuroscience*, 51(5):1234–1241, 2020. 3

[18] Nina F Dronkers, Odile Plaisant, Marie Therese Iba-Zizen, and Emmanuel A Cabanis. Paul Broca's historic cases: high resolution MR imaging of the brains of leborgne and lelong. *Brain*, 130(5):1432–1441, 2007. 4

[19] Iain DeWitt and Josef P Rauschecker. Wernicke' s area revisited: parallel streams and word processing. *Brain and language*, 127(2):181–191, 2013. 4

[20] SB Hong, J Park, Y Moon, ML Grandmason, D Nowroski, and M Moon. Learning a foreign language in adulthood using principles of neuroscience. *ARC Journal of Neuroscience*, 2(1):10–13, 2017. 4

[21] Stefka H Marinova-Todd, D Bradford Marshall, and Catherine E Snow. Three misconceptions about age and l2 learning. *TESOL quarterly*, 34(1):9–34, 2000. 4, 5

[22] Catherine E Snow and Marian Hoefnagel-Höhle. The critical period for language acquisition: Evidence from second language learning. *Child development*, pages 1114–1128, 1978. 4

[23] Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008. 5

[24] H Valerie Curran. Declarative and nondeclarative memory. *Encyclopedia of Psychopharmacology*, pages 1–7, 2014. 5

[25] Saeid Sanei and Jonathon A Chambers. *EEG Signal Processing*. John Wiley & Sons, 2013. 6

[26] Giovanni M Di Liberto, James A O'sullivan, and Edmund C Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015. 6, 77, 78

[27] Akshara Soman, CR Madhavan, Kinsuk Sarkar, and Sriram Ganapathy. An EEG study on the brain representations in language learning. *Biomedical Physics & Engineering Express*, 5(2):025–041, 2019. 6, 58

[28] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018. 6, 77, 80

[29] Akshara Soman, Prathibha Ramachandran, and Sriram Ganapathy. ERP evidences of rapid semantic learning in foreign language word comprehension. *Frontiers in Neuroscience*, page 178, 2022. 6, 77, 79

[30] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014. 6

[31] Steven J Luck and Emily S Kappenman. *The Oxford handbook of event-related potential components*. Oxford university press, 2011. 6

[32] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 7

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 8

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 9

[35] Catherine T Best, M Tyler, O Bohn, and M Munro. Nonnative and second-language speech perception. *Language experience in second language speech learning*, pages 13–34, 2007. 9

[36] Hilla Jakoby, Abraham Goldstein, and Miriam Faust. Electrophysiological correlates of speech perception mechanisms and individual differences in second language attainment. *Psychophysiology*, 48(11):1517–1531, 2011. 9

[37] Jeremy M Anglin, George A Miller, and Pamela C Wakefield. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186, 1993. 10

[38] Robert J Sternberg. Most vocabulary is learned from context. In *The nature of vocabulary acquisition*, pages 89–105. Psychology Press, 2014. 10, 11

[39] Patricia K Kuhl, Barbara T Conboy, Sharon Coffey-Corina, Denise Padden, Maritza Rivera-Gaxiola, and Tobey Nelson. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):979–1000, 2008. 10

[40] Paul Bloom. *How children learn the meanings of words*. MIT press, 2002. 11

[41] Jane B Childers and Michael Tomasello. Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental psychology*, 38(6):967, 2002. 11

[42] Gedeon O Deák. Hunting the fox of word learning: Why “constraints” fail to capture it. *Developmental Review*, 20(1):29–80, 2000. 11

[43] Sandra R Waxman and Amy E Booth. Principles that are invoked in the acquisition of words, but not facts. *Cognition*, 77(2):B33–B43, 2000. 11

[44] Albert Costa and Mikel Santesteban. Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and l2 learners. *Journal of memory and Language*, 50(4):491–511, 2004. 11

[45] Ulrike Halsband. Bilingual and multilingual language processing. *Journal of physiology-Paris*, 99(4-6):355–369, 2006. 11

[46] Arturo E Hernandez, Mirella Dapretto, John Mazziotta, and Susan Bookheimer. Language switching and language representation in spanish–english bilinguals: An fmri study. *NeuroImage*, 14(2):510–520, 2001. 11

[47] Viorica Marian, Michael Spivey, and Joy Hirsch. Shared and separate systems in bilingual language processing: Converging evidence from eyetracking and brain imaging. *Brain and language*, 86(1):70–82, 2003. 11

[48] Manuela Friedrich and Angela D Friederici. N400-like semantic incongruity effect in 19-month-olds: Processing known words in picture contexts. *Journal of cognitive neuroscience*, 16(8):1465–1477, 2004. 11

[49] Manuela Friedrich and Angela D Friederici. Lexical priming and semantic integration reflected in the event-related potential of 14-month-olds. *Neuroreport*, 16(6):653–656, 2005. 11

[50] Janne von Koss Torkildsen, Tuva Sannerud, Gro Syversen, Rune Thormodsen, Hanne Gram Simonsen, Inger Moen, Lars Smith, and Magnus Lindgren. Semantic organization of basic-level words in 20-month-olds: An erp study. *Journal of Neurolinguistics*, 19(6):431–454, 2006. 11

[51] Katherine E Travis, Matthew K Leonard, Timothy T Brown, Donald J Hagler Jr, Megan Curran, Anders M Dale, Jeffrey L Elman, and Eric Halgren. Spatiotemporal neural dynamics of word understanding in 12-to 18-month-old-infants. *Cerebral Cortex*, 21(8):1832–1839, 2011. 11

[52] Joseph R Jenkins, Marcy L Stein, and Katherine Wysocki. Learning vocabulary through reading. *American Educational Research Journal*, 21(4):767–787, 1984. 11

[53] William E Nagy, Richard C Anderson, and Patricia A Herman. Learning word meanings from context during normal reading. *American educational research journal*, 24(2):237–270, 1987. 11

[54] Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975. 11

[55] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975. 11

[56] Peter Cariani and Christophe Micheyl. Toward a theory of information processing in auditory cortex. In *The human auditory cortex*, pages 351–390. Springer, 2012. 12

[57] Ernst Pöppel and Tom Trans Artin. *Mindworks: Time and conscious experience.* Harcourt Brace Jovanovich, 1988. 12

[58] Rufin VanRullen and Christof Koch. Is perception discrete or continuous? *Trends in cognitive sciences*, 7(5):207–213, 2003. 12

[59] Census of india 2011. https://censusindia.gov.in/census.website/data/census-tables, 2011. Accessed: 30 July 2023. 17

[60] Bjarke Frellesvig. *A history of the Japanese language*. Cambridge University Press, 2010. 17

[61] Shao Jie Shi and Bao Liang Lu. EEG signal classification during listening to native and foreign languages songs. In *4th Int. IEEE/EMBS Conf. on Neural Engineering*, pages 440–443, April 2009. 21

[62] Guy Vingerhoets, John Van Borsel, Cathelijne Tesink, Maurits van den Noort, Karel Deblaere, Ruth Seurinck, Pieter Vandemaele, and Eric Achten. Multilingualism: an fMRI study. *NeuroImage*, 20(4):2181–2196, 2003. 21

[63] Gerda Videsott, Bärbel Herrnberger, Klaus Hoenig, Edgar Schilly, Jo Grothe, Werner Wiater, Manfred Spitzer, and Markus Kiefer. Speaking in multiple languages: Neural correlates of language proficiency in multilingual word production. *Brain and language*, 113(3):103–112, 2010. 22

[64] I Berlad and H Pratt. P300 in response to the subject's own name. *Clinical Neurophysiology*, 96(5):472–474, 1995. 22

[65] Z Radicevic, M Vujovic, Lj Jelicic, and M Sovilj. Comparative findings of voice and speech: language processing at an early ontogenetic age in quantitative EEG mapping. *Experimental brain research*, 184(4):529–532, 2008. 22

[66] Thomas F Münte, Eckart Altenmüller, and Lutz Jäncke. The musician's brain as a model of neuroplasticity. *Nature Reviews Neuroscience*, 3(6):473–478, 2002. 22

[67] Henriette Van Praag, Gerd Kempermann, and Fred H Gage. Neural consequences of enviromental enrichment. *Nature Reviews Neuroscience*, 1(3):191–198, 2000. 22

[68] Lee Osterhout, Andrew Poliakov, Kayo Inoue, Judith McLaughlin, Geoffrey Valentine, Ilona Pitkanen, Cheryl Frenck-Mestre, and Julia Hirschensohn. Second-language learning and changes in the brain. *Journal of Neurolinguistics*, 21(6):509–521, 2008. 22

[69] Ping Li, Jennifer Legault, and Kaitlyn A Litcofsky. Neuroplasticity as a function of second language learning: anatomical changes in the human brain. *Cortex*, 58:301–324, 2014. 22

[70] Johan Mårtensson, Johan Eriksson, Nils Christian Bodammer, Magnus Lindgren, Mikael Johansson, Lars Nyberg, and Martin Lövdén. Growth of language-related brain areas after foreign language learning. *NeuroImage*, 63(1):240–244, 2012. 22

[71] Daniela Perani, Eraldo Paulesu, Nuria Sebastian Galles, Emmanuel Dupoux, Stanislas Dehaene, Valentino Bettinardi, Stefano F Cappa, Ferruccio Fazio, and Jacques Mehler. The bilingual brain. proficiency and age of acquisition of the second language. *Brain: A Journal of Neurology*, 121(10):1841–1852, 1998. 22

[72] Kirsten Weber, Morten H Christiansen, Karl Magnus Petersson, Peter Indefrey, and Peter Hagoort. fmri syntactic and lexical repetition effects reveal the initial stages of learning a new language. *Journal of Neuroscience*, 36(26):6872–6880, 2016. 22

[73] Wolfgang Butzkamm. We only learn language once. the role of the mother tongue in FL classrooms: death of a dogma. *Language Learning Journal*, 28(1):29–39, 2003. 22

[74] Christine E Potter and Jenny R Saffran. The role of experience in children's discrimination of unfamiliar languages. *Frontiers in Psychology*, 6(1587), 2015. 23

[75] Chantel S Prat, Brianna L Yamasaki, Reina A Kluender, and Andrea Stocco. Resting-state qEEG predicts rate of second language learning in adults. *Brain and language*, 157:44–50, 2016. 23

[76] Tonio Ball, Markus Kern, Isabella Mutschler, Ad Aertsen, and Andreas Schulze-Bonhage. Signal quality of simultaneously recorded invasive and non-invasive eeg. *Neuroimage*, 46(3):708–716, 2009. 27

[77] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins. The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Frontiers in Neuroinformatics*, 9:16, 2015. 27, 59

[78] C Kothe. The artifact subspace reconstruction method, 2013. 28, 59

[79] Steven J Luck. *An Introduction to the Event-related Potential Technique*. MIT press, 2014. 28, 59

[80] Chih Chung Chang and Chih Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. 28

[81] Lawrence R Rabiner and Ronald W Schafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978. 29

[82] JG Lourens. Passive sonar detection of ships with spectrograms. In *Proc. of the South African Symp. on Communications and Signal Processing, (COMSIG 90)*, pages 147–151. IEEE, 1990. 29

[83] M Parrot, JJ Berthelier, JP Lebreton, JA Sauvaud, O Santolik, and J Blecki. Examples of unusual ionospheric observations made by the demeter satellite over seismic regions. *Physics and Chemistry of the Earth, Parts A/B/C*, 31(4-9):486–495, 2006. 29

[84] Gert Van Hoey, Wilfried Philips, and Ignace Lemahieu. Time-Frequency analysis of EEG signals. In *Proc. of the ProRISC Workshop on Circuits, Systems and Signal Processing*, 1997. 29

[85] Steven J Schiff, David Colella, Gary M Jacyna, Elizabeth Hughes, Joseph W Creekmore, Angela Marshall, Maribeth Bozek-Kuzmicki, George Benke, William D Gaillard, Joan Conry, et al. Brain chirps: spectrographic signatures of epileptic seizures. *Clinical Neurophysiology*, 111(6):953–958, 2000. 29

[86] Steven W Carruthers. Pronunciation difficulties of japanese speakers of english: Predictions based on a contrastive analysis. *Hawaii Pacific University TESOL Working Paper Series*, 4(2):17–24, 2006. 30

[87] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in Speech Recognition*, pages 65–74. Elsevier, 1990. 34

[88] Frederik Stouten and J.P Martens. On the use of phonological features for pronunciation scoring. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 329–332. IEEE, 2006. 35

[89] Tatsuya Kawahara, Masatake Dantsuji, and Yasushi Tsubota. Practical use of English pronunciation system for Japanese students in the call classroom. In *Eighth Int. Conf. on Spoken Language Processing*, pages 1689–1692, 2004. 35

[90] Joseph Tepperman and Shrikanth Narayanan. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In *Int. Conf. on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*, volume 1, pages 937–940. IEEE, 2005. 35

[91] Liang Yu Chen and Jyh Shing Roger Jang. Automatic pronunciation scoring with score combination by learning to rank and class-normalized dp-based quantization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1737–1749, 2015. 35

[92] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Fourth Int. Conf. on Spoken Language, ICSLP Proceedings*, volume 3, pages 1457–1460. IEEE, 1996. 35, 36

[93] Kikuo Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pages 7–12, 2003. 35

[94] Daniel Povey et al. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on Automatic Sspeech Recognition and Understanding*. IEEE Signal Processing Society, 2011. 35

[95] Lawrence R Rabiner and Biing Hwang Juang. *Fundamentals of Speech Recognition*, volume 14. Prentice Hall Englewood Cliffs, 1993. 35, 36

[96] Antony W Rix, John G Beerends, D.S Kim, Peter Kroon, and Oded Ghitza. Objective assessment of speech and audio quality technology and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1890–1901, 2006. 36

[97] Cort Horton, Ramesh Srinivasan, and Michael D' Zmura. Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'. *Journal of Neural Engineering*, 11:046015, 2014. 41

[98] Alain de Cheveigné, Daniel DE Wong, Giovanni M Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor. Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172:206–216, 2018. 41, 78

[99] *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. 42

[100] Eve V Clark. What's in a word? On the child's acquisition of semantics in his first language. In *Cognitive development and acquisition of language*, pages 65–110. Elsevier, 1973. 47

[101] Marta Kutas and Steven A Hillyard. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161, 1984. 48

[102] Marta Kutas and Steven Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980. 49

[103] John CJ Hoeks, Laurie A Stowe, and Gina Doedens. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, 19(1):59–73, 2004. 49

[104] Suzanne Dikker and Liina Pylkkanen. Before the N400: Effects of lexical–semantic violations in visual cortex. *Brain and Language*, 118(1-2):23–28, 2011. 49

[105] Anna C Nobre and Gregory Mccarthy. Language-related field potentials in the anterior-medial temporal lobe: II. effects of word type and semantic priming. *Journal of Neuroscience*, 15(2):1090–1098, 1995. 49

[106] Marta Kutas, Helen J Neville, and Phillip J Holcomb. A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology Supplement*, 39:325–330, 1987. 50

[107] Dénes Szűcs, Fruzsina Soltész, István Czigler, and Valéria Csépe. Electroencephalography effects to semantic and non-semantic mismatch in properties of visually presented single-characters: the N2b and the N400. *Neuroscience letters*, 412(1):18–23, 2007. 50

[108] Arti Nigam, James E Hoffman, and Robert F Simons. N400 to semantically anomalous pictures and words. *Journal of cognitive neuroscience*, 4(1):15–22, 1992. 50

[109] Marta Kutas and Steven A Hillyard. Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & cognition*, 11(5):539–550, 1983. 50

[110] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62:621–647, 2011. 50

[111] Colin Brown and Peter Hagoort. The processing nature of the N400: Evidence from masked priming. *Journal of cognitive neuroscience*, 5(1):34–44, 1993. 50, 74

[112] Katherine A DeLong, Thomas P Urbach, and Marta Kutas. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117, 2005. 50, 74

[113] Tim Curran, Don M Tucker, Marta Kutas, and Michael I Posner. Topography of the N400: Brain electrical activity reflecting semantic expectancy. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 88(3):188–209, 1993. 50, 74

[114] Phillip J Holcomb. Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology*, 30(1):47–61, 1993. 50, 74

[115] Cyma Van Petten and Marta Kutas. Interactions between sentence context and word frequency in event-related brain potentials. *Memory & cognition*, 18(4):380–393, 1990. 50

[116] Kara D Federmeier and Marta Kutas. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495, 1999. 50

[117] Steven J Luck, Edward K Vogel, and Kimron L Shapiro. Word meanings can be accessed but not reported during the attentional blink. *Nature*, 383(6601):616–618, 1996. 50

[118] Kara D Federmeier and Marta Kutas. Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research*, 8(3):373–392, 1999. 50

[119] Edward W Wlotko, Chia-Lin Lee, and Kara D Federmeier. Language of the aging brain: Event-related potential studies of comprehension in older adults. *Language and Linguistics Compass*, 4(8):623–638, 2010. 50

[120] Michelle Leckey and Kara D Federmeier. The P3b and P600 (s): Positive contributions to language comprehension. *Psychophysiology*, 57(7):e13351, 2020. 50

[121] Charles A Perfetti, Edward W Wlotko, and Lesley A Hart. Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1281, 2005. 50

[122] Anna Mestres-Missé, Antoni Rodriguez-Fornells, and Thomas F Münte. Watching the brain during meaning acquisition. *Cerebral Cortex*, 17(8):1858–1866, 2007. 51

[123] Laura Batterink and Helen Neville. Implicit and explicit mechanisms of word learning in a narrative context: an event-related potential study. *Journal of cognitive neuroscience*, 23(11):3181–3196, 2011. 51

[124] Arielle Borovsky, Marta Kutas, and Jeff Elman. Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296, 2010. 51

[125] Arielle Borovsky, Jeffrey L Elman, and Marta Kutas. Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development*, 8(3):278–302, 2012. 51

[126] Angela D Friederici. Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2):78–84, 2002. 51

[127] Eva M Moreno, Kara D Federmeier, and Marta Kutas. Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and language*, 80(2):188–207, 2002. 51, 52

[128] Lee Osterhout and Phillip J Holcomb. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6):785–806, 1992. 51

[129] Peter Hagoort, Colin Brown, and Jolanda Groothusen. The syntactic positive shift (SPS) as an erp measure of syntactic processing. *Language and cognitive processes*, 8(4):439–483, 1993. 51

[130] Angela D Friederici. The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and language*, 50(3):259–281, 1995. 52

[131] Edith Kaan, Anthony Harris, Edward Gibson, and Phillip Holcomb. The P600 as an index of syntactic integration difficulty. *Language and cognitive processes*, 15(2):159–201, 2000. 52

[132] Darren Tanner, Sarah Grey, and Janet G van Hell. Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology*, 54(2):248–259, 2017. 52

[133] Seana Coulson, Jonathan W King, and Marta Kutas. Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and cognitive processes*, 13(1):21–58, 1998. 52

[134] WC McCallum, SF Farmer, and PV Pocock. The effects of physical and semantic incongruites on auditory event-related potentials. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 59(6):477–488, 1984. 52

[135] Eva M Moreno, Antoni Rodríguez-Fornells, and Matti Laine. Event-related potentials (erps) in the study of bilingual language processing. *Journal of Neurolinguistics*, 21(6):477–508, 2008. 52

[136] Marco D Comerchero and John Polich. P3a and P3b from typical auditory and visual stimuli. *Clinical neurophysiology*, 110(1):24–30, 1999. 52

[137] Monica Fabiani, Demetrios Karis, and Emanuel Donchin. Effects of mnemonic strategy manipulation in a von restorff paradigm. *Electroencephalography and clinical neurophysiology*, 75(1-2):22–35, 1990. 52

[138] John Polich, Christine Ladish, and Tim Burns. Normal variation of p300 in children: age, memory span, and head size. *International Journal of Psychophysiology*, 9(3):237–248, 1990. 52

[139] Fabrizio Vecchio, Sara Määttä, et al. The use of auditory event-related potentials in alzheimer's disease diagnosis. *International journal of Alzheimer' s disease*, 2011, 2011. 52

[140] John Polich. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007. 52

[141] John Polich. Theoretical overview of p3a and p3b. *Detection of change: Event-related potential and fMRI findings*, pages 83–98, 2003. 53

[142] Cady K Block and Carryl L Baldwin. Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior research methods*, 42(3):665–670, 2010. 53

[143] Ben Olah. English loanwords in japanese: Effects, attitudes and usage as a means of improving spoken english ability. *Bunkyo Gakuin Daigaku Ningen Gakubu Kenkyū Kiyo*, 9(1):177–188, 2007. 55

[144] Nobuko Chikamatsu, Shoichi Yokoyama, Hironari Nozaki, Eric Long, and Sachio Fukuda. A Japanese logographic character frequency list for cognitive science research. *Behavior Research Methods, Instruments, & Computers*, 32(3):482–500, 2000. 55

[145] Rolf A Zwaan, Diane Pecher, Gabriele Paolacci, Samantha Bouwmeester, Peter Verkoeijen, Katinka Dijkstra, and René Zeelenberg. Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic bulletin & review*, 25(5):1968–1972, 2018. 58

[146] Marc Brysbaert. How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. *Journal of cognition*, 2019. 58

[147] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007. 58

[148] Guanghui Zhang, David R Garrett, and Steven J Luck. Optimal filters for erp research ii: Recommended settings for seven common erp components. *bioRxiv*, 2023. 59

[149] Peter Hagoort and Colin M Brown. ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia*, 38(11):1518–1530, 2000. 63

[150] Gina R Kuperberg, Tatiana Sitnikova, David Caplan, and Phillip J Holcomb. Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive brain research*, 17(1):117–129, 2003. 73

[151] Harm Brouwer, Hartmut Fitz, and John Hoeks. Getting real about semantic illusions: rethinking the functional role of the p600 in language comprehension. *Brain research*, 1446:127–143, 2012. 74

[152] Janet G Van Hell, Carla B Fernandez, Gerrit Jan Kootstra, Kaitlyn A Litcofsky, and Caitlin Y Ting. Electrophysiological and experimental-behavioral approaches to the study

of intra-sentential code-switching. *Linguistic Approaches to Bilingualism*, 8(1):134–161, 2018. 74

[153] Carla B Fernandez, Kaitlyn A Litcofsky, and Janet G van Hell. Neural correlates of intra-sentential code-switching in the auditory modality. *Journal of Neurolinguistics*, 51:17–41, 2019. 74

[154] Saeid Sanei and Jonathon A Chambers. *EEG Signal Processing and Machine Learning*. John Wiley & Sons, 2021. 78

[155] Nai Ding and Jonathan Z Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1):78–89, 2012. 78, 79

[156] Michael J Crosse, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10:604, 2016. 78

[157] Daniel DE Wong, Søren A Fuglsang, Jens Hjortkjær, Enea Ceolini, Malcolm Slaney, and Alain De Cheveigne. A comparison of regularization methods in forward and backward models for auditory attention decoding. *Frontiers in Neuroscience*, 12:531, 2018. 78

[158] Jonas Vanthornhout, Lien Decruy, Jan Wouters, Jonathan Z Simon, and Tom Francart. Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, 19:181–191, 2018. 78, 84

[159] Alain De Cheveigné, Malcolm Slaney, Søren A Fuglsang, and Jens Hjortkjaer. Auditory stimulus-response modeling with a match-mismatch task. *Journal of Neural Engineering*, 18(4):046040, 2021. 78, 94

[160] Bernd Accou, Mohammad Jalilpour Monesi, Jair Montoya, Tom Francart, et al. Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural

network. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1175–1179. IEEE, 2021. 78, 94

[161] David Poeppel and M Florencia Assaneo. Speech rhythms and their neural foundations. *Nature reviews neuroscience*, 21(6):322–334, 2020. 78, 84

[162] Damien Lesenfants, Jonas Vanthornhout, Eline Verschueren, and Tom Francart. Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech. *Journal of Neural Engineering*, 16(6):066017, 2019. 78

[163] Emily S Teoh, Farhin Ahmed, and Edmund C Lalor. Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment. *Journal of Neuroscience*, 42(4):682–691, 2022. 78

[164] Giovanni M Di Liberto, Daniel Wong, Gerda Ana Melnik, and Alain de Cheveigné. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *Neuroimage*, 196:237–247, 2019. 78

[165] Damien Lesenfants, Jonas Vanthornhout, Eline Verschueren, Lien Decruy, and Tom Francart. Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations. *Hearing Research*, 380:1–9, 2019. 78

[166] Corentin Puffay, Jana Van Canneyt, Jonas Vanthornhout, Tom Francart, et al. Relating the fundamental frequency of speech with EEG using a dilated convolutional network. *arXiv preprint arXiv:2207.01963*, 2022. 78

[167] Corentin Puffay, Bernd Accou, Lies Bollens, Mohammad Jalilpour Monesi, Jonas Vanthornhout, Tom Francart, et al. Relating EEG to continuous speech using deep neural networks: a review. *arXiv preprint arXiv:2302.01736*, 2023. 78

[168] Christian Brodbeck, L Elliot Hong, and Jonathan Z Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983, 2018. 78

[169] Miika Koskinen, Mikko Kurimo, Joachim Gross, Aapo Hyvärinen, and Riitta Hari. Brain activity reflects the predictability of word sequences in listened continuous speech. *NeuroImage*, 219:116936, 2020. 79

[170] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62:621–647, 2011. 79

[171] Bojana Mirkovic, Stefan Debener, Manuela Jaeger, and Maarten De Vos. Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications. *Journal of neural engineering*, 12(4):046007, 2015. 79

[172] Vidhi Sinha Akshara Soman and Sriram Ganapathy. Enhancing the EEG speech match mismatch tasks with word boundaries. *Proceedings of Interspeech 2023*, pages 526–530, 2023. 80, 99

[173] Kyle Gorman, Jonathan Howell, and Michael Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011. 82

[174] Arnaud Delorme and Scott Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004. 83

[175] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995. 84

[176] Zachary M Smith, Bertrand Delgutte, and Andrew J Oxenham. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90, 2002. 84

[177] Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. *eneuro*, 5(2), 2018. 84

[178] Octave Etard and Tobias 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*, 39(29):5750–5759, 2019. 84

[179] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 85

[180] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 85

[181] Mohammad Jalilpour Monesi, Bernd Accou, Tom Francart, and Hugo Van hamme. Extracting Different Levels of Speech Information from EEG Using an LSTM-Based Model. In *Proc. Interspeech 2021*, pages 526–530, 2021. 86, 88, 94

[182] Yulong Li, Dong Zhou, and Wenyu Zhao. Combining local and global features into a siamese network for sentence similarity. *IEEE Access*, 8:75437–75447, 2020. 89

[183] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006. 92

[184] Diogo Almeida and David Poeppel. Word-specific repetition effects revealed by MEG and the implications for lexical access. *Brain and language*, 127(3):497–509, 2013. 107