Brain Signal Recognition using Deep Learning

by
Sahil Datta

A thesis submitted for the degree of Doctor of Philosophy

College of Engineering, Design and Physical Science
Department of Electronic and Computer Engineering

January 6, 2022

Declaration of Authorship

I declare that this thesis titled, Brain Signal Recognition using Deep Learning and the work presented in it are my own. No part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution. I confirm that:

- This work was done entirely while in candidature for a research degree at Brunel University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.

Signed: Sahil Datta

Date: 16/08/2021

Following is the list of accepted/submitted publications resulting from the thesis:

- Datta, S. and Boulgouris, N.V., 2021. Recognition of grammatical class of imagined words from EEG signals using convolutional neural network. Neurocomputing, 465, pp.301-309.
- Datta, S., Holmberg, J.J. and Antonova, E., Electrode Selection and Convolutional Attention Network for Recognition of Silently Spoken Words from EEG Signals. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (in press).
- Datta, S., Holmberg, J.J., Aondoakaa, A.S. and Antonova, E., Recognition of Silently Spoken Word from EEG Signals using Dense Attention Network (DAN). In ICASSP 2021-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (Submitted).

Abstract

Brain Computer Interface (BCI) has the potential to offer a new generation of applications independent of muscular activity and controlled by the human brain. Brain imaging technologies are used to transfer the cognitive tasks into control commands for a BCI system. The electroencephalography (EEG) technology serves as the best available non-invasive solution for extracting signals from the brain. On the other hand, speech is the primary means of communication, but for patients suffering from locked-in syndrome, there is no easy way to communicate. Therefore, an ideal communication system for locked-in patients is a thought-to-speech BCI system.

This research aims to investigate methods for the recognition of imagined speech from EEG signals using deep learning techniques. In order to design an optimal imagined speech recognition BCI, variety of issues have been solved. These include 1) proposing new feature extraction and classification framework for recognition of imagined speech from EEG signals, 2) grammatical class recognition of imagined words from EEG signals, 3) discriminating different cognitive tasks associated with speech in the brain such as overt speech, covert speech, and visual imagery. In this work machine learning, deep learning methods were used to analyze EEG signals.

For recognition of imagined speech from EEG signals, a new EEG database was collected while the participants mentally spoke (imagined speech) the presented words. Along with imagined speech, EEG data was recorded for visual imagery (imagining a scene or an image) and overt speech (verbal speech). Spectro-temporal and spatio-temporal domain features were investigated for the classification of imagined words from EEG signals. Further, a deep learning framework using the convolutional network and attention mechanism was implemented for learning features in the spatial, temporal, and spectral domains. The method achieved a recognition rate of 76.6% for three binary word pairs. These experiments show that deep learning algorithms are ideal for imagined speech recognition from EEG signals due to their ability to interpret features from non-linear and non-stationary signals. Grammatical classes of imagined words from EEG signals were also recognized using a multi-channel convolution network framework. This method was extended to a multi-level recognition system for multi-class classification of imagined words which achieved an accuracy of 52.9% for 10 words, which is much better in comparison to previous work.

In order to investigate the difference between imagined speech with verbal speech and visual imagery from EEG signals, we used multivariate pattern analysis (MVPA). MVPA provided the time segments when the neural oscillation for the different cognitive tasks was linearly separable. Further, frequencies that result in most discrimination between the different cognitive tasks were also explored. A framework was proposed to discriminate two cognitive tasks based on the spatio-temporal patterns in EEG signals. The proposed method used the K-means clustering algorithm to find the best electrode combination and convolutional-attention network for feature extraction and classification. The proposed method achieved a high recognition rate of 82.9% and 77.7%.

The results in this research suggest that a communication based BCI system can be designed using deep learning methods. Further, this work add knowledge to the existing work in the field of communication based BCI system.

Acknowledgements

Doing a PhD was a great learning experience, it has motivated me to keep learning and growing in all areas of my life. I want to thank my supervisor Dr. Nikolous Boulgouris for his support during this research. I want to thank Prof. Michael Wright for sharing his knowledge and for training me on how to use the EEG lab. I want to thank my fellow doctoral researchers for making this PhD a fun experience. Most of all, I am grateful to my family for always being there for me, for their love, and support throughout this research.

Contents

List of Figures

List of Tables

List of Abbreviations

ALS	Amyotrophic Lateral Sclerosis
ANN	Artificial Neural Network
AR	Auto-Regressive
BCI	Brain Computer Interface
BSS	Blind Source Separation
CBAM	Convolutional Block Attention Module
CNV	Contingent Negative wave
CNN	Convolutional Neural Network
CRL	Common Representation Learning
CSP	Common Spatial Patterns
DNN	Deep Neural Network
DSP	Digital Signal Processing
DT	Decision Tree
DWT	Discrete Wavelet Transform
DTW	Dynamic Time Wrapping
ECOG	Electrocardiography
EEG	Electroencephalography
EKG	Electrocardiogram
EMD	Empirical Mode Decomposition
EMG	Electromyography
ERD	Event Related Desynchronises
ERP	Event Related Potential
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
fMRI	Functional Magnetic Resonance Imaging
fNRI	Functional Near-Infrared Spectroscopy
GA	Genetic Algorithm
HEOG	Horizontal Electrooculograph
HHT	Hilbert Huang Transform
HMM	Hidden Markov Models
Hz	Hertz
IMF	Intrinsic Mode Functions
KNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
LOO	Leave One Out
LSO	Leave Subject Out
LSTM	Long Short Memory Units
LTO	Leave Trial Out
LWO	Leave One Word Out
MC-CNN	Multi-Channel Convolutional Neural Network
MEG	Magnetoencephalography
MRI	Magnetic Resonance Image
MVPA	Multi-Variate Pattern Analysis

NB	Naive Bayes
PET	Positron Emission Tomography
PCA	Principle Component Analysis
PSD	Power Spectral Density
PSO	Particle Swarm Optimization
RF	Random Forests
RMS	Root Mean Square
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machines
STFT	Short Time Fourier Transform
SNR	Signal-to-Noise Ratio
TG	Temporal Generalization
VEOG	Vertical Electro-oculograph

Chapter 1

Introduction

The electroencephalogram (EEG) technology have made it possible to measure brain activity in a non-invasive manner, which can be used for communication especially by people with motor disabilities. A speech-based brain computer interface (BCI) would provide a mean of communication to the people unable to speak. However, EEG signals are difficult to interpret due to low signal-to-noise ratio (SNR), which limits the ability of EEG-based BCI system for real world application. On the other hand, advancement in deep learning techniques have made it possible to solve complex problems in many areas such as natural language processing, speech recognition, and computer vision (Zhang et al., 2019c). The unique ability of deep learning to learn high-level complex patterns makes it an ideal candidate for analyzing EEG signals and for designing a "thought-to-speech" BCI system.

1.1 Research Problem

Speech recognition is one of the most remarkable achievements in past few years, it has been intensively used by lawyers, doctors, and is available in commercial devices such as phone, laptops, and computers. In addition, speech recognition is faster than human typing on a keyboard (29 bits per second on average) (Herff & Schultz, 2016). However, there are certain problems that even speech recognition cannot solve. In America alone, 7.5 million people suffer from some sort of communication disability (on Deafness & communication Disorder, 2019), some of them can neither communicate vocally nor physically, being totally locked in. Locked in patients have healthy cognitive abilities to think and reason but are unable to move or speak. On the other hand, an EEG based brain computer interface (BCI) technology does not need speech input to produce a response as it can be controlled by thoughts. Brain computer interface is a technology which does not require input from muscles, rather it uses brain signals to command an electronic device. A thought-to-speech interface would be able to serve as a mode of communication for people with severe motor disability. Applications of thought-to-speech BCI can also be used in situations where making sound is not an option, for example, answering an important phone in a library or a meeting. Further, EEG signals of a person can also be used as bio-metric which cannot be forged. BCI technology offers several possibilities to change the way humans interact with their environment and have become one of the most exciting area of research in the last few years.

Regardless of these abilities, practical implementation of this technology in the real world has been limited to monitoring sleep and improving learning rate with products like Dreem and Halo Sport (Insights, 2019). This is because of the non-stationary and non-linear nature of EEG signals making them difficult to interpret. The non-stationary nature refers to the inter and intra-trial variability of EEG signals. Most of the machine learning techniques are designed for stationary data and are not optimal for learning non-linear trends in EEG signals. Therefore, a classifier trained on data from one session would be ineffective on recognizing EEG data recorded from another session (Alsaleh, 2019). Further, EEG signals have low signal-to-noise ratio with several noise sources such as muscle movement, eye movement or blinks and environmental noise. Apart from these, there are several factors causing deformations in EEG signals, such as variation in spatial information due to subjects having slight difference in head size and shape (Bashivan et al., 2015). Brain signals variations are also result of neurophysiological state during recording such as level of attention and emotional state. This is also followed by the type

of modality (cognitive activity) used to elicit brain signals e.g. hand movement, showing images and/or covert speech. Traditional BCI uses event related potentials (ERP) signals or EEG signals of motor imagery as input to the system, which have been the point of interest for research (Bashivan et al., 2015; Lawhern et al., 2018; Tabar & Halici, 2016). There are ERP based BCI systems such as P300 spellers, that allow text entry with a virtual keyboard using eye gaze (Nijboer et al., 2008). A P300 response is an positive deflection of the brain wave 300*ms* after the stimulus onset (Picton, 1992), these spellers have achieved good accuracy and a reliable speed (Chen et al., 2015; Guger et al., 2009). However, these systems are still relatively slow and does not recognize the word directly. Also, there are BCI systems that use electromyography (EMG) signals (Bocquelet et al., 2016; Guenther et al., 2009; Maier-Hein et al., 2005) to recognize silent speech, but this technique is not very effective with locked-in patients as it uses facial muscles movement.

Silent speech communication systems have been implemented using video recording of tongue and lip movement, but the method is only effective with smaller number of classes (Schultz et al., 2017). There is a need to develop a more intuitive brain computer interface, hence, the idea of using covert speech (also referred as: imagined speech, mentally spoken speech, silent speech) as input to the BCI may improve the speed and reliability of the system. Imagined speech has been defined by (Schultz et al., 2017) as "*internalised process in which one thinks in pure meaning*" which makes it even more difficult to work with. Therefore, this thesis focus on developing methods that can be used to design an EEG based BCI, for recognizing thoughts from the brain without any overt action.

1.2 Motivation

The motivation for this work initiates from the limitations and gap in knowledge identified in the previous imagined speech and brain computer interface studies.

1.2.1 Limitations of Speech Imagery Research

Speech production and processing in the brain is widely investigated using brain imagining methods such as the functional magnetic resonance imaging (fMRI), electrocardiography (ECOG), electroencephalography (EEG), and magnetoencephalography (MEG) (Angrick et al., 2019; Dash et al., 2020; Fonken et al., 2020; Palmer et al., 2001). Although, a large body of research has been done on language in brain, research in speech-based BCI system using non-invasive technology, for example EEG, is still in its infancy (Alsaleh, 2019). The work done so far suffer from several limitations and gaps in the area building a speech-based BCI system. One of the main problem is limited publicly available EEG datasets for covert speech, which leads to limited research as recording EEG signals requires proper recording equipment, human participants, and designing an optimal data recording protocol. Further, most studies on language based BCI model have focused on EEG signals produced by mentally spoken syllables, phonemes, and vowels (Arjestan et al., 2016; Matsumoto & Hori, 2014), which are not practical for implementing a language based BCI system because in daily life we do not use syllables and phonemes to construct words. Even the publicly available datasets contain imagined phonemes/syllables from EEG signals with only few words (Nguyen et al., 2017; Zhao & Rudzicz, 2015). Moreover, lack of research with words has led to an unexplored area of recognizing grammatical classes of words using EEG signals. This could help in building an hierarchical language model for a thought-to-speech BCI. In addition, several studies have used block recording for data collection (Torres-García et al., 2013; Wester, 2006), however previous studies have suggested that this approach leads to temporal correlation in the EEG signals (Porbadnigk et al., 2009).

The Majority of studies investigated EEG signals for covert speech (imagined speech) using traditional feature extraction methods, these methods although powerful suffer from some limitations such as inability to lean features and trends automatically in the input data. On the other hand, EEG signals are a produced by non-linear interactions between neurons in the brain (Hornero et al., 2009) and have low signal-to-noise ratio (SNR), therefore it is difficult to untangle the component of interest from

background activity in the EEG activity. Traditional machine learning techniques have been successful in recognition of motor imagery tasks from EEG signals, however their performances have been low in recognition of covert speech compared to the deep learning techniques (Saha & Fels, 2019; Sharon & Murthy, 2020; Zhao & Rudzicz, 2015). The deep learning methods have state-of-the-art performance in many areas, especially in recognition even with EEG-based BCI systems (Antoniades et al., 2016; Bashivan et al., 2015). Further, deep learning can be used to capture high-level representations in EEG signals which can help increase EEG-based BCI performance due to inter-trial variability. Further, the combination of non-linear activation functions and depth of deep learning models makes them suitable for BCI systems. So far, deep learning methods used with speech imagery EEG signals have been explored in time-domain or traditional features created using techniques such as wavelet transform and representing information as a vector (Panachakel et al., 2019; Sereshkeh et al., 2017). This vector presentation reduces the local component of the information processed in different regions of the brain. Processing of raw EEG data from multiple electrodes offer many advantages, like end-to-end learning and requires no prior feature selection. However, EEG signals have multiple channels and in order to extract useful information from the multi-dimensional imagined speech signal, complicated structures have been employed so far with large number of parameters (Saha & Fels, 2019; Sharon & Murthy, 2020). Majority of studies have implemented their methods under binary classification techniques, whereas a language based BCI system should have a larger vocabulary of words for efficient communication.

In the studies mentioned above a based BCI have always be associated with covert speech. On the other hand, research on thinking has associated verbal representation with mental imagery (Guo et al., 2020; Petsche et al., 1992) which suggests that imagined speech in the brain might accompanied by visual imagery. Moreover, imagined speech of an object such as "*orange*" might lead imagination of spatial information of the object itself (Lee et al., 2019). EEG signals for imagined speech are acquired while the subject attend to a visual cue (DaSalla et al., 2009a; Nguyen et al., 2017), which may be regarded as performing imagined speech and/or visual imagery (Clevert et al., 2015). Therefore, there is a need to study visual imagery as a potential cognitive input to a language based BCI system.

Based on the limitations and research gaps outlined above, the research question of this thesis is:

"*How to achieve high recognition rate for mentally spoken words from EEG signals using machine learning techniques?*"

1.3 Objective

The aim of this thesis is to develop a robust and efficient method for recognition of imagined speech from EEG signals. The most reliable form of communication is speech; therefore this study investigates about word processing in humans under different modalities (cognitive activities), using EEG and machine learning technology. There is a need to develop a more intuitive brain computer interface, hence, the idea of using EEG signals for covert speech (inner speech) as input to the BCI may improve the speed and reliability of the system. In order to recognize covert (imagined) speech from EEG signals there are certain factors that are to be explored, such as EEG datasets for imagined speech and machine learning (deep learning framework) methods are needed to be implemented. However, in order to explore a robust language based BCI system, certain objectives need to be achieved:

1. To design an experimental protocol to acquire EEG signals for imagined speech belonging to different grammatical classes and build an EEG database for variety of words. EEG signals were also acquired for other modalities (tasks) such as visual imagination of image associated with mentally spoken words.

2. To investigate methods for feature extraction and classification of EEG signals for imagined speech.

3. To develop machine learning and deep learning framework that can be used for recognition of imagined speech from EEG signals.

4. To develop a framework for recognizing the grammatical class of the imagined words (covertly spoken) from EEG signals.

5. To discriminate EEG signals produced by imagined speech task with EEG signals from other language based tasks such as visual imagery and overt speech.

1.3.1 Ethical Approval

EEG recording in itself is invasion of human thoughts, in other words the recorded data consist of personal information about the subject. EEG can be used to detect any brain conditions such as dementia, sleep disorder, epilepsy, and encephalopathy, but this research does not aim to detect any brain disorder. All the subjects were informed that their identity will be confidential, their name will not be mentioned anywhere and will be referred to as subject number. Research Ethics Committee of Brunel University, College of Engineering, Design, and Physical Sciences reviewed and approved this research, by approving the participant information sheet and subject's informed consent under reference number 7361-LR-Sep/2017-8301-1. The letter of approval has been attached in Appendix A.

1.4 Structure of the Report

The structure of the thesis is as follows:

- **Chapter 2** provides an in-depth review of brain computer interface (BCI) systems for communication application and deep learning methods. This chapter includes a discussion about the production and processing of language in brain areas. Further, literature on imagined speech recognition with EEG signals is provided along with state-of-the-art methods. A short discussion about association of imagined speech and visual imagery. Finally, gaps in previous studies and possible areas of improvement are discussed.

- **Chapter 3** discusses the experimental setup used for recording EEG signals from human participants (subjects). The motivation of recording the a new EEG dataset is discussed in this chapter along with reasons what is different from previous studies that recorded EEG dataset for imagined speech. The chapter also discusses the time-frequency feature used throughout the thesis and reasons for choosing it over other features.

- **Chapter 4** provides an analysis to differentiate between EEG signals of imagined (covert) speech with respect to visual imagery and overt speech. The chapter also propose a electrode selection methods along with deep learning structure for spatio-temporal feature learning from EEG signals.

- **Chapter 5** presents classification results for imagined words from EEG signals under binary condition. The chapter shows the analysis with linear features and discusses their limitations. Further, a method is proposed to overcome the drawbacks of linear features by using dynamic time warping (DTW).

- **Chapter 6** presents and evaluate an electrode selection method for EEG signals. This chapter investigate the performance of deep learning techniques for recognition of imagined words from EEG signals under binary conditions. This chapter also perform classification between imagined speech and non-speech activity.

- **Chapter 7** first, presents with recognition of grammatical class of imagined words from EEG signals. Second, a multi-stage recognition method is proposed for recognition of imagined words

from EEG signals under multi-class classification task. The chapter also presents the performance of proposed method on publicly available EEG dataset.

- **Chapter 8** discusses the contributions of this research. It also discusses the limitations of the present study and possible future work.

Chapter 2

Background and Literature Review

2.1 Introduction

This chapter provides a detailed overview of brain computer interface (BCI) system for silent speech communication. This chapter contains a background of basic concepts and techniques used in brain computer interface technology. This chapter provides an overview of the brain region involved in language processing. Then stages of BCI system are discussed along with the brain imagining technique EEG used in this thesis. Then the description is provided about machine learning and the deep learning techniques. This is followed by the literature on imagined speech recognition using electroencephalography (EEG) signals and techniques used in past studies.

2.2 The Human Brain

2.2.1 Neurons in The Brain

In the human brain, neurons are the nerve cells that process and transfer, electrical and chemical signals. EEG captures signals that are produced by the firing of neurons in the brain, a neuron has a dendrite that receives electrical potential from other neurons and axons that transport electrical signals to other neurons. Each neuron is linked to the other by a connection of axon and dendrite, which acts as transmitter and receiver in a neuron. This link is known as a synapse, these synapses are of two kinds excitatory synapses which tend to increase the potential in the neuron, and inhibitory synapses which tend to reduce the potential of a neuron.

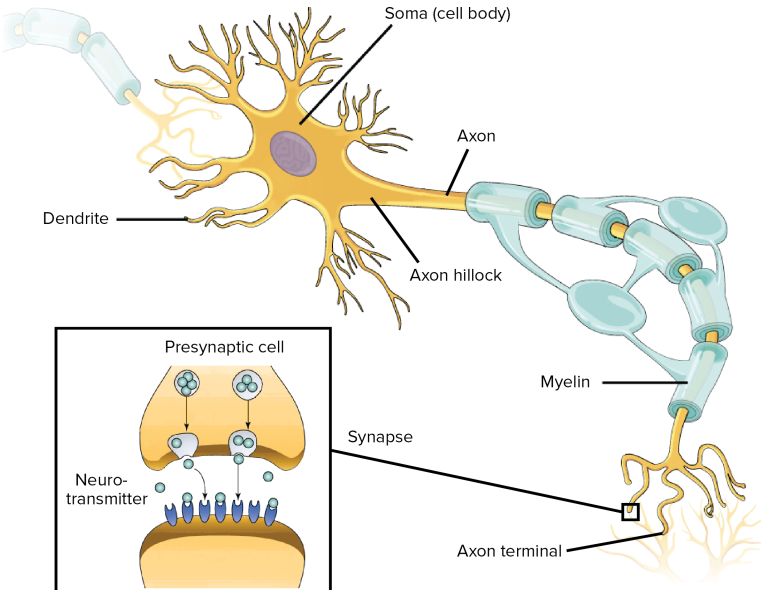


FIGURE 2.1: Structure of a Neuron (khan academy, 2016).

An action potential is evoked in case the excitatory potential which is positive potential exceeds a certain threshold. This potential is then transferred to other neurons. The potential inside a neuron is negative as its cell contains ions guarded by a membrane that does not let the ions go out of the cells or inside it. There are two ways by which the cell maintains a negative potential; active and passive. The cell body has protein which opens and closes for Na^+ and K^+ ions, because of diffusion the K^+ ion moves out of the cell body changes the electric potential to positive. In active strategy, Na^+ is pumped out which makes the cell more negatively charged. During an action potential, this transfer of ion takes place making cell membrane positive caused by the flow of Na^+ into the cell, shown in figure 2.2. This activity is believed to be measured with EEG.

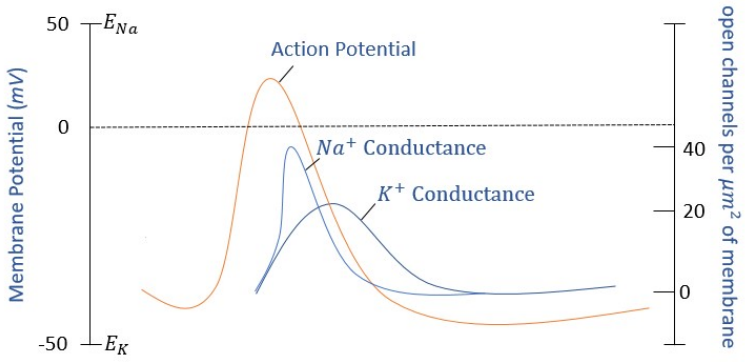


FIGURE 2.2: Permeability of potassium and sodium ions during an action potential.

2.2.2 Speech in The Brain

Speech can be thought of as a continuous flow of words. Phenomena of speech production and areas involved in the human brain are still controversial topics, although the *Perisylvian language area* is considered to play an important role in speech production. It is known to be more dominant (96%) in right-handed people, than (70%) left-handed people (Alsaleh, 2019). The Perisylvian area consists of regions around the Sylvian fissure on the left hemisphere of the brain, it includes the Broca's area, Wernicke's area, the temporal lobe (auditory cortex), and the parietal lobe (the angular gyrus).

- **Broca's area:** Paul Broca was the first to investigate the word production in the inferior frontal gyrus of the brain. Broca's area is on the left side of the brain, it is responsible for the formation of sound, shown in the figure 2.3. Broca's area is believed to be recruited during different stages of word production such as phonological processing, phonetic encoding, and articulator coordination (Flinker et al., 2015). Broca's area has been found to be most active prior to the production of speech, which has been associated with the planning of speech articulation and information transfer to different regions such as motor area (Herff et al., 2015).

- **Wernicke's area:** Another important part involved in speech processing is Wernicke's area which was discovered by German scientist Carl Wernicke which is situated on the left side of the brain as shown in figure 2.3. Broca's and Wernicke's area are connected by a junction of nerves known as arcuate fasciculus. The Wernicke's area plays an important role in translating auditory input to overt and also covert speech output (Pei et al., 2012). Though, the participants differed in both the conditions. The superior temporal gyrus and superior temporal sulcus (Wernicke's area) are said to be involved in the understanding and production of speech.

- **Frontal and Temporal lobes:** Although, Broca's and Wernicke's areas are considered as the classical language regions of the brain. There are studies suggesting the involvement of brain areas outside these two areas in the processing of words belonging to different grammatical classes i.e., nouns and verbs (Preissl et al., 1995). The frontal lobe has been associated with verb processing

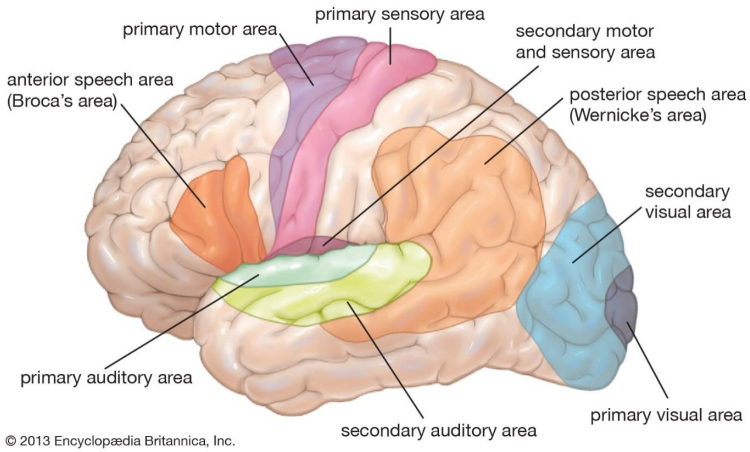


FIGURE 2.3: The Human Brain with important speech processing regions (Encyclopaedia Britannica, 2020).

(Shapiro et al., 2001), apart from that frontal lobe is also known to play a role in language processing (Demb et al., 1995; Gernsbacher & Kaschak, 2003). Neural activity in the frontal and temporal lobe have been found to be most activated during the processing of nouns and verbs (Schilling et al., 2020). Also, the frontal and temporal region have been found to be activated during spoken and imagined vowels and consonants (Pei et al., 2012). The motor cortex has also been associated with language, speech, and sound, transformation of motor commands into speech occurs in the primary motor cortex (Al-Fahoum & Al-Fraihat, 2014). The motor cortex and temporal lobe (auditory area) have been linked to being activated by action verbs (Popp et al., 2019). Behavior such as planning, sequencing of behaviour and cognition occurs in the frontal region of the brain. One of the functions of the frontal lobe is to memorize the result of each step in cognitive task in order to decide how to further execute the task. The left anterior quadrant is known for its role in language production, the right anterior quadrant that is involved in non-verbal function (Petsche et al., 1992). During the verbal tasks, power is observed to be distributed mainly in the frontal and parietal regions (Hwang et al., 2005). Sensorimotor cortex and some regions in the superior temporal gyrus related to language and auditory function have also shown activity after speech onset (Herff et al., 2015). Further, speech comprehension in the brain is known to be processed by Heschl's gyrus also known as primary auditory cortex (Martin et al., 2014). Speech perception and production integration can be represented by sensorimotor integration (SMI) which is the foundation of verbal communication (Jenson et al., 2014). However, the SMI involved in imagined and actual speech activities is different, which is evidence for different neurophysiological phenomena (Jenson et al., 2014).

- **Parietal and Occipital lobe:** Often imagination has been associated with verbal thinking (speech), according to the proportional theory the verbal thinking and imagination in the brain take abstract proportions which have neither verbal nor imagination form (Petsche et al., 1992). Therefore, it is important to consider brain areas involved in the perception of visual information such as images and mental imagery. These brain areas are the frontal lobe, temporal lobe, and parieto-occipital lobe. When information is presented visually there is often synchrony between parietal and occipital region (von Stein et al., 1999). Visual speech interaction and visual speech information is processed in Occipital lobe (Alsaleh, 2019). The occipital region is also known to play an important role in the processing of nouns (Preissl et al., 1995). Further, the visualization condition has shown coherence changes in left occipital and posterior temporal region of the brain. Mental imagery is lateralized to the left hemisphere of the brain, whereas rotation of the images involves the posterior part of the brain (Petsche et al., 1992). Although there is a conflict about the lateralized effect of mental imagery, some studies found that the left posterior was more involved than the

right posterior when analyzing verbally elicited mental imagery (Petsche et al., 1992). However, mental tasks that can affect local coherence are more in the anterior region of both the left and right hemisphere (Pei et al., 2011). Thinking with language has been linked to left frontal activity, whereas thinking with images implies a right frontal contribution (Petsche et al., 1992).

2.3 Brain Computer Interface

A brain computer interface (BCI) is a technology that uses mental activity (*thoughts*) to communicate or send commands to an external device such as a computer. BCI is a direct interface between an external device and the human brain, its objective is to provide an alternative mode of communication without any physical movement. Therefore, BCIs can provide means of communication for people with motor disabilities. For example, people with less severe motor disabilities can use it to control a wheelchair (Graimann et al., 2008), and it is useful in cases where healthy users find conventional means of communication difficult (Allison et al., 2007). Further, symptoms from autism, stroke, attention, and emotional disorder could also be reduced with the help of BCI technology (Gadhoumi et al., 2016; Kouijzer et al., 2009). People with sever motor disabilities have been able to communicate (Nijboer et al., 2008), draw pictures (Münßinger et al., 2010), and control robots (Tonin et al., 2011). BCI system takes brain signals induced by specific tasks such as speech imagery (internal speech) which are recorded from electrodes on the head and these signals are classified into different commands to control the external device. BCIs are mainly of two types; invasive in which electrodes recording the brain signal are implanted over the brain (through surgery) and non-invasive where the electrodes are placed on the scalp (Wolpaw et al., 2002). In this research non-invasive, EEG method is used for recording the brain signals. A BCI system has four stages:

1. **Brain Signal Acquisition**: Measuring brain activity is first and most critical part of BCI system. There are many ways of recording the brain signals, many BCI have been developed using non-invasive methods such as magneto-encephalogram (MEG), functional magnetic resonance imaging (fMRI) (Bocquelet et al., 2016; Guenther et al., 2009; Maier-Hein et al., 2005).

2. **Pre-processing**: Pre-processing enhances the signal quality without loss of information. Physiological signals can be contaminated due to many factors during recording, therefore, this stage is important to clean the data.

3. **Feature Extraction**: At this stage important characteristics of the recorded signals are extracted. In other words, information which encodes a particular command is extracted from the brain signals. Examples of features are event-related potentials (ERP), or time and/or frequency domain features (Bashashati et al., 2007; Bostanov, 2004; Farina et al., 2007).

4. **Classification**: After pre-processing and feature extraction final step is classification. At the classification stage, a particular set of features is assigned a class that refers to a particular mental state or command. There are many classification methods used in BCI system such as K-nearest neighbor (KNN), support vector machines (SVM), linear discriminant analysis (LDA), and neural networks.

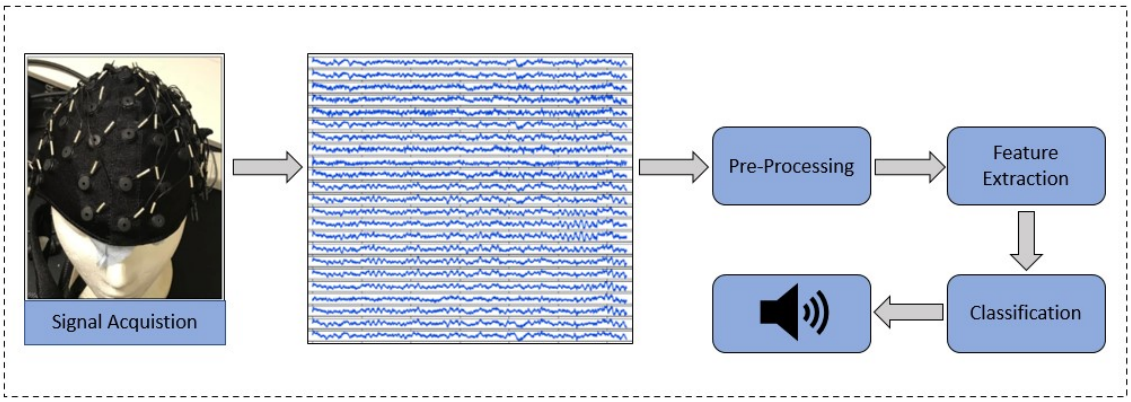


FIGURE 2.4: Stages of BCI system for imagined speech communication. At the first stage, the EEG signals are acquired from the user. These signals are further pre-processed using filters and artifact rejection techniques. At the third stage, features are extracted from the filtered signals which are used for classification. At the last stage, the BCI produces synthetic speech or text output.

2.4 Electroencephalogram (EEG)

Electroencephalogram (EEG) is the technology that measures the electrical activity of the brain. The recorded activity is the increase in potential due to information transfer between the neurons in the brain. EEG was invented by Hans Berger in 1929 (Wester, 2006), where he performed the first human EEG recording, although similar work had been done on animals since 1870. Currently, EEG is used by scientists and physicians to perform neurological diagnoses in order to detect neurological diseases, such as epilepsy, head injury, and sleep disorder (Siuly, 2012).

EEG signals are recorded using non-invasive electrodes usually made of silver-silver chloride (Ag-AgCl), with the advancement of technology, some EEG recording equipment has wireless data transmission facilities. Mostly, Ag-AgCl disc electrodes approximately of $1cm$ diameter are used, for better performance, the electrode disc is filled with conductive gel which bonds the scalp and electrode through hair, this also helps in reducing the impedance of the recorded signal. These electrodes are connected to an amplifier. Electrodes are distributed by a standard 10-20 system for electrode placement on the scalp (Herwig et al., 2003). This was developed by International EEG Federation in order to reproduce the recorded data, 10 and 20 here refer to the fact that electrodes are at either 10 or 20 percent distance of front-back or left-right of the skull. The EEG recording system mainly has an EEG cap, amplifier, and software where the digitized signals are filtered. EEG recording can have multiple electrodes spread at different locations on the scalp, most commonly ranging from 14 to 64 electrodes, where each electrode is called a channel. During an EEG recording each channel capture an electrical signal, therefore EEG signals are also referred to as multi-channel or multi-dimensional signals.

There are certain reasons that make EEG most suitable for a though to speech BCI system. Primarily, EEG have a high temporal resolution, although some other technologies such as positron emission tomography (PET) or magnetic resonance image (MRI) have a high spatial resolution, however, they have a low temporal resolution. This is an important property for recording covert speech which requires capturing changes over time (Wester, 2006). Another advantage is that EEG is easy to use and non-invasive, unlike some other technologies such as ECOG it does not require the subject to undergo any electrode implant in the brain.

The potential of the EEG activity is very small in a range of μV because of which the component of interest is easily contaminated by high potential artifacts. These artifacts can be produced by many reasons for example electromagnetic interference from computers, or other electronic devices present at the place of recording. Another source of artifacts can be subjects themselves, movements such as eye blinks, muscle movement, heartbeat produce artifacts. Therefore, raw EEG signal does not allow

quantitative analysis which requires EEG signals to be processed using filters. EEG measures brain signals at different points in the brain which comprise of different frequencies.

2.4.1 Brain Waves in EEG

As mentioned above there are five main brain frequencies (also referred to as brain waves) namely delta δ (0.5-4Hz), theta θ (4-8Hz), alpha α (8-13Hz), beta β (13-30Hz) and gamma γ (30-200Hz). These brain waves are generalized to play a major role in specific activities, but studies over the years have shown them to behave distinctly in different activities eradicating the misconception of generalization of their behavior to particular activities. Knowledge regarding the brain and its behavior is still in its infancy, phenomena like brain plasticity make it even more difficult to understand the complexities of the human brain and the underlying brain waves. Hence, after thoughtful consideration discussion on role of frequencies has been restricted to speech activity, which is the subject of interest in this study.

Role of brain waves in speech: We studied the role of different frequencies in language and speech processing as they play a specific part at different stages, this eventually helped us understand the importance of different frequencies during this research. Following is a literature survey of the frequencies involved in speech processing:

Delta : Its power is more during the verbal response task, which indicates that delta might play a critical role in perceptual and cognitive processing verbal and non-verbal linguistic tasks (Ding et al., 2016). The significant role of delta oscillation (0.1-4Hz) in decoding imagined speech have been observed in (Dash et al., 2021), however, the best performance was achieved using all the frequencies. Further, increased power in delta band is also associated with drowsiness (Majumder et al., 2019).

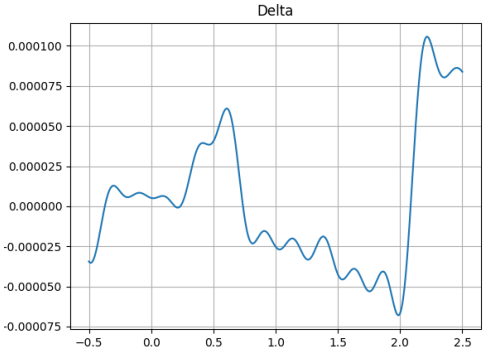


FIGURE 2.5: Delta band in EEG signal.

Theta : A recent study (Kösem & Van Wassenhove, 2017) mentioned that theta waves (4-8Hz) are reactive to phonetic features of speech, they also play role in the reconstruction of phonemes and processing co-articulation cues which helps in the construction of word within a speech.

Alpha : ERD (event related de-synchronisation) has been indicated as a sign of feedback to primary motor cortex (PMC) during speech production. It is considered to represent somatosensory activity. In the study (Jenson et al., 2014), it has been shown to represent auditory feedback, which makes sense about how auditory and somatosensory region provides feedback to PMC while speech is being produced. Alpha ERS/ERD phenomena are observed during speech perception, in the study (Jenson et al., 2014) ERS activity was observed before the onset of the auditory stimulus which was followed by ERD in low alpha (8-10Hz) and high alpha (11-13Hz). Increased alpha activity is often related to concentration tasks and memory workload, which helps in processing the auditory speech and reproducing it as spoken language. Alpha waves during covert speech are weak

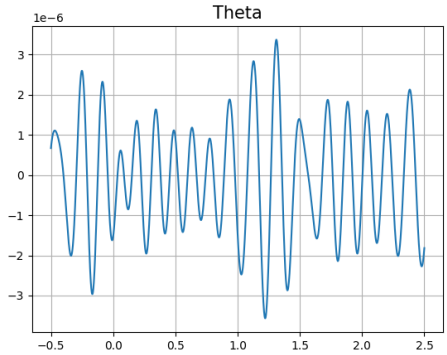


FIGURE 2.6: Theta band in EEG signal.

compared to during overt speech (Jenson et al., 2014). Further, alpha waves also play an important role in attention

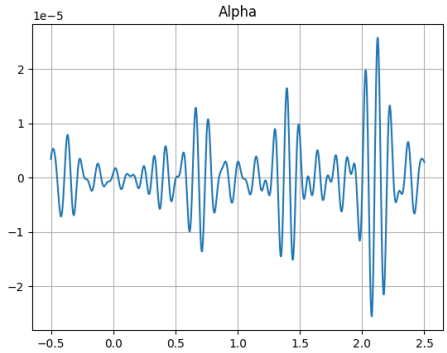


FIGURE 2.7: Alpha band in EEG signal.

Beta : Beta waves in speech can be considered as a result of muscle movement, it is also speculated to play a role in the generation of feed forward control through PMC and internal modeling (feedback) in addition to motor activity. In other words, it is used in discrimination of speech (Bowers et al., 2013) which is then transferred to auditory region for further processing and speech production. The work in (D'Zmura et al., 2009) mentioned that beta waves provided the most information about the EEG signal of imagined speech (syllable).

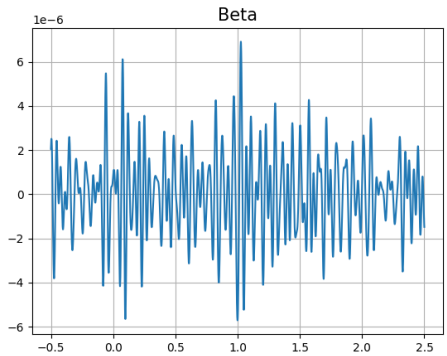


FIGURE 2.8: Beta band in EEG signal.

Gamma : Overt speech production is associated to high gamma changes (70-150Hz) in the temporal lobe (middle and superior parts), supramarginal gyrus, Broca's area, Wernicke's area, premotor cortex, and primary motor cortex (Pei et al., 2011). Whereas the covert speech is also associated with changes in high gamma frequency in the superior temporal lobe and supramarginal gyrus (Pei et al., 2011). The study (Hwang et al., 2005) showed that stimulus present interval (SPI) and inter stimulus interval (ISI) has a rapid effect on gamma frequency; it decreases below the baseline and then very rapidly return to the baseline level.

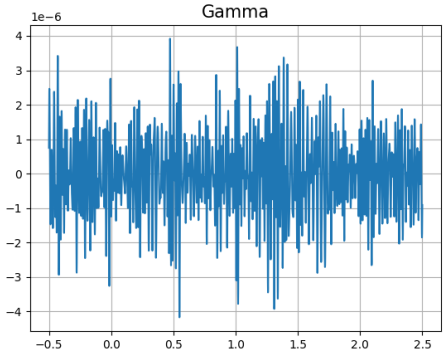


FIGURE 2.9: Gamma band in EEG signal.

2.4.2 Event Related Potentials (ERP)

EEG signals recorded from the scalp are sum of local field potentials (source activity), and non-source activity such as scalp muscle, eye movement, heartbeat, electrode, and environmental noise (Coles & Rugg, 1995). In other words, recorded EEG does not comprise of all the activity from the neurons activated by particular tasks, but also comprises of background activity having weak coherence with task related activity. As mentioned above EEG recording is electrical activity produced by neurons, these signals in passing from one neuron to another suffer conduction by interceding conductive media such as skull, skin, membrane, grey, and white matter, which attenuate the signals (Coles & Rugg, 1995). Therefore, signal averaging over many trials and electrodes is used to eliminate the sources that does not directly contribute to time-locked events (Coles & Rugg, 1995). *When a stimulus is presented, then the voltage change occurs in the EEG signal which is known as event related potential (ERP)* (Coles & Rugg, 1995). The stimulus which gives rise to an ERP, signifies the action generated in the neurons i.e., change of potential by the stimulus. The electrical field activity of a sizable population of neurons is represented by the ERP.

A time domain representation of ERP is a plot of voltage at each time point. Voltage changes that are time locked to certain events within an epoch are called ERP. The advantage of averaging over behavioral study is that it can observe and monitor the unattended stimuli (Luck, 2005). Averaging of trials offer another advantage i.e., it reduces the residual noise in the averaged signal as compared to the single trail, but this works to a certain extend. ERPs have been used to control communication-based BCI systems, some of the ERPs used in BCI applications are SSVEP, N200, P300, N400.

- **SSVEP**: Steady state visually evoked potentials are brain responses and occur at the same fundamental frequency as the presented stimulus. In other words, if a visual stimulus at a frequency ranging from 3.5Hz is presented, the brain will produce an event at a similar frequency (Alsaleh, 2019). Text-based BCI using SSVEP have been explored in past and recent research has achieved text classification based on SSEVP up to 90% (Abdelnabi et al., 2019). However, despite the success SSVEP induced by visual stimulus can cause fatigue and makes it impractical for a communication-based BCI application.

- **P300**: The P300 is an event related potential that can be detected 300*ms* after the stimulus onset (Picton, 1992). This ERP depends on the level of attention and confusion (when an unexpected order stimulus is presented), it is mainly observed at the parietocentral region of the brain. It is known to be the neural signature of required change by the mental model to make an appropriate response (Linden, 2005). The P300 speller is a widely used BCI, however, suffers from certain limitations such as low real-time P300 detection accuracy and has shown difficulty for people who cannot control gaze (Linden, 2005).
- **N200**: Motion related visually evoked N200 is another ERP that has been used for controlling a BCI (Hong et al., 2009). The author developed a speller based on N200 ERP which is a negative deflection around 180 to 300*ms* after the stimulus onset. The accuracy of the N200 speller was comparable to the P300 speller with a smaller number of training data.
- **Hybrid BCI**: A BCI should be able to detect multiple mental activities making it more suitable. Hence, hybrid BCIs that can detect multiple activities or combination of activities have been proposed (Alsaleh, 2019). A combination of P300 and SSVEP was proposed as a hybrid BCI system, where the system detects the level of attention by checking SSVEP activity (Linden, 2005). Further, some hybrid BCIs have combined P300 and MI where the user has navigated through a virtual house by imagining left or right hand movement (Su et al., 2011).

2.5 Machine Learning

Machine learning techniques provide computers the ability to learn patterns from the given dataset. Many problems can be solved using machine learning techniques such as classification, regression, clustering etc. In BCI systems it plays an important role by extracting information from EEG recording and make predictions. Machine learning techniques have been used widely in neuroscientific research (Das et al., 2010; Herff et al., 2015; Knops et al., 2009; Sturm et al., 2016). There are mainly three types of machine learning algorithms:

1. **Supervised Learning**: In this type of learning, a data-set known as training data along with labels which refer to outcomes are fed to the algorithm. Using this data, a training process starts where the algorithm learns to differentiate characteristics between different labels (outcomes). The learning process continues till a desired level of accuracy is achieved, then the algorithm is fed new and unlabeled data to predict the label. Example of supervised learning algorithms are K-NN, logistic regression, and decision tree.
2. **Unsupervised Learning**: The data does not have known labels. The algorithm looks for a structure in the data and organizes similar looking data together. Examples of this kind of problem are clustering, dimensionality reduction algorithms are K-means, and the Apriori algorithm.
3. **Reinforcement Learning**: This type of learning is focused on goal-directed learning through interaction. Reinforcement learning is the process of determining what to do—how to map situations to actions—in order to maximise the magnitude of a reward. The algorithm is not taught which activities to do, but must determine which behaviours produce the greatest reward through trial and error (Sutton & Barto, 2018).

In this thesis, we used all three types of learning. For classification purposes we used supervised learning algorithms like support vector machines (SVM), and K-nearest neighbor (K-NN). Further, dimensionality reduction were implemented as linear discriminate analysis (LDA) and deep neural networks (DNN). Linear classifiers such as K-NN and SVM were used because these methods seem ideal with limited data and easier to implement (Muller et al., 2003). However, EEG signals are non-stationary and complex in nature making it difficult to recognize, therefore deep learning methods were also used in this work.

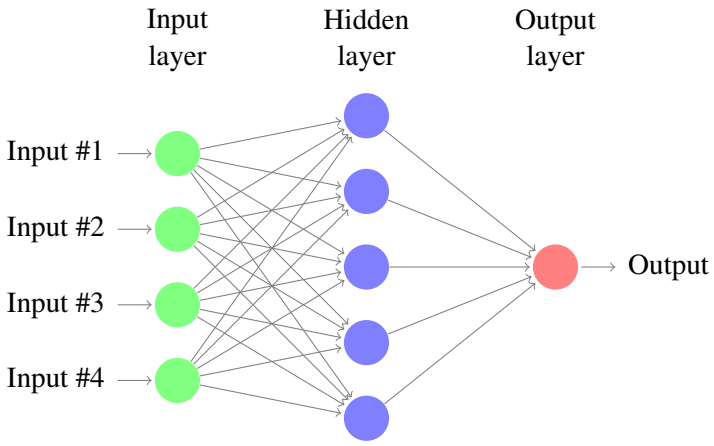


FIGURE 2.10: The architecture of a feed forward network.

2.6 Deep Learning

BCI using traditional feature extraction and machine learning methods have been successful such as recognition of motor-imagery tasks from EEG signals, however, the classification accuracy using standard techniques remains between 60-80% (Zhang et al., 2019c). Further, recognition of imagined speech from EEG signals is more difficult due to the fast processing of speech in the brain (Alsaleh, 2019). In addition, EEG signals suffer from a low signal-to-noise ratio due to the non-stationary nature of EEG signals, which cannot be dealt with using standard feature extraction methods (Zhang et al., 2019c). On the other hand, deep learning has performed better than standard machine learning techniques in different areas. Also, in recent years it has been used to analyze and interpret EEG signals (Hartmann et al., 2018; Zhang et al., 2019b).

Deep learning is a branch of machine learning that is based on the idea of artificial neural networks. The basic building block of neural networks nodes, which is inspired by the biological neurons in the human brain. The basic operation of these nodes is to transfer information to other nodes present in the network. It is a technique that learns high-level features contained in the data. Deep learning models contain several layers, where each layer process features from the previous one. A deep neural network contains non-linearities in form of activation functions which makes them robust towards learning non-linear trends in the given dataset. In this thesis, three types of neural networks are used, that are discussed in the following subsections.

2.6.1 Feed Forward Neural Network

The feed forward neural network is also known as artificial neural network (ANN) is inspired by the way the human brain works (Rosenblatt, 1958). Multiple "neurons" are connected to each other, where neurons are the building blocks of a neural network, these are computational units that perform basic functions such as applying non-linearity to the input. The feed forward network is shown in figure 2.10. The feed forward networks mostly consist of an input, output layer one or more hidden layer. The layers beyond the input layers are called the hidden layers and are made up of neurons, each neuron weights its input. An activation function is used to sum the weighted input to provide a mapping of the input to the output of the neuron. The threshold at which a neuron is activated depends on the strength of the output of the activation function. The weights of the neurons are updated using back-propagation algorithm which reduces the error between target and predicted outcome. The feed forward network although a very powerful algorithm is not effective in solving complex problems requiring spatial and temporal precision.

2.6.2 Convolutional Neural Network

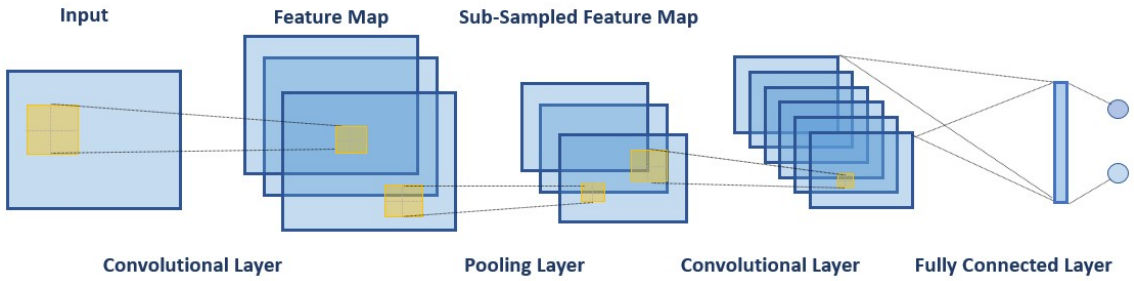


FIGURE 2.11: The architecture of a convolutional neural network showing feature map creation from the input matrix followed by pooling layer and fully connected (dense) layer.

The convolutional neural network (CNN) (LeCun et al., 1998) learn the pattern in the data using convolutional operation. The CNN architecture can vary in the number of layers, ranging from single layer (shallow) architecture to a very deep network with several consecutive layers (Abdel-Hamid et al., 2014; Simonyan & Zisserman, 2014). The CNN's initial layers capture low level features and more complex high-level features are learned in the deeper layers. The CNNs mainly has three main hyper-parameters: filters, strides, and padding.

Filter: At the first stage, the filter captures spatial information from a multi-dimensional input (e.g., image, spectrogram) by performing a convolution operation between the region of the input matrix and the filter (also known as receptive field). The values inside the filter are known as weights. The filter slides over the entire input matrix (image) to extract features, each position of the filter corresponds to activation of the neuron and is collected on a two-dimensional feature map. If the input is a multi-dimensional matrix for example an image with multiple channels, then a neuron is the summation of convolutional operation across all the channels for the same region. The ability of the CNN to extract features using filters is particularly useful, as different feature maps can represent activity at different spatio-temporal windows. The value of the feature on the i^{th} row and j^{th} column of the k^{th} feature map at a given layer is obtained as (Tabar & Halici, 2016):

$$y_{ij}^{k} = h(a) = h((W^{k} * x)_{ij} + b_k) \tag{2.1}$$

where x refers to the input, h to the activation function, $*$ denotes the convolutional operation, b_k refers to the bias value, and W^k is the weight matrix of filter k, with $k = 1, 2, ..., K$, K = 32, 64 or 128.

Stride: Stride specifies how far our sliding filter window will travel when the filter function is applied. Each time the filter function is applied to the input, a new depth column is created in the output map. Lower stride values results in more strongly overlapping receptive fields between the columns, resulting in increased output volumes (Patterson & Gibson, 2017). On the other hand, higher stride values result in reduced spatial overlap and smaller output volumes.

Padding: For the convolution operation, zero padding was used in order to preserve the spatial resolution of the input. Zero-padding is defined as:

$$p = \frac{h_r - 1}{2} \tag{2.2}$$

where p is padding and h_r is receptive field size. Here, h refers to the activation function. The third stage is the pooling or sub-sampling calculates the summary statistic of the local patch in the feature map. The summary statistic is usually average or maximum and creates a lower-level perception of the features by reducing high level details in the feature map which helps avoid over-fitting in the network. The architecture of the CNN is shown in figure 2.11.

2.6.3 Activation Functions

In deep learning models an activation function specifies how the weighted sum of the input is converted to an output from a node or nodes in a layer of the network. The network performance is largely impacted by the choice of activation functions. The activation functions have improved the ability of neural networks to learn complex features (Qian et al., 2018). In this work CNNs were implemented with non-linear activation functions. In CNNs non-linearity is applied to the feature maps using the non-linear activation function, this stage is often called the detector stage (Goodfellow et al., 2016). There are many different types of activation functions; however, this work used four activation functions, defined as follows:

- **ReLU**: The rectified linear unit (*ReLU*) (Nair & Hinton, 2010) is the most widely used activation function. The *ReLU* are effective in alleviating the vanishing gradient problems in the neural networks (Clevert et al., 2015). The values The *ReLU* is defined as:

$$R(a) = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases} \tag{2.3}$$

  where a refers the input to the activation function.

- **ELU**: The exponential linear unit (*ELU*) (Clevert et al., 2015) is defined as:

$$ELU(a) = \begin{cases} a & \text{if } a \geq 0 \\ c(e^a - 1), & \text{otherwise} \end{cases} \tag{2.4}$$

  where c is a parameter.

- **Sigmoid**: The *sigmoid* activation function is usually used in deep learning models for binary prediction. When used in the hidden layer activation function can cause network gradient to saturate (Nielsen, 2015). The *sigmoid* activation is given as:

$$sigmoid(a) = \frac{1}{1 + e^{-a}} \tag{2.5}$$

  where a is defined in (2.1).

- ***Softmax***: The $softmax$ activation is used in deep learning models to make classification prediction. The softmax activation normalizes the input vector such that the sum of all values in the vector is one. The $softmax$ function is define as:

$$\alpha_a = \frac{\exp(a_i)}{\sum_{i=1}^{K} \exp(a_i)} \tag{2.6}$$

  where α is the $softmax$, a_i is the input vector, K is the number of classes.

2.6.4 Recurrent Neural Network

The recurrent neural networks (RNN) (Bishop, 2006; Schuster & Paliwal, 1997) are powerful neural networks for processing sequential data. The RNNs have loops, feedback, and memory which makes them robust when working with temporal data. The RNN can be considered as an extension of the feed forward network where the output of each neuron is also passed latterly to the next neuron in the same layer which is known as recurrent connections. This adds memory to the network and allows the network to learn features across the sequence of the input, therefore the future hidden state of the network

is dependent on the current state. The long-short memory units (LSTM) (Hochreiter & Schmidhuber, 1997) are the most widely form of RNNs, which are trained using backpropagation algorithm. The LSTM cell architecture can be described mathematically (Zhang et al., 2019b) as:

$$i_t = \sigma(W_i \cdot [h_{t-1} + b_i], \tag{2.7}$$

$$f_t = \sigma(W_f \cdot [h_{t-1} + b_f]), \tag{2.8}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t + b_c]), \tag{2.9}$$

$$o_t = \sigma(W_c \cdot [h_{t-1}, x_t + b_o]), \tag{2.10}$$

$$h_t = o_t * \tanh(C_t), \tag{2.11}$$

where h_t is the output of LSTM hidden state vector for t_{th} cell which corresponds to the t_{th} time point, C_{t-1} is cell state at time t, x_t is the input to the cell, h_t is the hidden state of LSTM at time t, W_i,W_f,W_o,W_c are the weights, and b_i,b_f,b_o,b_c are the biases. The output from LSTM from all the hidden states was calculated as:

$$h_t = LSTM(s_t), t \in [1, T] \tag{2.12}$$

2.6.5 Attention Mechanism

The concept of attention mechanism is inspired by the human brain's ability to focus on selective things. It was initially introduced to overcome the problem of processing long sequences (Bahdanau et al., 2014). However, it has been implemented to improve performance of deep learning models for several tasks such as natural language processing, image captioning, and computer vision (Vaswani et al., 2017; Woo et al., 2018; Xu et al., 2015). In representation learning models like the ones used in this work, attention mechanism is used to provide more weights to certain features which provides the most discriminative information.

To implement the attention in our work, we used the self-attention layer, where the output from parallel dense layers were used as input to attention layer to create a more informative global feature map g. The resultant global feature map has more weights assigned to discriminative information which contributes to classification of imagined words. The output from parallel dense layers from all the hidden states was calculated as:

$$d_t = Dense(s_t), t \in [1, T] \tag{2.13}$$

where d_t is the output of single dense layer which corresponds to the t_{th} time point, s_t is the hidden state (features) at time t and T is the number of nodes in a dense layer. The attention mechanism was implemented using a dense layer, where the first layer had one neuron and the $tanh$ activation function. The output of first layer is defined as:

$$u_t = \tanh(W_s d_t + b_s) \tag{2.14}$$

where W_s is the weight and b_s is the bias of the fully connected layer with single neuron. The output u_t was passed through a the $softmax$ activation function to estimate α_t known as the normalized importance vector, which was calculated as follows:

$$\alpha_t = \frac{\exp(u_t)}{\sum_{t=1}^{T} \exp(u_t)} \tag{2.15}$$

$$g = \sum_{t} \alpha_t d_t \tag{2.16}$$

where g is the global feature vector produced by the attention layer. The self-alignment layer is trained using back-propagation algorithm and learn important time points in the feature-map using the gradient of the cost function (Bahdanau et al., 2014).

2.6.6 Network Training

Training is a process, where a deep learning algorithm adjust the weights and biases (initialized randomly or using some other method), making some smaller and others larger, so assigning significance to certain features and reducing importance of others. This assists the network in determining which features are associated with particular outcomes and adjusts the network's weights and biases appropriately (Patterson & Gibson, 2017). The weights and biases are adjusted with the help of optimization algorithm such as the stochastic gradient descent (SGD).

All the networks in this research were trained on NVIDIA Tesla T40 GPU, using the *Keras* library (Chollet et al., 2015) with *Tensorflow* backend (Abadi et al., 2016). The networks were trained with the Adam optimization algorithm. Due to weight sharing in convolutional networks, the gradient at different layers can vary widely (Bashivan et al., 2015). Therefore, networks were implemented with a slower learning rate for training. Further, in order to avoid the problem of unstable gradient (Simonyan & Zisserman, 2014), the network implemented in this research used *He* weights initialization method (He et al., 2015).

2.7 Silent Speech Processing in Literature

Speech is the primary means of human communication and it is an ideal modality for a BCI communication system. A silent speech interface or brain-to-text interface would be able to serve as a mode of communication for people with severe motor disability, for example locked-in patients (section 1.1) (Herff & Schultz, 2016). Analysis of neural mechanism for imagined speech has been studied using different brain imaging techniques such as fMRI, fNRI, ECOG, MEG, and EEG have been used for recognition and analysis (Dash et al., 2021; Heger et al., 2015; Huang et al., 2002; Stephan et al., 2020). Analysis and interpretation of EEG signals produced during speech has been the topic of interest for the scientific community for a long time. Studies more than two decades ago (Suppes et al., 1998; Suppes et al., 1997) combined MEG and EEG attempting to classify averaged EEG signal produced during speech. EEG offer more advantages compared to other brain imagining techniques, EEG devices are cheap, portable, and easy to use. In addition to costs, EEG headsets are available with dry electrodes and wireless communication ability which are more practical in daily use applications. This section provides an literature review of studies done in imagined speech recognition using EEG signals.

2.7.1 Recognition of Syllables

Different aspects of speech have been investigated, the study in (D'Zmura et al., 2009) recorded EEG signals for imagined syllables "*/ba/*" and "*/ku/*" from six subjects to understand the contribution of different frequency bands involved. The study was focused on distinguish the role of alpha, beta, and theta frequency bands in language processing and comprehension in the brain. In order to achieve this, EEG signal envelopes were analyzed for each electrode in Theta (3-8Hz), Alpha (8-13Hz) and Beta (13-18Hz). These envelopes were used to construct the matched filters by averaging the trials for an envelope and condition. Also, the study computed power spectral density (PSD) for all the conditions in different frequency bands. Using two different approaches the author concludes beta band activity to be most informative in comparison to the alpha and theta bands. However, the inter-subject study did not provide accurate information indicating variation in information between subjects and condition. Similarly, the

study in (Brigham & Kumar, 2010a) performed recognition of EEG signals recorded for mentally spoken syllables "*/ba/*" and "*/ku/*" and achieved above chance level accuracy.

EEG signals for speech imagery syllables have not only been studied for a though-to-speech interface, but also have been proposed as an alternative to motor imagery. The work in (Brigham & Kumar, 2010b) suggests that speech imagery EEG signals are a better alternative to EEG signals of motor imagery for bio-metric identification. They aimed to use EEG signals produced during imagined speech of two syllables "*/ba/*" and "*/ku/*" for subject identification. The author mentioned that using syllables instead of words as stimulus would avoid semantic effects in EEG signals. The author used auto-regressive coefficients for feature learning and linear SVM to achieve an accuracy of 99.7%.

The study (Wang et al., 2013a) aimed to distinguished EEG signals for speech imagery of Chinese characters and proposed speech imagery to be suitable for asynchronous BCI system. EEG signal were recorded for mentally spoken Chinese language characters. For feature extraction purpose the study used common spatial patterns (CSP) and cross-correlation function to calculate the eigenvalues of the EEG signals. The extracted eigenvalues from CSP and cross-correlation were classified using SVM classifier.

2.7.2 Recognition of Vowels

Along with syllables, EEG signals produced during mentally spoken vowels have also been studied. The work in (DaSalla et al., 2009a) proposed an algorithm for the purpose of speech prosthetic. EEG signals were recorded for vowel (speech) "*/a/*" and "*/u/*" imagery task and resting state as control task. Grand averaging of the EEG signal was performed in time domain to visualize speech potentials (ERP), common spatial patterns were used for feature extraction and non-linear SVM was used for classification, with classification rate ranging between 68% to 78%.

The work in (Chi et al., 2011) investigated phonemes production using different vocal articulation (tongue, lips, nasal, jaw, and fricative), EEG signals were recorded on different days from the same subject. Five classes were classified in pairwise manner, for classification of imagined phoneme production author used naive Bayes and linear discriminant analysis (LDA). The method achieved a recognition rate of 80%.

Apart from English language, EEG signals produced during mentally spoken vowels in Japanese language have been investigated (Matsumoto & Hori, 2014). The study used two different classification algorithm SVM with Gaussian kernel (SVM-G) and RVM with Gaussian kernel (RVM-G), common spatial patterns (CSP) and adaptive collection (AC) were used for feature learning. The recognition accuracy rate achieved by both the SVM-G and RVM-G were in the same range 77%-79%. The study concluded SVM-G as a better choice because RVM-G had a higher calculation cost and required more training data.

On the other hand, some studies used both vowels and syllables to this topic, (Arjestan et al., 2016) used three syllables, six vowels, and no-activity as a control state to decode the covert and overt speech from EEG signals. In this study the authors further point out the advantages of using covert speech for recognition of speech.

The work with phonetics was taken a step ahead in (Zhao & Rudzicz, 2015) by checking the presence of different phonological categories like presence of nasal, bilabial, high-front vowel and high-back vowel, using EEG signals along with six other modalities. EEG signals were recorded while subjects mentally spoke seven phonemes (*/uw/*, */iy/*, */tiy/*, */diy/*, */m/*, */n/*, */piy/*) and four words (*pat*, *pot*, *knew*, *gnaw*). Statistical features like (mean, variance, median, standard deviation) along with entropy, skewness, kurtosis, and energy were calculated. The author used SVM with quadratic kernel (SVM-quad) and radial basis function (SVM-rbf) to perform classification for five tasks, achieving an accuracy of above chance level for four out of five tasks.

2.7.3 Recognition of Words

The first attempt of recognizing imagined speech for mentally spoken words from EEG signals was in (Suppes et al., 1997). However, no new work was done until (Wester, 2006) acquired EEG signals for silent speech and speech gestures such as mumbling and whispering. This author used short time Fourier transform (STFT), delta coefficients for feature extraction, linear discriminant analysis (LDA) for dimensionality reduction, and hidden Markov models (HMM) for classification achieving an above chance accuracy. Additionally, results from electrodes over Broca's, Wernicke and Homunuculus areas were also evaluated, but the results using all the electrodes achieved better recognition rates.

In the study (Porbadnigk et al., 2009) the author suggests that the order in which the stimulus is presented can impact the recognition rate of imagined speech. EEG signals were recorded by presenting the visual stimulus in different orders (random, blocks). It was concluded that presenting words in sequence (blocks) during EEG recording leads to better recognition rate, the author mentioned that this increased the concentration and hence reduce the noise in the signal. It was also mentioned that better recognition rate could be due to temporal correlated artifacts in the signals, as the recognition rate reduced when the words are presented in random order. The study in (García et al., 2012) proposed a channel selection method for recognition of unspoken speech from EEG signals. The author used EEG signals produced during covertly spoken Spanish words: "*arriba (up)*", "*abajo (down)*", "*izquiedra (left)*", "*derecha (right)*", "*slecciobar(select)*", similar to past studies (Porbadnigk et al., 2009; Wester, 2006) presented the stimulus in blocks. The proposed technique produced recognition rate of 68% with seven electrodes compared to 70% with all the electrodes. Also, the author mentioned that the unspoken speech window for each speaker and for each word can vary, therefore during recording and pre-processing all the signals in the EEG dataset should be transformed to equal length.

Further, EEG signals are affected by the length of signal recording (Levy, 1987), and therefore an appropriate window size should be chosen for a trial. In addition, overt speech-based EEG signals accompanied by muscle movement makes it even more difficult to understand the underlying neural mechanism. EEG signals from the same subject can vary from one day to the other leading to change in recognition rate (Chi et al., 2011), making experiments with longer imagined words or speech difficult.

Despite these drawbacks studies have shown to successfully analyze and classify unspoken speech for an online BCI. The work in (Sereshkeh et al., 2017) used EEG signals for mentally spoken words: "*yes*", "*no*" and no-activity as control state for binary classification task as a preliminary step towards a covert speech based online BCI system. The study recorded EEG signals in two sessions, the first session was the training session, and the second session was the online testing session. Statistical features using DWT were calculated and artificial neural network (ANN)- multi-layer perceptron (MLP) was used for classification. The results were evaluated under binary setting for word "*No*" and control condition with average accuracy for 12 subjects of 75.7%. For "*No*" and "*Yes*" an accuracy rate of 63.2% was achieved. These results were reported to be higher than results obtained using K nearest neighbor (KNN), SVM, and Naive Bayes.

Work with internal speech was further extended to bilingual unspoken speech by (Balaji et al., 2017), EEG signals from five subjects were recorded during imagined words in Hindi: "*Haan*" and "*Na*" and English words:"*Yes*" and "*No*". In the experiment subjects were asked ten questions in both the languages and subjects had ten seconds to answer the questions. Subjects were proficient in both Hindi and English languages. EEG data was subject to dimensionality reduction using principle component analyses (PCA) and trained different classifier: SVM, random forest (RF) artificial neural network (ANN), and AdaBoost for Hindi and English language. ANN outperformed other classifiers and achieved an average recognition rate of 75.3%.

Similarly, advanced work was done by (Nguyen et al., 2017) which recorded EEG signals for imagined speech from words of varying length from 15 subjects. Covert speech was recorded for vowels: "*/a/*", "*/i/*", and "*/u/*", short words: "*in*", "*out*" and long words: "*Cooperate*" and "*independent*". This was the first time EEG signals produced by imagined speech for words of varying length. The study proposed an algorithm based Riemannian manifold for features extraction and used relevance vector

machines for classification. The results were evaluated under four binary conditions: classification of vowels, short words, long words and short long words, and proposed method achieved an accuracy of 49%, 50.1%, 66.2% and 80.1%. The study achieved best recognition rate with classification between words of varying length ("*Short-Long words*"). Further, the authors suggested that recognition of imagined speech for shorter words is difficult in comparison to imagined speech for longer words.

The work in (Saha & Fels, 2019) proposed a hierarchical deep feature learning method by combining CNNs, RNNs, and auto-encoder to recognize the long words dataset used in (Nguyen et al., 2017). The author proposed that imagined speech related cognitive processes could be captured by a channel cross covariance matrix and used it as feature in order to have better spatio-temporal representation of EEG data. The CNN and RNN were trained in parallel for extracting spatial and temporal information, the features from these two networks were combined and fed to an auto-encoder for dimensionality reduction. The proposed method achieved an average accuracy of 79%, which is an increase of 13% compared to previous work in (Nguyen et al., 2017) on the same EEG database. Further, in (Saha et al., 2019) this method was also applied on another publicly available dataset *Kara One* from the study (Zhao & Rudzicz, 2015). The proposed technique outperforms past methods by 67.15% in one of the five recognition task and achieving highest accuracy on the dataset. This shows ability of deep learning techniques in effectively learning behavior of EEG signals.

(Krishna et al., 2019) proposed automatic speech recognition model using EEG signals for vowels and speech signals. The author used three different set of features acoustic features, acoustic and EEG features combined, and EEG features only. Mel-frequency cepstral coefficients (MFCC) were used as acoustic features and gated recurrent units (GRU) was used for classification. Distillation technique was used for training the GRU. Best recognition rate of 96.3% was achieved when features from EEG and MFCC were combined. Further, the author proposed electrodes T7, T8, FC5 and P7 showed to be the most informative electrodes. However, the application of the method is limited for people suffering with locked-in syndrome due to use of speech/audio signals.

(Sharon & Murthy, 2020) proposed a method that uses multi-phasal information in the EEG data as an alternative to multi-modalities information such as speaking, imagination and speech related articulately movement. The work used inter-phase information by using common representative learning (CRL) which integrates information from multiple models of the EEG data. In order to achieve this correlation network (CorrNet) which maximizes correlation between different modes were used along with DNN to achieve state-of-the-art recognition rate on publicly available EEG dataset *Kara One*.

Most of the work in the literature have been focused on binary classification of words or phonemes. However, in recent years there has been a growing body of research performing multi-class classification of EEG signals of unspoken speech. The work in (Panachakel et al., 2019) used discrete wavelet transform with deep neural network to classify publicly available Kara-One EEG dataset (Zhao & Rudzicz, 2015) and achieve a classification accuracy of 57.1%. Another work (Pawar & Dhage, 2020) also used Kernel based machine learning method to classify EEG signals from mentally spoken words under multi-class setting.

2.8 Methods proposed for recognition of Imagined speech

In order to design a BCI for imagined speech recognition it is important to learn the discriminative features representing different classes. In the literature many methods have been proposed to learn discriminative features from EEG signals. A popular technique that many studies have used is the Discrete Wavelet Transform (DTW) to decomposed the signals and extract features such as energy, entropy, and power spectral density to perform classification (Balaji et al., 2017; Panachakel et al., 2019; Sereshkeh et al., 2017). The work in (García et al., 2012) classified EEG signals for imagined speech for five Spanish words using different classifiers. The results were obtained for five classes using discrete wavelet transform (DWT) feature extraction method, for classification purposes naive Bayes (NB), Random forests (RF), SVM and bagging RF were used. In (Dash et al., 2021) MEG signals for mental speech were

decomposed to seven levels using DWT, which were reconstructed using high pass coefficient from level two to seven. From the decomposed signals root mean square (RMS) features were extracted and concatenated from different frequency bands. The study proposes the delta frequency band to be the most discriminative, however the best results were obtained when features from all the bands were concatenated.

Similarly, common spatial patterns (CSP) (Koles et al., 1990) have been very successful in learning spatial patterns from multi-dimensional EEG data. CSP have been used along other techniques for feature learning of imagined speech, the work in (Wang et al., 2013a; Wang et al., 2013b) used CSP along with cross-correlation matrix to recognize mentally spoken Chinese characters. Further, the study in (Arjestan et al., 2016) performed classification of EEG signals for overt and covert speech, ad the features were extracted using intrinsic mode functions (IMF). Further, eigen vectors were extracted from IMFs using common spatial pattern (CSP) and two different feature vectors were used for classification purpose.

Machine learning techniques such as SVM have also been widely used for classification of imagined speech EEG signals. The studies in (Arjestan et al., 2016; Matsumoto & Hori, 2014) used combination of CSP for feature extraction and SVM for classification of EEG signals produced during mentally spoken vowels. Other techniques such as auto-regressive (AR) modeling, Riemannian manifold, and empirical mode decomposition (EMD) have achieved above chance level recognition rate (Brigham & Kumar, 2010a; Nguyen et al., 2017).

However, most feature extraction methods suffer from limitations, for example CSP performs better with two classes (Sun & Zhang, 2006). Auto regressive modeling can suffer from poor spectral estimation of signal (Al-Fahoum & Al-Fraihat, 2014) and are unable to adapt to variations within a given class. Adaptation to variations is particularly important in imagined speech recognition, where the semantic variations lead to changes in the processing of words in the brain (Vigliocco et al., 2011).

On the other hand, deep learning has exhibited excellent performance in various recognition tasks (Graves et al., 2013; Jiang & Yin, 2015; Krizhevsky et al., 2012), including the recognition of imagined speech from EEG signals. The work in (Panachakel et al., 2019) used deep learning to perform multi-class classification of mentally spoken phonemes and words to achieve 57% accuracy rate. An Artificial Neural Network (ANN) was used to classify bilingual unspoken speech in (Balaji et al., 2017) from 11 classes and recognition rate was above chance level. The accuracy of recognizing long words was improved by a 23% in (Saha & Fels, 2019) using a hybrid network designed of CNN, recurrent neural network (RNN) and auto-encoder. Likewise, the work in (Sharon & Murthy, 2020) proposed a classification framework that uses inter-phasal information by implementing the common representation learning (CRL) in neural network to recognise imagined speech using publicly available “Kara One” EEG dataset. This study achieved state-of-the-art results under binary classification task. Studies in literature and methods used are shown in Table 2.1 and Table 2.2.

TABLE 2.1: Studies in the literature that used EEG signals for imagined speech recognition.

Reference	Task	Brain Area	Method	Performance
(Suppes et al., 1997)	Recognition of 7 words: First, second, third, yes, no, right, and left	F7, T3, T5, FP1, F3, C3, P3, FZ, CZ, FP2, F4, C4, P4, F8, T4, T6	Averaging and prototyping and least square criterion	Above chance level of 1/7
(Wester, 2006)	Different modalities: i.e. imagined speech, and silent mumbling	Primary motor cortex, the Broca's and Wernickes area	LDA and HMM	Above chance of 50% level
(D'Zmura et al., 2009)	Binary classification between /ba/ and /ku/	All	Matched Filter	not mentioned
(Porbadnigk et al., 2009)	Classification of words to access order of word presentation during EEG signal recording	Broca's area	Hidden Markov Model (HMM)	Above chance level (Block mode), Chance level (other modes)
(Brigham & Kumar, 2010a)	Binary classification between /ba/ and /ku/	not mentioned	Auto regressive modeling and K-NN	61%
(Brigham & Kumar, 2010b)	Binary classification between /ba/ and /ku/ for subject identification	Not mentioned	Auto regressive modeling, Power spectral density (PSD) and SVM classifier	99.76%
(Yoshimura et al., 2011)	Classification between vowels /a/, /u/ and control state	Brodmann ares 1, 2, 3, 4, 6, 9, 22, 39, 40, 41, 42, 44, and 45	Sparse logistic regression and Variation approximation (SLR-VAR)	61.2%
(Chi et al., 2011)	Classification of articulation class of imagined phonemes	All except occipital and far frontal regions	naive Bayes and linear discriminant analysis (LDA)	Significantly above chance rate
(García et al., 2012)	Multi-class classification between 5 Spanish words	F7, FC5, T7, P7 (Wernickes area)	Naive Baye classifier, Random Forest, SVM and Bagging-RF	Above chance level
(Wang et al., 2013a)	Classification of Chinese characters	ALL areas, left hemisphere (two setups)	Common spatial patterns (CSP), cross-correlation, and support vector machines (SVM)	Between 79.33% to 88.26%
(Wang et al., 2013b)	Classification of Chinese characters	ALL areas, left hemisphere (two setups)	Common spatial patterns (CSP) and support vector machines (SVM)	Between 73.25% to 95.56%
(Matsumoto & Hori, 2014)	Classification between Japanese vowels /a/, /e/, /i/, /o/, and /u/	All	CSP, support vector machine with Gaussian kernel (SVM-G), Relevance vector machine with Gaussian kernel (RVM-G)	SVM: 77%, RVM: 79%
(Sarmiento et al., 2014)	Classification between /a/, /o/ (open), /e/ mid, /i/, /u/ (closed) vowels	Broca's and Wernickes area	Support Vector Machines (SVM)	84%-94%
(Zhao & Rudzicz, 2015)	Binary classification to check presence of C/V, $\pm Nasal, \pm Bilab, \pm /uw/, \pm /iy/$	T7, FT8, FC6, C5, C3, CP3, C4, CP5, CP1, P3	deep-belief network (DBN) and support vector machine quad (SVM-q)	SVM-q: C/V:18%, $\pm Nasal$: 63.5%, $\pm Bilab$: 56.6%, $\pm /iy/$: 59.6%, $\pm /uw/$: 79.1%
(Yamaguchi et al., 2015)	Classification between: rock, paper or scissors and spring, summer, autumn, winter	Pre-motor cortex, supplementary and Broca's area	Hidden Markov model (HMM) and Gaussian mixture model	29%-100% difference between words.

TABLE 2.2: Studies in the literature which performed recognition of imagined speech using EEG signals in the last five years.

Reference	Task	Brain Area	Method	Performance
(Arjestan et al., 2016)	Classification of vowels, syllables and resting state	Not mentioned	common spatial pattern (CSP), empirical mode decomposition (EMD) and SVM	Syllables: 76.4% (best), Vowel: 76.6% (best)
(Sereshkeh et al., 2017)	Binary Classification between "yes", "no" and rest state	ALL	discrete wavelet transform and multi-layer perceptron (MLP)	"yes" vs "no": 63.2%, covert speech vs rest: 75.7%
(Nguyen et al., 2017)	Classification of vowels, short, and long words	F5, FT7, FC5, FC3, TO7, CP5, CP3, P5 (Broca's and Wernickes area)	Riemannian manifold, and relevance vector machine classifier	Vowels: 49%, Short words: 50.1%, Long words: 66.2%, S-L: 80.1%
(Paul et al., 2018)	Recognition of vowels in Hindi language	F3, F4, F7, F8, T3, T4, T5, T6, C3, C4, P3, P4	time domain features and support vector machine (SVM)	82.8%, 75.8% and 59.4%
(Saha & Fels, 2019)	Classification of Long words: Independent, Cooperate	All	Hierarchical Deep learning	79%
(Saha et al., 2019)	Binary classification to check presence of C/V, $\pm Nasal, \pm Bilab, \pm /uw/, \pm /iy/$	All	Hierarchical Deep learning	C/V:85.2%, $\pm Nasal$: 73.4%, $\pm Bilab$: 75.5%, $\pm /iy/$: 73.3%, $\pm /uw/$: 81.9%
(Panachakel et al., 2019)	Classification of 7 phonemic prompts and 4 words	C4, FC3, FC1, F5, C3, F7, FT7, CZ, P3, T7, C5	Discrete Wavelet Transform (DWT) and deep neural network (DNN)	57.15%
(Sharon & Murthy, 2020)	Binary classification to check presence of C/V, $\pm Nasal, \pm Bilab, \pm /uw/, \pm /iy/$	All	Correlation network	C/V:89.3%, $\pm Nasal$: 76.5%, $\pm Bilab$: 75.6%, $\pm /iy/$: 80.3%, $\pm /uw/$: 82.5%
(Bakhshali et al., 2020)	Recognition of 4 words: pat, pot, knew and gnaw	T7, C5, C3, C1, CP5, FC1, FC3, FC5, FC6, FT7, F1, F3, F5, F7, FT8, TP7	Cross-spectral density and Riemannian distance	90.2%
(Pawar & Dhage, 2020)	Recognition of words: left, right, up and down	Prefrontal cortex, Broca's area, Wernickes area right inferior frontal gyrus	Kernel Based extreme machine learning	multi-class: 49.77%, binary: 85.5%

2.9 Grammatical Classes in the Brain

It is known that grammatical classes of words used in a sentence effects its processing, moreover, in all languages words belong to the two most fundamental classes of nouns and verbs (Vigliocco et al., 2011; Yu et al., 2012). There have been many studies aiming to distinguish between nouns and verbs in the brain. For example, the work in (Tsigka et al., 2014) aimed to differentiate between processing of nouns and verbs in the brain during silently reading of noun/verb homonyms. In this study the author recorded MEG signals from participants while they read noun/verb homonyms. The processing of noun and verbs was associated with four components or events between 0-725ms after the stimulus onset. Neural activation of noun and verbs have also been investigated using even related potentials (ERPs), which are used in neuroscience to detect the onset of an event produced by a particular activity (Luck, 2005). Investigation using ERP reported verbs to be processed faster in the left-hemisphere compared to in the right hemisphere, while processing of nouns showed no such behavior (Xiang & Xiao, 2009).

The processing of verbs in the brain produced greater activation in comparison to processing of nouns mainly in two brain regions: the right posterior parietal areas and the centro-parietal regions. Some work has proposed stronger activation elicited by processing of verbs could be because of its association with motor activity (Preissl et al., 1995). Also, verbs are known to evoke stronger events (high amplitude) in the frontal and central cortical region of the brain which have motor cortices (Yang et al., 2017). On the other hand, nouns have been reported to have stronger visual association (object words) and activate left anterior, left temporal, and occipito-parietal cortices (Damasio & Tranel, 1993; Yang et al., 2017). Further, processing of nouns depends on a stream of neural substrate at the anterior temporal region (Preissl et al., 1995). Therefore, differences in processing of nouns and verbs have been associated with meaning of the words being processed. Processing of nouns and verbs in the brain at early stage (106-329ms) occurs at lower frequency range (24-98Hz) while higher frequencies (116-137Hz) are involved at later stage (411-430ms).

Although, many studies have investigated in order to distinguish between brain activities produced by nouns and verbs, to the best of our knowledge no work so far has tried to distinguish between EEG signals for mentally spoken nouns and verbs using machine learning techniques. Since the problem has not been solved using machine learning techniques, further research is crucial in order to discover more effective ways to distinguish between brain responses produced by these two main grammatical classes.

2.10 Thinking with Images

Do we think with speech or images ? Often visual imagination has been associated with speech and language. According to the proportional theory, the verbal processing and imagination in the brain takes abstract proportions which have neither verbal nor imaginary form (Petsche et al., 1992). *Mental imagery is an ephemeral (short termed) internal visual portrayal (representation) of an object or activity using the information stored in the long termed memory (Li et al., 2010).* Imagery of an object is said to activate the same part of the brain as the word for the given image (Xie et al., 2020). The neuro-anatomical regions used for visual perception and mental imagery have been found to be similar: the occipital, temporal, parietal cortices. However, different hemisphere specializes in activities associated with mental imagery. Conceptually, speech in mind can be considered as impression or visual imagery of sentences, this had been proven by the study (Suppes et al., 1999) which investigated about brain waves produced by visual images and their names. The study showed that for visual words and images, the brain produced EEG signals that are almost identical.

Recent study argued that thought production is comprised by meaningful linguistic and/or visual imagery representations (Amit et al., 2017). Whether written words, perceived images, imagined images, or pictures activate the same area in the brain or not has been subject of interest for several studies. Research around the brain's ability to change and re-enact spoken words and images has been proposed as mechanisms of larger scale neural network and complex top-down processes (Ganis et al., 1996;

Kosslyn et al., 2001). It has been shown that covert speech (imagined speech) is accompanied by visual imagery (Amit et al., 2017). Further, it is known that images placed towards the end of a sentenced are processed similarly to words (Ganis et al., 1996). However, this study also suggested a partial non-overlap of underlying neural mechanism activated during the two tasks (Ganis et al., 1996).

According to (Esfahani & Sundararajan, 2012), visual imagination can be divided into two important processes; creating the image and holding the image in the mind. In the work (Petsche et al., 1992), it has been suggested that the inter-hemispheric and the right fronto-temporal coherence occur in order to maintain the image during the visualization of a concept. Studies have shown that there are two types of imagery; holistic which refer to an overall representation of the object (an outline) and partial which is more difficult to restore or recreate as it focuses on certain features of the object. According to (Li et al., 2010), holistic imagery is more easily evaluated compared to partial imagery. Speech is a complex activity which involves different tasks simultaneously such as vision, mental imagery, auditory feedback of own speech and/or other's speech. It is important to understand that some parts of the brain are involved more in visual perception, whereas others, such as the anterior temporal cortex, is more related to complex visual memory (von Stein et al., 1999). This can influence the brain waves which are accompanied by these tasks. Therefore, in this theses speech in mind is studied under different modalities, such as covert speech and visual imagery i.e., mental imagery of an object.

2.11 Limitations of Previous Research

Despite the increasing interest from the research community regarding speech-based BCI systems, the area of silent speech interface is still in its infancy. Recognition of imagined speech using EEG signals is a difficult task and to design a BCI for communication there is a need to address gaps in the literature that have been highlighted i this section. The limitation of the past research have been discussed as follows:

- **Low accuracy**: Despite the increasing interest from the research community regarding speech based BCI system, the area of silent speech interface is still in its infancy. There are ERP based BCI systems such as P300 spellers, that allow text entry with a virtual keyboard using eye gaze, these work on P300 ERP. These spellers can be providing reliable speed with a information transfer rate of 5.32 bits per second (Chen et al., 2015). However, these systems are still relatively slow, does not recognize the word directly, and require high attention level.

- **Limited Vocabulary**: Most of the previous studies that used EEG signals of covert speech, focused mostly on imagined vocalization of phonemes, vowels */a/* and */u/* and/or syllables */ba/* and */ku/* (Arjestan et al., 2016; Brigham & Kumar, 2010a). Recognition of imagined syllables and vowels is motivating however, it cannot be used for a thought to text BCI application. Recent experiments have focused on recognition of mentally spoken words (Hwang et al., 2016; Nguyen et al., 2017; Sereshkeh et al., 2017; Wang et al., 2013b). However, these studies have used a smaller set of words to be vocalized covertly while recording EEG signals, limited to two or three words such as "*Yes*" and "*No*".

- **Experimental Design Choices**: The experimental design for some studies included mental repetition of the same word multiple times (Nguyen et al., 2017). There are two issues with this approach. Mental repetition creates temporal effects (Porbadnigk et al., 2009) and daily conversation does not include repetition of the same word multiple times. In addition to mentally repetition, the length of recording an EEG trials have been longer (Sereshkeh et al., 2017; Zhao & Rudzicz, 2015). In EEG activity, a longer epoch length provides better resolution to identify the small changes in the EEG signal (Levy, 1987), but longer epochs causes time delay before the new information is presented, reducing the response time of a BCI. This questions the reliability of the proposed methods, when it comes to analyzing the complex nature of imagined speech from EEG signals.

- **Lack of work with Grammatical Classes**: Studies have focused on recognition of individual words; however, these words belong to different grammatical classes such as nouns and verbs. All verbal communication have a primary object and characteristics associated with the object, which are linguistically reflected in nouns and verbs (Crepaldi et al., 2011). A communication based BCI could benefit from having a larger dictionary of words belonging to different grammatical classes. Therefore, there is a need for discriminating imagined speech EEG signal belonging to different grammatical class. To the best of our knowledge no work so far has used machine learning methods to distinguish between grammatical classes.

- **Recognition Methods**: In the area of Brain Computer Interface many techniques have been suggested in order to recognize the covertly vocalized speech. Further, most of the other research has been done as binary classification task, where recognition was performed between two covertly spoken words (Balaji et al., 2017), and (Sereshkeh et al., 2017). Even under such circumstances some techniques work better in recognizing words of varying length (long words vs short words), but the accuracy reduces when trying to recognize words of same length (Nguyen et al., 2017). Further, due to complex nature of EEG signals a more robust feature learning and classification techniques are required (Zhang et al., 2019c). As a result, it becomes apparent that a technique is required that recognize mentally spoken words irrespective of their length and which can be practically implemented to recognize multiple-classes.

2.12 Research Design and Methodology

This work aims to overcome the limitations of the previous studies. In order to achieve this, we conducted this research study in following stages:

- **Literature Survey**: To begin, a comprehensive literature analysis was undertaken on the imagined speech recognition from EEG signals evaluating previous studies in order to gain a thorough understanding of the experimental design for EEG experiment and techniques used in literature for analysis of EEG signals.

- **Recording New EEG Database**: We designed an experiment to record EEG signals for mentally spoken words. In the experiment participants were presented with only "words" rather than syllables and phonemes. The experiment contained larger vocabulary of words, with the words presented as stimulus belonging to two main grammatical classes of nouns and verbs. This made it possible to analyze of grammatical classes of mentally spoken words. Further, in the experiment subjects were asked to mentally speak the words only once as soon as the stimulus appeared on the screen, in comparison to multiple repetition of the words (Nguyen et al., 2017).

- **Deep Learning for Recognition**: Deep learning has achieved state-of-the-art performance in most recognition tasks and have the ability learn representations in the data automatically (Goodfellow et al., 2016). Therefore, in this work we have investigated deep learning methods for recognition of imagined speech from EEG signals. We also used other methods for recognition and shown improvement on those results using deep learning frameworks. High recognition rate was achieved by the proposed deep network and the response time of these algorithm is fast enough to be used in real world applications.

2.13 Summary

The background and literature review chapter presents a comprehensive overview of existing research on the BCI, language processing in brain, EEG signals, and machine learning methods. The chapter begins by an overview on signals produced in the human brain and language processing within the brain.

Further, the chapter provides a comprehensive background on brain computer interface and its application in the area of thought-to-speech technology. Following this brief overview of machine learning and deep learning methods was provided. This section of the chapter outlines deep learning techniques used in some of the past studies and more importantly in this research. In addition, the chapter discusses in detail previous studies and methods employed for recognition of imagined speech from EEG signals. Limitations of past research have been highlighted with respect to EEG experimental design and methods used for recognition of speech from EEG signals.

Chapter 3

Brain Signal Acquisition and Pre-Processing

This chapter presents a new EEG data compiled for overt speech, covert speech, visual perception, and visual imagination tasks. The chapter discusses factors taken in consideration when designing the experimental paradigm to record EEG signals from human participants. Further, pre-processing steps and feature extraction method are discussed. The feature extraction section highlights the importance of presenting information in time-frequency domain. The chapter is structured as follows; section 3.1, provides an overview of the guidelines followed for experimental design, hardware, and software setup. Section 3.2, provides an overview of the experimental procedure and different tasks involved in it. Section 3.3, provides an overview of the pre-processing, artifact reduction, electrode interpolation and EEG signal filtering. In section 3.4, feature extraction and time-frequency analysis is discussed.

3.1 Motivation

A thought-to-speech BCI can be means of communication for people suffering from neuro-muscular disabilities such as paralysis and locked-in syndrome. Furthermore, recent advances in invasive BCI technology have shown promising results with monkey playing computer games with thoughts (gaurdian, 2021). Language is the most common mode of communication among humans, therefore a language based BCI is most intuitive for communication. In addition, a language based BCI offers the user with more options compared to cognitive task such as motor imagination or selective attention (Alsaleh, 2019). In the past there have been many studies which explored the recognition of language in the brain using neural time series data (EEG, ECOG signals) (Alsaleh, 2019; Herff, 2016). Some have tried to investigate differences and similarities between speech modalities such as overt speech and imagined (covert) speech (Martin et al., 2014). Most of the studies so far have recorded brain activity during mental repetition of phonemes, and/or syllables. Whereas studies with complete words has been limited to mental repetition of words like "*Yes*" and "*No*" (Sereshkeh et al., 2017). From a linguistic point of view each word can be further categorize as a noun or verb, therefore investigating grammatical class of words can be helpful in designing an efficient language-based BCI system. However, due to lack of publicly available EEG database and limited studies with words, grammatical classes of mentally spoken words has not been investigated in BCI research. On the other hand, verbal thinking (imagined speech) have been linked to visual imagery (imagination) (Xie et al., 2020). Further, investigation of verbal thinking, and imagination is limited only to discriminating EEG signals during visual perception and imagined speech (Alsaleh, 2019). Past studies have proposed visual imagery as a potential input for a language based BCI system (Lee et al., 2019). However, to the best of our knowledge no EEG database is publicly available for visual imagery and imagined speech task.

Therefore, this chapter presents a new EEG database for investigating brain activity during imagined speech and visual imagination. The study recorded EEG signals for four modalities: overt speech (verbal thinking), covert speech (imagined speech), visual perception, and visual imagination tasks. In contrast to previous studies EEG signals for imagined and overt speech were recorded from a larger dictionary

of 10 words, containing five nouns and five verbs. This was done because in almost all the languages have nouns and verbs as the two most basic grammatical categories. To discriminate between words belonging to different grammatical classes, we choose 10 words shows in Table 3.2. Further, to investigate comparison between verbal thinking and mental imagery, EEG signals were recorded while the subject performed imagination an object or a picture. The presented imagined picture where similar to object previously presented as a word during imagined speech task. Furthermore, compared to previous studies (Nguyen et al., 2017; Sereshkeh et al., 2017), the present work used a more natural method for recording imagined speech by mentally speaking the presented words only once rather than repeating multiple time. This chapter discusses the experimental paradigm, and the system used in recording the EEG signals. This chapter also focus on feature extraction method used in order to obtain a more informative representation of the raw EEG signals (separated trials).

A well recorded EEG data is considered as half study completed. In order to design the experiment for recording brain signals for imagined speech, we followed certain guidelines to get rid of influences during recording such as environmental disturbance, body movements, and additional thoughts. The aim of the research is recognition of covert (imagined) speech from EEG technology. In order to achieve this goal, there were certain influential factors which were taken into consideration such as recording environment, length of a trails, and tasks to performed. Recording the EEG data is the most important task of any EEG based research, "a good data set has no substitute" (Luck, 2005). Pre-processing is only useful if the recorded data is of good quality. Following were the factors considered while designing and conducting the EEG recording experiment:

1. In order to maintain consistency of the psychological conditions, same words were used in both overt and covert speech tasks. The Psychological condition changed, but not the stimuli. This was done because different stimulus can lead to variation in psychological response and comparing two different cognitive task (imagined and verbal speech) becomes difficult.

2. The experimental conditions were varied within the trial blocks rather than changing them between the trial blocks. In other words, the stimulus was presented in random order rather than presenting the same stimulus (word) several times which have been suggested by the previous studies to produces temporal effects (Porbadnigk et al., 2009).

3. The EEG lab used for recording the brain signal had a moderate temperature, this was done because high temperature could lead to sweating, increasing in potential of the recorded signals.

4. There was a gap of one second between each stimulus presentation, this was done in order to avoid overlapping of the EEG waveform from individual trials. If the interval between stimulus is too short, then potentials from the previous stimulus might contribute to the reaction (potential) evoked by the new stimulus (Luck, 2005).

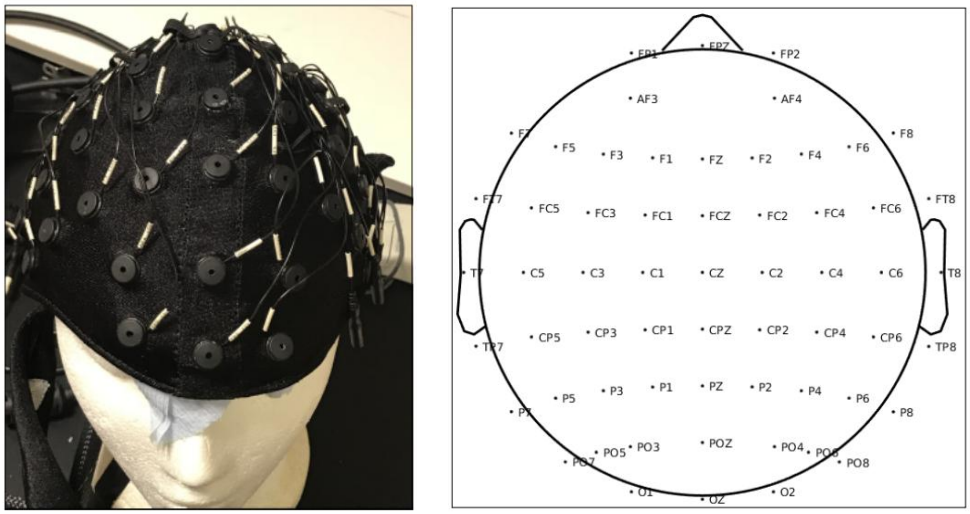
5. In order to reduce the number of eye blinks in the EEG signal, short duration trials were designed.

3.1.1 Recording Setup

Recording was performed during both the day and night times, in a LAB specifically designed for EEG based experiments. The room was maintained dark during all the recording sessions, lab was based in Brunel University London, United Kingdom. The subjects were asked to sit on a chair approximately one meter away from a computer screen, the screen was connected to two computers. One computer was used to present stimulus to the screen in front of the subject using the E-prime software, and the other system was connected to the EEG head-cap recording the EEG signals using Neuroscan Curry-8 software. The signals were recorded using the Neuroscan 64 channel Quik cap of extended 10-20 system. The subjects were asked not to make any kind of moment during the recording, if subject made any mistake or did not perform the task correctly, recording was restarted. Subjects were free to withdraw from the study at any time during the recording.

3.1.2 EEG Headset and Software

The E-prime software, available for behavioral research was used for designing the stimulus presentation for the experiment, labelling different classes, and activities. E-prime was used to presenting data in random order to subjects and timing the gap between two stimulus presentations. The stimulus was presented in random order to avoid impact of the temporal effects (Porbadnigk et al., 2009). The Neuroscan 64 channel Quik cap had 64 EEG electrodes (Ag/AgCL) and four non-EEG electrodes: VEOG (vertical electrooculograph), HEOG(horizontal electrooculograph), EKG (electrocardiogram), and EMG (electromyograph), which were used to detect and remove artifacts. The electrodes in the headset are conductive plates which picks up the electrical activity from the scalp. Further, to ensure good contact between scalp and the electrodes, an abrasive gel (conductive electrolyte) was injected in the electrodes which also helps in lowering the impedance of the recorded signal.



(a) (b)

FIGURE 3.1: (a) 64 channel Quik cap (b) Position of 64 electrodes on the scalp.

In order to reduce noise in the EEG signals and extract important events, the impedance of the skin should be reduced below $5K\Omega$ before attaching the electrodes (Luck, 2005). To achieve this, the outer layer of the dead skin cells was removed using alcohol pads from the face and head regions; mastoids, above & below the eye, and the temples. At the skin below and above the left eye electrodes VEOG and HEOG were placed, these vertical and horizontal channels were used to detect and remove eye blinks from the recorded EEG signals. Two reference electrodes (M1 and M2) were place at mastoid. It is difficult to remove dead skin using alcohol pads from the scalp because of the hair, therefore abrasive needle was rubbed gently on the scalp to displace the top layer of the dead skin cells. The EEG cap was connected to the synamp amplifier (shown in figure 3.2) which amplified and digitized the EEG data at 1000 $Hertz$ sampling rate, specification of the amplifier is mentioned in Table 3.1. Faraday cage was not used during EEG signals recording. The amplifier was connected to the system where signals were being recorded.

TABLE 3.1: Specifications of the amplifier used for recording.

Bandwidth	DC 3500Hz
Resolution	24 Bit
Common Mode Rejection (CMRR)	$> 110\ dB$
Noise (peak-to-peak)	$< 0.5\mu V\,(DC\ mode)$
Input Range	$400mV\,(DC\ mode), 1.9mV\,(AC\ mode)$
Sampling Rate (Hz)	$1000Hz$

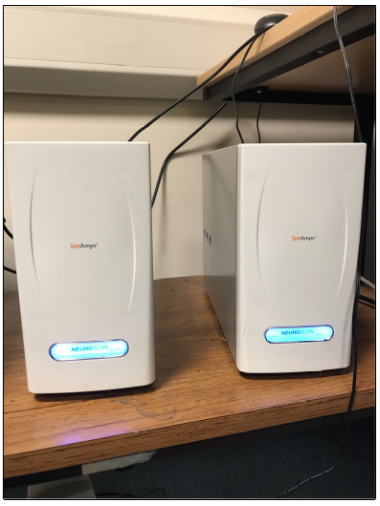


FIGURE 3.2: The amplifier (left) and power supply (right) that were used in digitization of the EEG data.

3.2 Experimental Protocol

3.2.1 Subjects

A new database was compiled by studying 17 healthy subjects, 10 male and 7 female. All subjects were fluent in English, 12 of them were not native English speakers. A subject can be regarded as a parameter which needs to be in an ideal position to produce noiseless EEG signal. All subjects finished recording without using any medication or drug. All the subjects were aged above 18 years, youngest subject was 22 years old and oldest was 70 years. None of the subject had any neurological or speech related disorder. Subjects were told in advance not to make any physical movement, because slightest movement made causes fluctuation in the EEG recording and muscle movement causes electrode amplitude to increase. Movement during overt speech were suggested to be minimized, eyes should be focused on the screen, eye blinks and horizontal eye movements were detected using VEO and HEO electrodes, clinching of teeth was also asked to be avoided. Any movement made during the recording session was noted, subjects were asked to stay relaxed, but alert during the recording. A subject sat on a wooden chair, 1 meter away from a computer screen where the stimulus was presented. Faraday cage was not used in the experiment. Participants were instructed to keep a normal pitch of their voice during overt production was told to be normal, but it varied for every subject. Intensity of pitch decreased as the trial progressed, but not in all cases. All the subjects provided feedback about the experiment and if they made any mistake during the covert tasks, the particular trials were removed. Some subjects mentioned problems with imagining the concepts, although there was no specific concept in particular. Due to noisy and artifact contaminated data for 5 subjects only 12 subjects were used in the analysis throughout this thesis.

3.2.2 Experimental Design and Task Performed

EEG signals for four different modalities (activities) were recorded from 17 participants between the age of 21 to 70 years. All the subjects were asked to perform four tasks in each session, each session lasted 13 seconds (shown in figure 3.3). For one subject, ten such sessions were recorded for ten different stimulus (words and images), with a break after five sessions. Therefore, recording for each subject lasted approximately 45 minutes. Every session was divided into four modalities (tasks). The modalities correspond to different categories which are most appropriate for a communication based BCI system: speech and visual imagery of the object/scene. However, most of the work in this thesis focus on speech imagery. The four task (modalities) participants performed were:

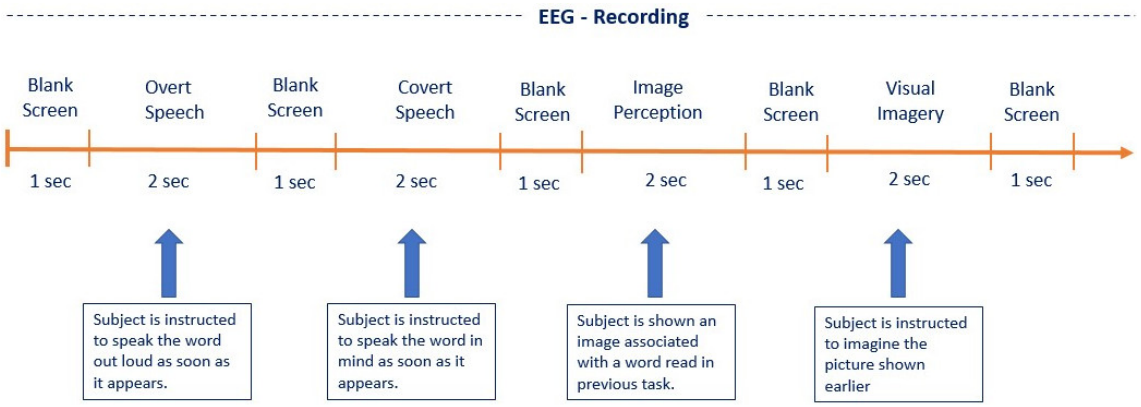


FIGURE 3.3: The sequence in which stimulus for four modalities was presented to the subjects. The figure represents a session in the experimental paradigm for a given word.

1. **Overt Speech**: This was the first task in each session which involved subjects to speak the word out loud as it appeared on the computer screen.

2. **Covert Speech**[1]: In this task, the subjects were asked to mentally read the word as soon as it appeared on the screen.

3. **Image Perception**: Subjects saw a picture of object or activity represented by the word presented in the earlier tasks (over and covert) of the same session. For example, if the word appeared in overt and covert speech task was "*Apple*", then image of an apple was presented on the screen for 2 seconds.

4. **Visual Imagery**: Subjects had to imagine the picture shown in the "Image Perception" task on the blank computer screen.

The Sequence of tasks

- First, a blank screen appeared for one second before the stimulus onset and then word was presented for two seconds, subjects were told to perform *overt speech task*. The word appeared in capital letters, black in color with white background presented on a computer screen 1 meter away from the subject.

- It was followed by blank screen for 1 seconds and the subject had to perform *covert speech task*.

- The covert speech task was followed by a blank screen for 1 second and *image perception task*, where the picture of the object appeared for 2 seconds with white background.

- The perception was followed by a blank screen and then the participants performed *visual imagery task*. The sequence in which the tasks were performed is shown in figure 3.3.

White background was chosen for the stimulus presentation to avoid potential due to visual stimulus (Luck, 2005). Ten different words and object/scene pictures were presented in total, each word, scene/object was presented 10 times, a total of 100 trials for each modality was recorded for a given subject. The words (stimuli) used in the experiment are shown in Table 3.2. The selected words were randomly chosen from the list of most frequently used words in spoken and written English (Leech, Rayson, et al., 2014).

[1] In some chapters of the thesis covert speech is also referred to as imagined speech and silently spoken speech.

TABLE 3.2: List of words used as stimulus.

Noun	*"Apple"*	*"Orange"*	*"Bottle"*	*"Football"*	*"Laptop"*
Verb	*"Carry"*	*"Laugh"*	*"Run"*	*"Swim"*	*"Write"*

3.2.3 Characteristics of the Recorded EEG Data-set

TABLE 3.3: Characteristics of the recorded EEG data.

Parameters	**Details**			
No of Subjects	17 (10 Males & 7 Females)			
No of Electrodes	64			
Sampling Rate	1000Hz			
Tasks per Subject	**Overt Speech**	**Covert Speech**	**Image Perception**	**Visual Imagery**
Trials	100	100	100	100
Classes	10	10	10	10
Class Type	Spoken Words	Imagined Words	Watching Pictures	Visual Imagery
Stimulus	Words	Words	Pictures	Blank Screen
Order of Presentation	Random	Random	Random	Random
Trials per Class	10	10	10	10
Trial Length	3 *sec*	3 *sec*	3 *sec*	3 *sec*
Grammatical Classes	2	2	-	-
Grammatical Class Type	Noun & Verb	Noun & Verb	-	-
Words for Noun Class	5	5	-	-
Words for Verb Class	5	5	-	-
Trials per Grammatical Class	50	50	-	-
Reading of Word	Once	Once	-	-

This section presents Table 3.3, which summarizes important parameters of the recorded EEG data. As shown in the table four tasks were performed by the subjects during EEG recording. Each task had 10 different stimulus presentation, therefore each task had 10 classes. The stimulus for each task was presented randomly as mentioned in section 3.1.

3.3 Public Database

In this thesis, analysis in Chapter 7 used a publicly available *Kara-One* EEG dataset (Zhao & Rudzicz, 2015). The data set contained EEG signals for four imagined words. The work in (Zhao & Rudzicz, 2015) recorded EEG signals along with the audio and facial information for imagined speech from 12 participants, of which 10 participants were native English speaker. EEG signals were acquired using 64 channel Neuroscan Quik cap with electrodes placed in the 10-20 system. Conductive electrolyte was injected in the electrodes to improve connectivity. All the EEG signals were recorded at a sampling rate of 1000Hz using the SynAmps RT amplifier. Further, a Microsoft Kinect camera was used to record participant's facial information and speech. Subjects were presented with 11 prompts (stimulus) with 7 phoneme/syllables (*/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/*) and 4 words (*pat, pot, gnaw, and knew*). For each word and phoneme/syllable, 12 trials were recorded. The EEG dataset was recorded from 12 participants, however data from four participants was discarded due to corrupted signals. Each trial consisted of four stages:

1. Resting state, where the participant was instructed to relax for 5 seconds.
2. Next the stimulus appeared (text) on the computer screen and auditory utterance associated with the stimulus.
3. Imagined speech stage where the participant imagined speaking the prompt for 5 seconds.

4. Speaking stage where the subject read the stimulus out load (overtly).

As this work focus on recognizing imagine speech from EEG signals, therefore audio and facial information present in the database were discarded. More information about the database can be found in (Zhao & Rudzicz, 2015). The words used as stimulus to record EEG signals are presented in Table 3.4.

TABLE 3.4: Words presented as stimulus during recording of EEG signals for covert speech.

Noun:	"*Pat*"	"*Pot*"
Verb:	"*Gnaw*"	"*Knew*"

3.4 Pre-processing

Recorded signals are contaminated by noise and artifacts, such as eye blinks, eye movements, breathing and muscle movement. In order to avoid physiological noise caused by muscle movement, subjects were asked to refrain from any kind of moment during the recording. However, some noise due to eye movement and eye blink cannot be avoided when a visual stimulus is presented. Hence, pre-processing was applied to avoid noise in the data. Further, the recorded EEG data is continuous, therefore another objective of pre-processing is to separate the data into sets of trials. Pre-processing was done in following steps:

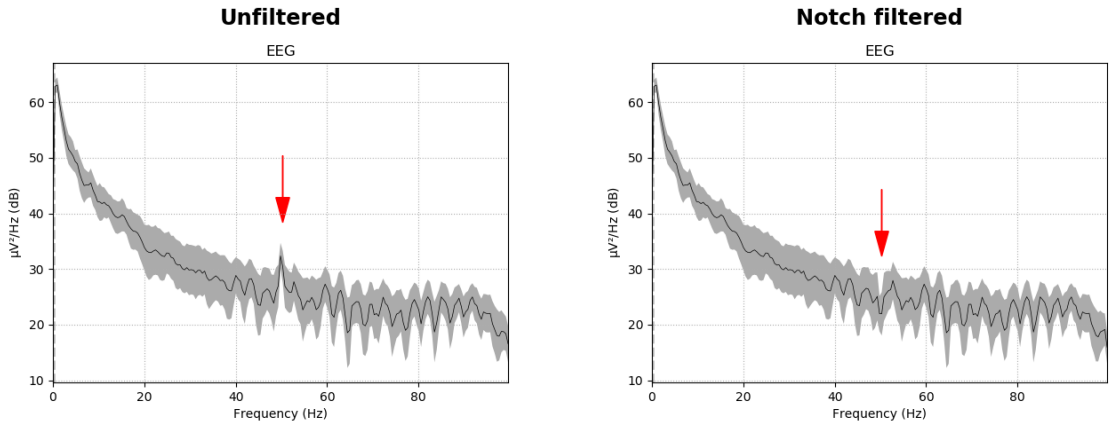
3.4.1 Baseline Correction

To represent a signal in two dimensions a constant base level is required, experimental procedure of recording the data takes hours, which can result in change in noise level, environmental or brain activity (Staljanssens, 2013). Baseline of the raw data was corrected in real time and during offline processing, using Neuroscan Curry 8 software. The mean per trial for each channel was calculated of the pre-trial data and the mean value was subtracted from the all the points in the data. This was done because the values prior to the stimulus onset are considered to have no event related activity. Although, it is essential to recognize that the pre-stimulus values are not completely neutral. Baseline correction was the first step in pre-processing. If it is performed after filtering or artifact rejection as it can impact important events in the signal.

3.4.2 Filtering

In EEG research, time domain information has always been point of interest, but temporal information of the EEG signal may get distorted by filtering. Filters can affect the timing and the amplitude of the EEG by adding artificial peaks. Therefore, we avoided using analog filters because the analog filters tent to change the latency of different frequencies by different amount, but this is not the case with digital filters (Luck, 2005). For this reason no filtering was performed before digitization of EEG signals. It is important to find an appropriate method to eliminate noise from the EEG signals.

Noise or artifacts, such as slow voltage shifts, occurs at low frequencies below 0.1Hz, hence a high-pass, zero-phase, finite impulse response (FIR) filter of order 62 of $0.01Hz$ was used to filter the raw EEG signal. Similarly, a notch filter was used to remove the harmonics of the $50Hz$ line (shown in figure 3.4). Most of the noise at the higher frequencies, such as noise due to muscle movement, was eliminated by means of EMG electrode. Artifacts due to eye movements, were corrected by measuring the peak-to-peak voltage of the VEOG signal along with the threshold voltage of $\pm 100\,\mu V$ and created an average artifact corresponding to eye blink. The averaged artifact from the VEOG channel is subtracted by eye



(a) (b)

FIGURE 3.4: Removal of 50Hz line noise using notch filter from the EEG signal: (a) before (b) after.

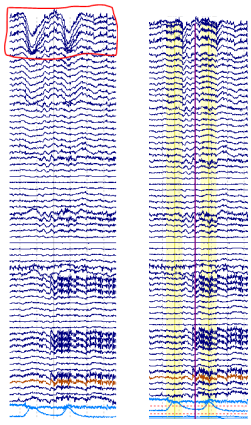
blink contaminated EEG channels, for the same time points when the VEOG peak-to-peak voltage was above the threshold, (shown in figure 3.5).

3.4.3 Electrode Interpolation

All the noisy electrodes were marked during the time of recording and later were inspected visually. Electrodes with poor connection to the scalp and presence of very high voltage around $\pm 100\,\mu V$ were interpolated i.e., the information in the noise contaminated electrode was estimated based on the EEG activity from two neighboring electrode.

3.4.4 Artifact Detection and Correction

Another artifact correction method was removing or correcting the bad sections of the continuous raw signal, which was corrected in the Curry 8 software. A threshold of $\pm 100\,\mu V$ was used to detect artifact in all channels and by checking the voltage in all the channels with respect to threshold voltage. The artifact contaminated region of the signal (from all electrodes) was corrected using principal component analysis (PCA). The data from several EEG epochs is averaged to create average artifact and PCA is performed on the averaged artifact which reveals components representing the artifact. Components contributing to artifact were removed to acquire the corrected EEG signal.



(a) (b)

FIGURE 3.5: Continuous data with: (a) Eye blink; (b) Corrected using VEOG electrode.

3.4.5 Trial Extraction

The continuous data contained labels and markers, where markers show the start of a trial. Labels revealed the task (section 1) performed and class (here class refers to the different stimulus (words)) presented in the given trial. Using these markers and labels the continuous data can be divided into separate trials. Trials belonging to a particular task and class were used as a set as shown in figure 3.6. Each class (word) had a total of 10 trials for a given subject. However, due to artifact reduction some of the subjects had less than ten trials. Each trial length was three seconds, with 0.5 seconds pre-stimulus and 2.5 seconds post-stimulus onset time. The trials were extracted using the MNE-Python library (Gramfort et al., 2013).

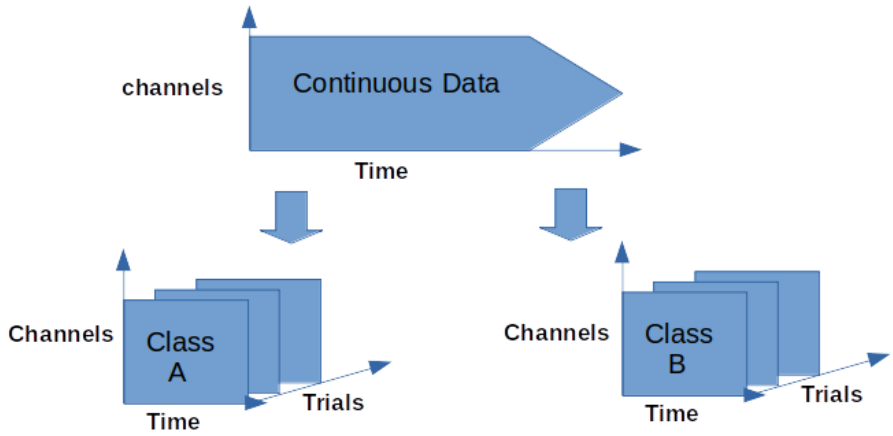


FIGURE 3.6: The process shows separate trails extracted from continuous data. This method was done separately for all the four tasks (section 1), class here refers to the word and object/scene picture presented as stimulus during the experiment.

3.5 Feature Extraction

Feature extraction is the process of extracting important features (characteristics) containing the information of interest from the original signal. The objective of feature extraction is compression of information contained in the EEG signal by eliminating information not relevant to the task or to the imagined word. It is most essential part of a BCI system to achieve high recognition because recognition rate of the system will degrade if useful features are not chosen. This work mainly used two types of features: Temporal features in chapter 4, 5 & 6, and Time-Frequency features in chapter 5, 6, & 7. Although frequency features have not been directly used for classification, we used frequency features for electrode selection in chapter 6.

There are many methods available to extract features from the given signal in order to achieve better classification rate. Feature extraction methods can be divided into three main categories:

1. **Temporal**: Signal amplitude can be considered as the most standard form of temporal feature (Lotte, 2008). This feature can be used after preprocessing of the raw signal using digital filters such as FIR and IIR filters. In many studies mean of the data from different trials is used to create the ERP and understand the underlying effects such as P300, N400 and ERD-ERS in case of motor imagery.

2. **Frequency**: Power spectral density (PSD) is a frequency domain feature; it is one of the most widely used methods in the field of BCI-EEG. Fourier transform of the raw signal is taken and power at specific frequency can be used as input to the classifier.

3. **Time-Frequency**: Wavelet transform, short time Fourier transform (STFT) and Hilbert Huang transform (HHT) are some examples of T-F methods to extract features from the signals. Properties such as energy, entropy, mean, and median can be evaluated from the different frequency bands at specific time instances, reducing the information of signal for classification.

Temporal features alone can suffer several limitations due to the non-stationary nature of EEG signals. For example, signals produced by different stimulus (different words in this research) may be similar in terms of head maps and neural activity, but dissimilar in terms of frequency characteristics (Roehm et al., 2004). On the other hand, frequency information alone does not provide any information about the temporal characteristics of the signal. Therefore, in this study EEG signals were transformed into time-frequency domain.

3.5.1 Time Frequency Analysis

Representation of EEG signals in a desired form allows for more accurate analysis. Characteristics of EEG signals can be studied using time-frequency analysis which provides information in both time and frequency domain. STFT is one of the most popular time-frequency analyses, it provides spectral information of the EEG signal at each time point. To calculate the STFT the original signal is segmented into short-time frames by performing temporal overlapping (Zabidi et al., 2012). Transforming raw EEG signals to spectrograms offer advantage because STFT provide important time-frequency information about the underlying activity. Spectrograms are capable of capturing energy modulation across spectro-temporal patterns which can be used to distinguish between different cognitive responses (Cohen, 2014). Spectrograms are the most common form of time-frequency representation, they can be calculated by applying STFT on the raw EEG signal and mapping it into the two-dimensions of frequency and time.

The data needs to be tapered during STFT, this is done in order to avoid the discontinuities in the signals or otherwise known as leakage. There are several windowing options, however Hann window was chosen because it tapers the data from zero at the beginning and at the end (Cohen, 2014). Figure 5.3 is produced from STFT method, length of the window is taken to be 256. Two consecutive windows had a temporal overlap of 87%. Shorter window length enhances temporal resolution however reducing the frequency resolution, therefore lower frequencies below 5Hz were not included in the analysis. Also, better temporal resolution will help detect onset or events that might be present in the signal. The STFT equation is given as:

$$S(t,f) = \sum_{n=0}^{N-1} s(n + tN'w(n))^{-j\frac{2\pi}{N}nf} \tag{3.1}$$

where $f = 0,1,....,N-1$, $S(t,f)$ is the time-frequency spectrogram, N is the window length, N' is the overlapping of the time window, $w(n)$ window method of N point sequence. Spectrograms can be calculated from (3.1) as:

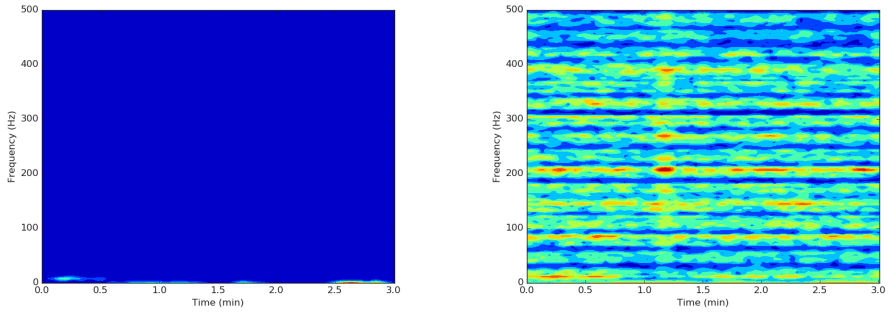
$$A(t,f) = \frac{1}{N}|S(t,f)|^2 \tag{3.2}$$

where $A(t,f)$ is the magnitude of $S(t,f)$ obtained in (3.1).

3.5.2 Baseline Normalization

EEG signals suffer from the 1/f phenomena, which means low power representation at the higher frequencies. This is due to the structure of neural network in the brain and slow speed at which neurons communicate (Demanuele et al., 2007). There are several limitations caused by it:

1. Incorrect power representation across the different range of frequencies.



(a) (b)

FIGURE 3.7: Spectrogram before and after baseline normalisation. Spectrogram before normalization (a) provide no useful information. However, (b) after baseline normalization, energy at different frequencies can be used to discriminate between mentally spoken words.

2. Comparison between two different frequency bands become difficult and probably incorrect due to lower power representation at higher frequencies.

3. Computing results over different trials and subject could be difficult.

4. Events can be misclassified based on background activities.

There are several methods to overcome the problem of low power representation at high frequencies (Demanuele et al., 2007; Hu et al., 2014). In this work, the 1/f phenomena were resolved using a short time window of 0.3 *sec* from pre-stimulus time period 0.5 *sec*, from -0.4 *sec* to -0.1 *sec* was averaged over all trials of the given word. Pre-trial time-period was considered to be the time-period with no event related activity, but due to the effects of windowing (overlapping pre-trial and post-trail time periods) a safer time window (-0.4 to -0.1 *sec*) was chosen. Baseline activity was divided with spectrogram overall time points and decibel conversion was performed. A baseline vector was calculated, comprising of frequencies averaged over the baseline time window.

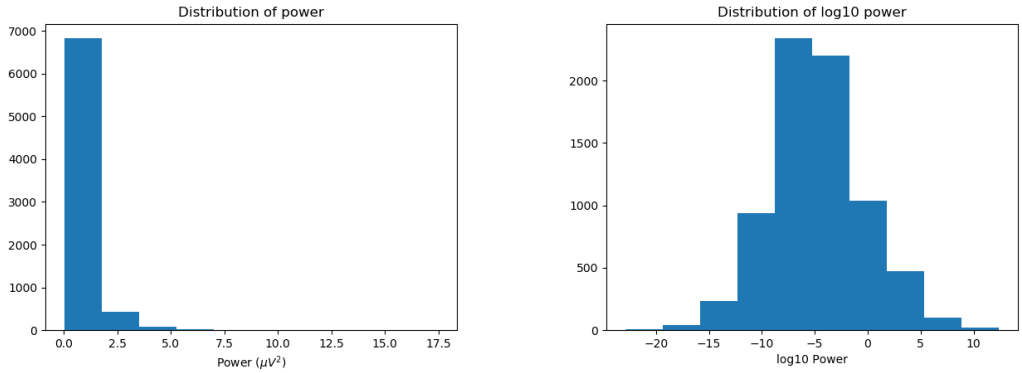
$$B(f) = \frac{1}{t_2 - t_1} \sum_{t_1}^{t_2} S_x(t, f)$$

where $B(f)$ is the baseline vector averaged across time axis in spectrogram, and the normalized spectrogram (in decibels) is defined as

$$S_{dB}(t, f) = 10 \log_{10}\left(\frac{S_x(t, f)}{B(f)}\right) \tag{3.3}$$

As seen in figure 3.7 (a), spectrogram that has not been normalized shows low power at higher frequencies (blue region). On the other hand, the normalized spectrogram, shown in figure 3.7 (b), exhibits power variations at different frequencies. Baseline normalization also helps untangle the task related activity from the non-task related background information. Power in spectrograms normalized using $\log_{10}$ is normally distributed as shown in figure 3.8.

The baseline normalized spectrogram show regions with high power (red patches) and other regions (blue patches) in the spectrogram refer to decrease in energy. Important information in the signal can be shared among different frequencies, therefore extracting components of interest from spectrogram can bring forth important information (features) and help distinguish between different classes.



(a) (b)

FIGURE 3.8: Distribution of power (a) shows non-normalized power distribution; (b) taking logarithm-base-10 of spectrograms distributes the data normally.

3.6 Summary

This chapter presented a new EEG dataset recorded for analysis and recognition of speech in brain. EEG signals for overt and covert speech were recorded from human participants. EEG signals for visual imagination and perception were also recorded, which has been associated with language-based thinking. In comparison to previous research, this study recorded EEG signals for larger vocabulary (i.e., 10 words) and the subject repeated the word mentally and oral only once. The words used as stimulus belong two grammatical class of noun and verb, allowing to investigate nature of EEG signals produced by two different grammatical classes. Important factors that were taken into consideration in order to design experiment for recording EEG signals are highlighted. The experimental protocol for all the modalities is presented. Further, in order to improve the signal quality for further analysis, EEG signals were processed using artifact rejection method, digital filtering, and electrode interpolation. Finally, the EEG signals were transformed into time-frequency form which can better represent characteristics of EEG signals in both temporal and spectral domain. The following chapter will investigate different techniques to recognize imagined speech using EEG signals.

Chapter 4

Discriminating between Imagined Speech and Other Speech Related Activities

In this chapter, classification was performed between imagined speech and two other speech related activities: visual imagery and overt speech. In the first part of this chapter, the temporal dynamics were investigated in the EEG signals recorded during covert speech, overt speech, and visual imagery task using temporal decoding method. Investigative analyses were performed using multi-variate pattern analysis (MVPA), to find the best time window and frequency range for distinguishing between two (covert & overt speech and covert speech & visual imagery) cognitive tasks. In the second part of this chapter, a framework is proposed for recognizing the cognitive activity from EEG signals. The proposed framework uses the K-means clustering for electrode selection and the convolutional-attention network for classification.

4.1 Introduction

Words or concepts in the brain can be produced by many different cognitive activities, such as imagined speech, visual imagery (imagining the picture associated with the word), and overt speech. In chapter 3, EEG signals were recorded for imagined words, spoken words, and imagination of the concept associated with word. The neural events produced by in these activities are triggered by the same concepts. This chapter investigates whether the events produced by different cognitive activities are dissociable. In order to achieve this, EEG signals from three cognitive activities are used in this chapter; covert (imagined) speech, visual imagery, and overt speech. This is important as covert (imagined) speech based BCI system should be able to distinguish between EEG signals produced by different cognitive activities.

When exploring mental representation in the brain, we refer to the brain's ability to create visual imagery during cognitive tasks (Petsche et al., 1992). These cognitive tasks are processed in a top-down manner i.e., visual images generated from the memory (Dentico et al., 2014). On the other hand, bottom-up processes are cognitive processes based on modal processing of external stimulus. These processes are taking place during perception and have shown to not vastly differ in terms of brain activity, for example when perceiving written words or images (Dentico et al., 2014; Petsche et al., 1992; Schendan & Ganis, 2012).

Although, these cognitive tasks have been proven to show dissociability between processing of overt and covert speech (Tian & Poeppel, 2010), the authors found that different areas in the brain were activated for the overt speaking task, responses were recorded on the bilateral frontal areas, assumed to be in relation to activity in the primary motor cortex, due to tongue movement when speaking aloud. In contrast, the parietal cortex was activated during the covert (imagery speech) task (Tian & Poeppel, 2010). However, the study did not distinguish the two tasks based on the temporal dynamics of neural activity. Similarly, even if visual imagery and covert speech have been shown to recruit similar or overlapping brain circuits and neural dynamics (Martin et al., 2014; Xie et al., 2020), yet the temporal dynamics identified by using electroencephalography (EEG) for exploring these neural mechanics are still less utilized and understood (Xie et al., 2020).

Therefore, to examine the temporal events of these processes, the dynamics in EEG signals are investigated for covert speech (imagined speech), visual imagery, and overt speech tasks. MVPA was used on the raw EEG data to find the most discriminative time and frequency information. In addition, a framework is proposed for discriminating between EEG signals produced by visual imagery & covert speech and overt & covert speech tasks using K-mean clustering and the CNN-attention network. The proposed framework used K-means clustering to select electrodes and evaluated the network on electrodes in different clusters. The electrodes selected using K-mean achieved an average accuracy of 82.9% and 77.7% for 12 subjects.

4.2 Dataset

The electroencephalogram (EEG) signals for three different modalities (tasks) were collected from 12 subjects. The three tasks were overtly spoken words, covertly spoken words, and visual imagery of the images associated with the words who were fluent in the English language. The three tasks are defined as :

- **Overt speech**: This was the first task in each session which involved subjects to speak the word out loud as it appeared on the screen.
- **Covert speech**: In covert speech task the subjects mentally read the word as soon as it appeared on the screen.
- **Visual imagery**: In visual imagery task subjects had to imagine the picture shown in the "Image Perception task" which was associated with the words in the overt and covert speech task.

Ten words, five nouns, and five verbs were randomly shown to the subjects. Each subject was given ten trials for each word, giving a total of one hundred trials for each modality. A trial lasted $3sec$ and was recorded at a sampling rate of $1000Hz$, resulting in an EEG signal of length $3000ms$ per electrode. For further details on the experimental protocol and EEG recording please refer to section 3.2.

4.3 Methods

In this chapter, several techniques are used for investigation of the difference between neural activity of different cognitive tasks. Methods such as MVPA and temporal generalization (TG) were used to investigate the neural behavior in time domain. EEG signals are recorded from multiple electrodes, processing of such high dimensional EEG data requires more resources and training time. Therefore, this analysis used K-means clustering to select electrodes and evaluated the proposed network on different group (clusters) of electrodes. This section provides an overview of these methods.

4.3.1 Temporal Decoding using MVPA

The central goal of neuroscientific research is to understand the processing of cognitive task in time (King & Dehaene, 2014). Therefore, in this chapter, MVPA was used to distinguish between three cognitive tasks. MVPA has the ability to decode information from multi-dimensional dataset and has been used extensively in neuroimaging research (King & Dehaene, 2014). MVPA has been used in several fMRI studies to provide insight about patterns triggered in different brain regions by the cognitive behavior. It processes the information at different time points separately which increase the signal-to-noise ratio, this offers advantage in comparison to event related potential (ERP) estimated by averaging (King & Dehaene, 2014). MVPA uses sliding window approach which can help detect temporal segment for optimal decodability.

4.3.2 Temporal Generalization (TG)

Further, the idea of temporal decoding can be extended to generalization across time, also known as temporal generalization (TG) (King & Dehaene, 2014), which can predict if the decodable information reoccurs at different time segments. TG is a comprehensive strategy for recognizing how mental processes are adjusted and notably altered (King & Dehaene, 2014). TG method trains a classifier to predict stimuli or conditions using the brain activity on a single time point in the EEG data and then evaluating the trained classifier across all conceivable time points. This is repeated for all potential training points in order to get a complete matrix of accuracy for each potential train/test points combination. The findings of this analysis can indicate when brain activity patterns are consistent (i.e., when the trained model works well across multiple time windows) and when they are inconsistent, allowing the user to follow the neural representations over time (Fyshe, 2020). A significant advantage of this technique is the use of pattern classifiers to time-resolved brain activity data (e.g., EEG or MEG) recordings. Accordingly, the TG technique determines at which moment a specific mental content becomes decodable within the brain activity, characterizing the time course of cognitive events. Most importantly, the trained classifiers generalize across time and illuminate the temporal structure of the various stages of information processing (King & Dehaene, 2014). This is a major advantage, as understanding how mental activities unfold might be considered one of the core goals within cognitive neuroscience (Dehaene & King, 2016).

4.3.3 K-Means Clustering

Clustering is a popular exploratory data analysis tool for understanding the structure of the data. It can be defined as the process of discovering sub-groups within the data, e.g., that data points belonging to the same sub-group (cluster) are similar, whereas, the data points belonging to different sub-groups (clusters) are dissimilar. For clustering, the K-means algorithm (MacQueen et al., 1967) was selected. The K-means algorithm partitions the data into K non-overlapping clusters. The K-means algorithm tries to find a partition between the clusters by minimizing the squared error between the mean of a cluster and the data points in it (Jain, 2010). The squared error is given as:

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \tag{4.1}$$

where x_i, $i,...,n$ are d-dimensional points that are to be divided into K-clusters. c_k is the k^{th} cluster and μ_k is the mean of the cluster c_k. The K-mean algorithm aims to minimize the sum squared overall the K clusters, given as:

$$J(C) = \sum_{k=1}^{K} \sum ||x_i - \mu_k||^2 \tag{4.2}$$

K-means begins to divide the data into K clusters and assigns patterns to clusters in order to minimize the squared error.

4.4 Decoding Temporal Points with Task Discriminative Information

4.4.1 Distinguishing between Covert Speech and Visual Imagery Tasks

For decoding analysis, the length of the EEG signal used was 3000ms, -500 ms from stimulus onset to 2500 ms (1000Hz sampling rate). A logistic regression (linear) classifier was trained and tested on each time point in the EEG signals separately, to classify all the trials under binary classification conditions. Therefore, the input to the classifier was a vector of length 64, i.e., time point from 64 electrodes. The evaluation was carried out in a leave-one-out cross validation manner, with 80% of the data used for training and 20% for testing. The decoding accuracy was averaged across all subjects for each time point, which provided a grand average temporal decoding vector. The temporal decoding accuracy is

shown in figure 4.1. The decoding accuracy provided information about the time when EEG signals for two cognitive activities could first be distinguished, and the time when each cognitive task was most distinct in terms of linear separability. For this analysis the performance of the classifier is assessed at chance level of 55.0% rather than 50.0%, this was done to have a stricter threshold for the performance evaluation.

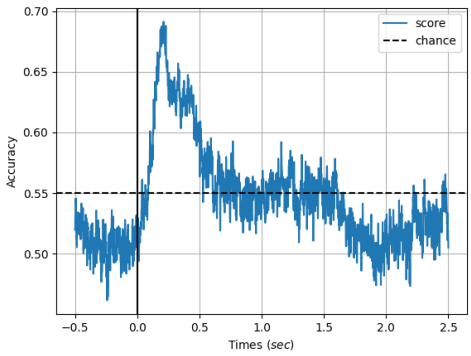


FIGURE 4.1: The timing of grand average decoding was above chance level at 69 *ms*, with a peak at 138 *ms* and 220*ms*. A smaller peak above chance level was observed between 2131 *ms* and 2457 *ms*.

Prior to and immediately following stimulus presentation, the grand average decoding accuracy fluctuated about chance level. At 69 *ms* the accuracy increased quickly and attained significance, followed by peaks at 138 *ms* and 220 *ms* and then a progressive decline.

Further, MVPA was used to investigate the contribution of different frequency bands in discriminating covert and visual imagery task. Raw EEG signals were band-pass filtered using FIR filter for different frequency bands: delta (0.5-3.5Hz), theta (4-7Hz), alpha (8-12Hz), beta (13-30Hz), gamma (31-80Hz), high-gamma (81-120Hz), and ripple (121-250Hz). The results are shown in figure 4.2. As can be seen, the highest accuracy of temporal decoding attained for each frequency band varies. The delta band had a peak at 408 *ms* and second peak at 2473 *ms*, accuracy at all time points for delta band was above chance level although the highest accuracy was 66.4%, whereas the theta band achieves the highest accuracy of 81.4% at 138 *ms* and a second peak at 2263 *ms*. However, the accuracy drops to chance level at around 780 *ms*. Accuracy for the alpha band (79.1%) peaked at 199 *ms*, followed by a decline to chance level shortly after 500 *ms* of the stimulus onset. Peak accuracy for the beta band was lower than for the alpha and theta bands (69.5%) at 120 *ms* and there was not a second peak after 200 *ms* as observed in other frequency decoding accuracy. The peaks in this decoding measure correspond to temporal points when the signals produced in the brain for two cognitive tasks (visual imagery and covert speech) varied the most. The decoding accuracy indicated linear separability of the two conditions in the different frequency bands, which means that visual imagery and covert speech for an object (concept) varies mostly in the theta frequency band. The highest accuracy was achieved by the theta band, the alpha and beta band also achieved high recognition rate. Higher frequencies, i.e., gamma and high-gamma frequencies achieved chance-level recognition, with no particular temporal region of high accuracy.

4.4.2 Temporal Behavior of Neural Activity for Visual Imagery and Covert Speech Task

The nature of neural oscillation was also estimated over time for covert speech and visual imagery task. A logistic regression classifier was trained to discriminate between covert speech and visual imagery from the EEG signals at time point t_x and tested on a time point t_y. In this manner a vector of accuracy (decoding) was obtained for each training time point, these vectors were concatenated to create a TG matrix. Each row of the TG matrix corresponds to the time the classifier was trained and each column to the time it was tested. Classification for all subjects was performed separately and the TG matrix was

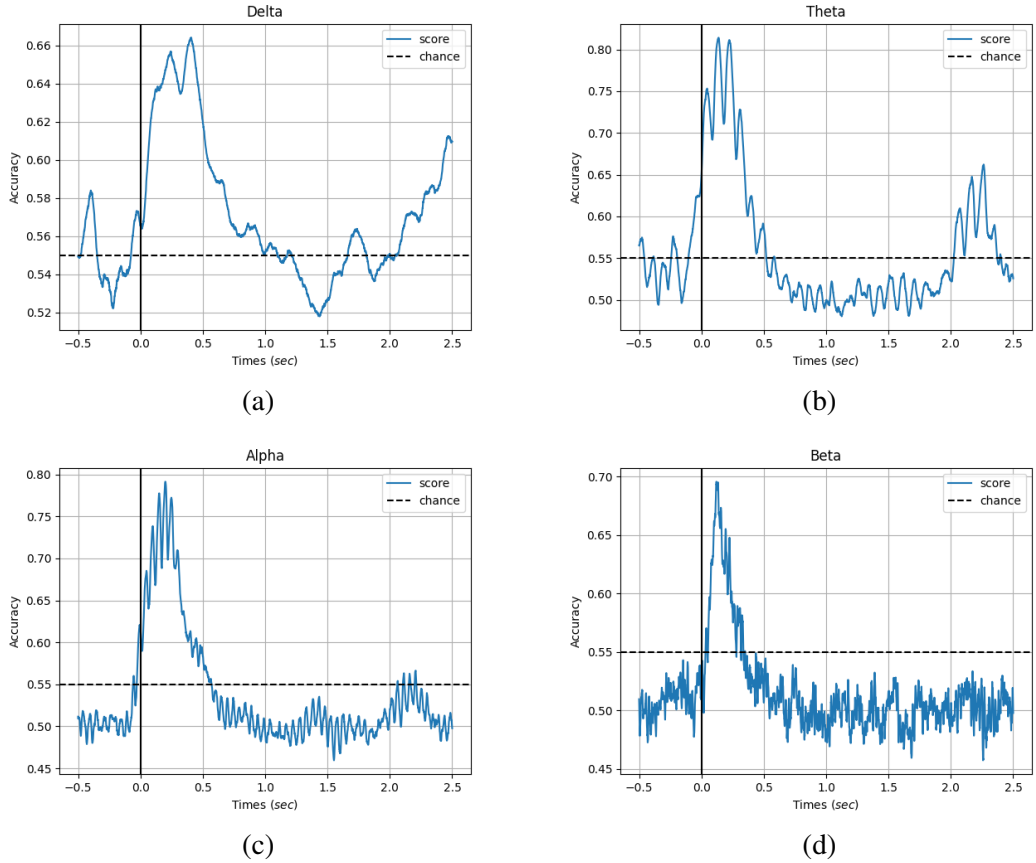


FIGURE 4.2: Temporal decoding between covert speech and visual imagery for different frequency bands. (a) Delta (b) Theta (c) Alpha (d) Beta.

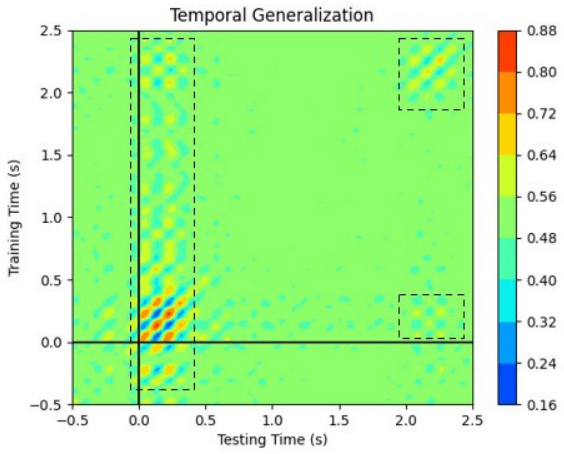


FIGURE 4.3: A logistic regression classifier was trained at time point t_x and was evaluated on its ability to generalize on another time point t_y. In this manner the classifier was evaluated for all trials and at all time points. The figure shows time-time decoding matrix averaged over all the subjects.

averaged. As best decoding results were achieved using the theta band activity, therefore calculating the time generalization matrix, we used EEG signals from the theta band. The TG matrix is shown in figure 4.3.

The brain activity between 0-0.5 *ms* in the TG matrix showed oscillatory behavior, i.e., an oscillatory

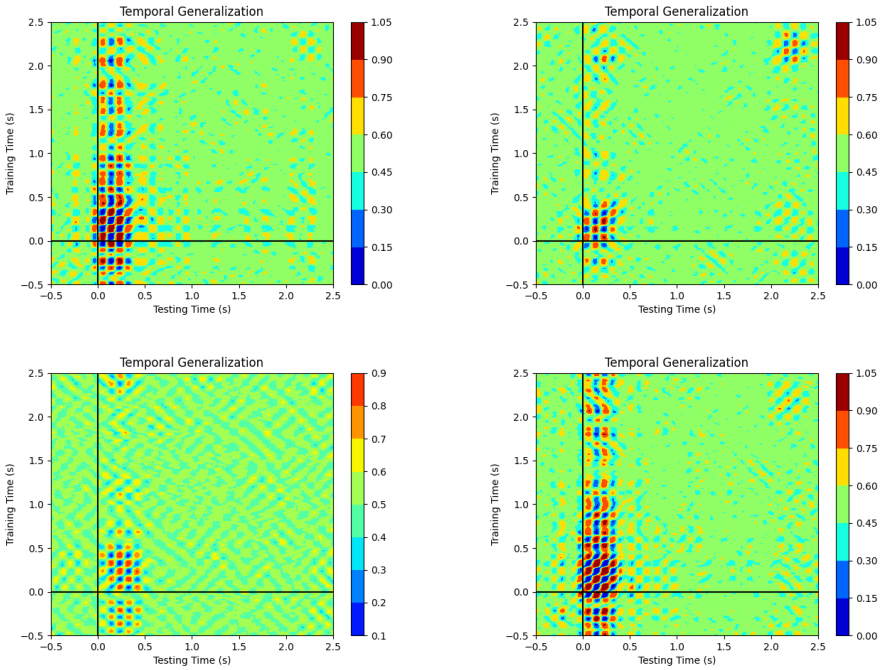


FIGURE 4.4: Temporal generalization matrix for some subjects showing strong oscillatory behaviour.

or reversing behavior of brain signals results in transient performance below chance level. Sustained oscillatory activation was observed. The classifier performed effectively, (55-64% decoding accuracy) when trained for time points between $\sim$ 600–2400 *ms* and tested on time points between $\sim$ 0-300 *ms*. This effect was time-limited to 500 *ms* testing time, but was also observed later between testing time points $\sim$ 2000-2400 *ms*. Some participants showed stronger oscillatory behavior during 0-500 *ms* after the stimulus onset. The theta band oscillation is known to indicate periodic reactivation ("replay") of the preserved information in the memory (Fuentemilla et al., 2010).

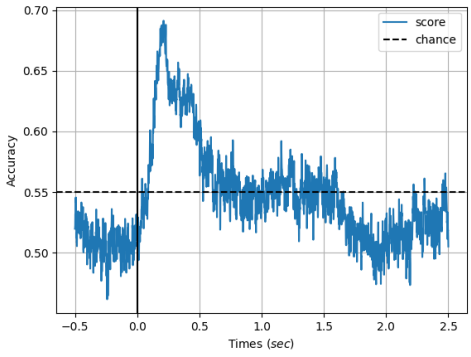


FIGURE 4.5: Grand average temporal decoding accuracy when discriminating between overt and covert speech task using MVPA. The peak accuracy was achieved at 215*ms*.

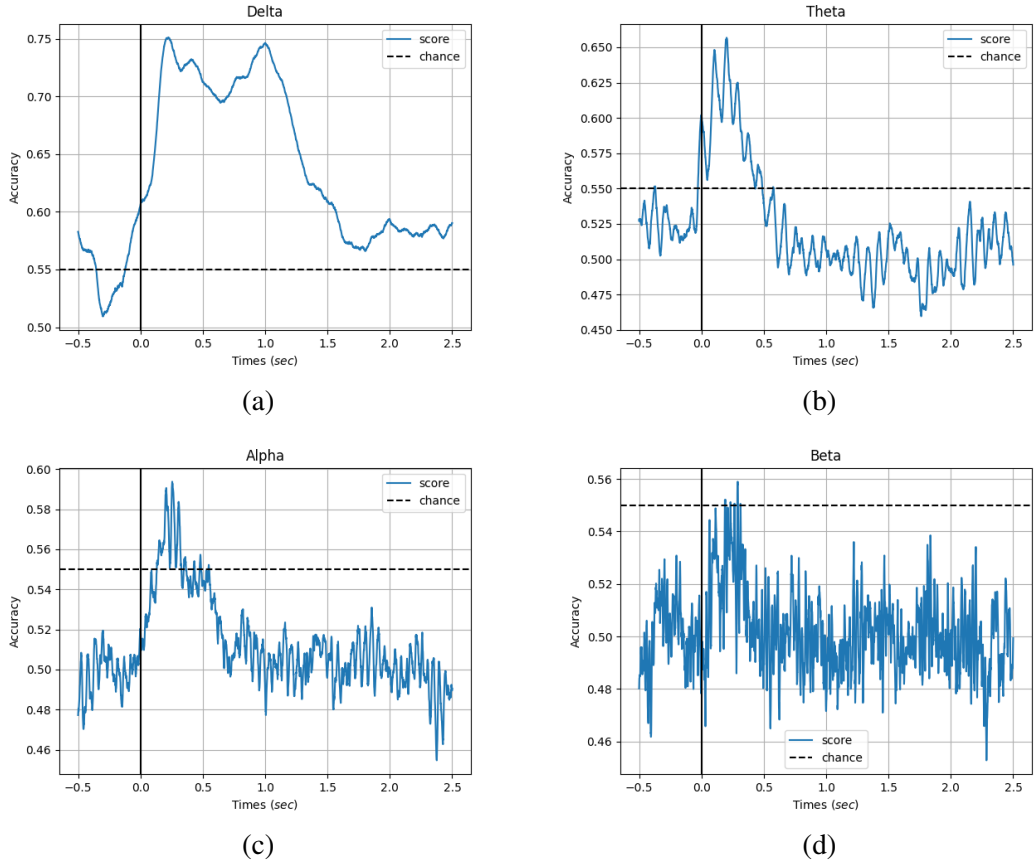


FIGURE 4.6: Temporal decoding between covert and overt speech for different frequency bands. (a) Delta (b) Theta (c) Alpha (d) Beta.

4.4.3 Decoding Temporal Points with Task Discriminative Information for Covert and Overt Speech

In the second analysis, temporal decoding was estimated to distinguish between covert and overt speech. A logistic regression (linear) classifier was trained and tested on each time point in the EEG signals separately. The data was split into 80% training and 20% testing. Input to the classifier was a vector of length 64, containing a time point from 64 electrodes. The decoding accuracy vector was measured for all participants separately, which were later averaged to obtain the global decoding accuracy vector. The decoding accuracy provided the temporal points when EEG signals produced during the overt and covert speech where most distinct. The temporal decoding vector is illustrated in 4.5. The grand average decoding accuracy ranged around chance level prior to the stimulus onset and above chance level after the stimulus presentation. At 0 *ms*, the decoding accuracy increased rapidly and peaked at 215 *ms* following a gradual drop.

Further, we evaluated the decoding accuracy for different frequency bands, where the delta band achieved highest recognition rate (approximate 75%). This could be due to the fact that the delta band's oscillations may reflect the processing of abstract language (Zhou et al., 2016a). Accuracy for delta band was always above chance level and had a longer peak latency. The decoding performance for different frequency bands are shown in figure 4.6. Decoding accuracy for beta band was below the chance level throughout.

4.4.4 Temporal Behavior of Neural Activity for Covert and Overt Speech Tasks

Along the decoding matrix, TG matrix was also calculated for covert and overt speech task. The TG matrix is shown in figure 4.7. As shown, some of the neural activity during the speech processing tasks was transient, i.e., the classifier performed well on neighboring time points, which is presented by the highest accuracy along the diagonal (orange color). On the other hand, the classifier did not generalize well for distant time points, lower accuracy away from the diagonal (green color). There was also evidence of off diagonal neural activity, illustrated by orange pattern in figure 4.7.

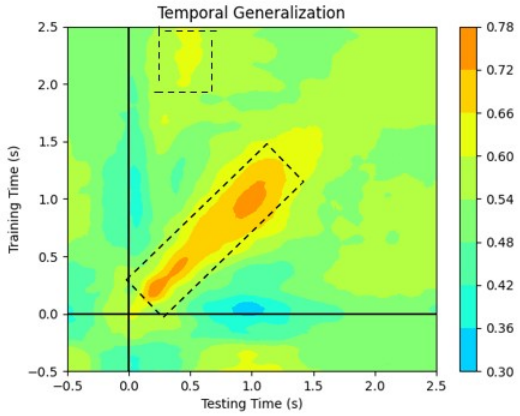


FIGURE 4.7: Temporal Generalization matrix for covert and overt speech. The TG matrix shows the highest accuracy along the diagonal and the sharp drop decoding accuracy away from the diagonal.

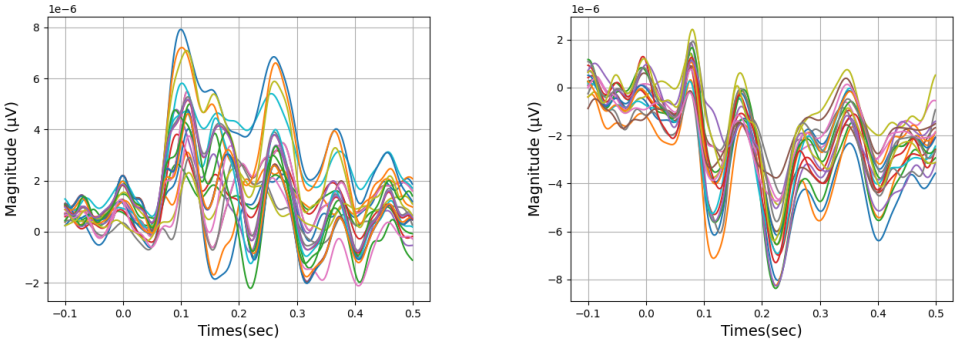
4.5 Spatio-Temporal Feature Learning using the Convolutional Neural Network (CNN)

This section proposes a system for identifying different cognitive tasks from EEG data, by learning spatio-temporal patterns. The system is composed of two major steps: the clustering of electrodes into subgroups using the K-means clustering algorithm, and classification using the CNN. The first step of clustering served as a dimensionality reduction technique and helped in finding components associated with different brain regions for the underlying brain activity. The second step involves feature extraction and classification using the CNN network. A single input to the network is defined as a matrix $D \in \mathbb{R}^{E \times T}$, where E is the number of electrodes and T is the time points. To capture short-term patterns in the EEG signals, input was divided into N windows, then the new input was $D \in \mathbb{R}^{N \times E \times T_w}$, where T_w is the number of time points in a given window. Each window was fed to a separate CNN and the extracted features were combined at later stage in the network. Further, the network contains a convolutional block attention module (CBAM) (Woo et al., 2018), which helped in highlighting the important features across the spatial (electrodes) dimension.

Further, from the epoch of $3sec$ a smaller time window of $600ms$ was used in this analysis, with time period -100 to $0ms$ before and 0 to $500ms$ after stimulus onset. This time period is shown to be significant in MVPA analysis figure 4.1. Then the electrodes were divided into three groups using K-means clustering algorithm, further, the data was standardized to zero mean and unit standard deviation. Lastly each EEG trial of length $600ms$ was further divided into N windows and fed to the CNN.

4.5.1 Electrode Selection using the K-Means Clustering Algorithm

Electrode selection is a challenging task in EEG studies. Selecting fewer EEG electrodes is more convenient for particle applications as it reduces resource consumption, minimizes the analysis time, and



(a) (b)

FIGURE 4.8: Patterns (clusters) in the EEG signal were observed in two electrode groups obtained by K-means clustering. The colored lines in the plot refers to pattern from each electrode separately (best viewed in color). The clusters were observed in two brain regions (a) Parieto-Occipital lobe, and (b) the Frontal lobe.

avoids overfitting caused by the usage of a high number of redundant electrodes. In addition, reducing the number of electrodes is an effective method for artifact reduction. In this chapter, the K-means clustering algorithm was used for electrode selection. The K-means algorithm was chosen because it was found that spatio-temporal patterns produced by neighboring electrodes in a given brain area were similar. Therefore, clusters of electrodes were used for evaluation of the proposed method.

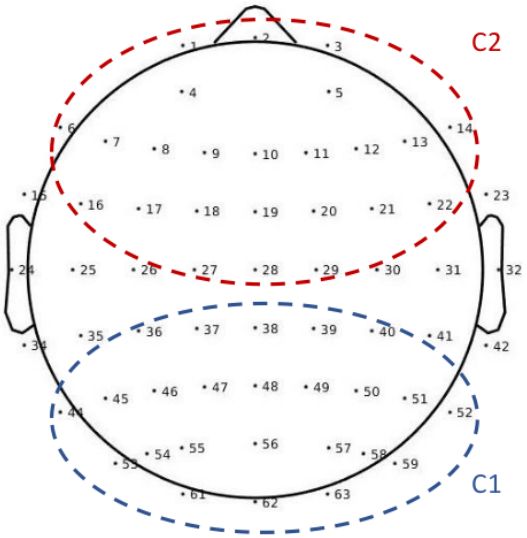


FIGURE 4.9: Approximation of electrodes in Cluster 1 (C1) and Cluster 2 (C2).

The input to the electrode selection method is training data $X \in R^{n \times E \times T}$, where n is the total number of training trials, E is the number of electrodes, and T is the time points in the signal ($600ms$). All the training trials of $600ms$ time window are averaged for each electrode separately, resulting in $X_{avg} \in \mathbb{R}^{E \times T}$. K-mean clustering is performed on the X_{avg} matrix, resulting in K electrodes group obtained from the set of E electrodes. In this analysis, several values of K were investigated however, best results were obtained for $K = 3$, i.e., the 64 electrodes were divided into three groups, where each group was associated an underling EEG signal pattern (component). EEG patterns are shown in figure 4.8. The three components or patterns were associated with three brain regions: (C1) Parieto-Occipital lobe, (C2) Frontal lobe, and (C3) Superior Temporal lobe (Broca's and Wernicke's area). However, for some subjects the cluster (C3) contained fewer electrodes. Subsequently, two clusters C1 & C2 and the brain regions associated with

them were further investigated during the classification stage. An approximation of electrodes in C1 and C2 is shown in figure 4.9.

4.5.2 Network Architecture

A single convolutional layer was used in each block. The convolutional layer in the first block filtered the data using 32 kernels with a receptive field of size 3×3 and stride of size 2×2, a process that can capture high-level spatio-temporal features from the EEG signals. The convolutional layers in the second and third block had 64 and 128 kernels of size 1×3 applied with a stride of size 1×2. The network performed spatio-temporal convolution in the first block, however, in the second and third block performed spatial convolution (across electrodes). To assist with regularization, the network employed dropout regularization with dropout rate of 20% and batch normalization approach (Ioffe & Szegedy, 2015; Srivastava et al., 2014). The network used *Relu* activation function (Nair & Hinton, 2010) in all the three blocks to learn non-linearities in the features.

The third block contained a convolutional block attention module (CBAM) (Woo et al., 2018). The CBAM helped the network to emphasize significant features along spatial (electrodes) information. To do this, the CBAM apply channel and spatial attention mechanism, allowing each branch to determine which channel and spatial region to assign more weights. CBAM leverage the inter-channel relationship to generate a channel attention map. This is done by aggregating the spatial information by performing the average-pooling and maxpooling across the spatial dimension, which results in average-pooled features F^c_{avg} and maxpooled features F^c_{max}. These features are forwarded to dense layers (MLP with one layer), then the output vectors are summed together (Woo et al., 2018). The channel attention is defined as:

$$M_c(F) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \tag{4.3}$$

where σ is the sigmoid function, W_0 and W_1 are the weights of MLP. The number of hidden units in the MLP are $\mathbb{R}^{\frac{C}{r}}$, where C is the number of channels and r is the reduction parameter. Subsequently, the weights of MLP becomes $W_0 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$. The second part of CBAM is the spatial attention, which highlights the most informative region in the feature map. In order to calculate spatial attention, maxpooling and average-pooling is applied across channel dimension and combine (concatenate) the pooled features to produce an effective feature (Woo et al., 2018). The combined feature map is fed to a convolutional layer which generate a spatial attention map providing more weights to the region of high importance. The average-pooled F^s_{avg} and maxpooled F^s_{max} features across channel dimension are given as:

$$\begin{aligned} F^s_{avg} &\in \mathbb{R}^{1 \times E \times T} \\ F^s_{max} &\in \mathbb{R}^{1 \times E \times T} \end{aligned} \tag{4.4}$$

where E and T refers to the dimensions of the 2D input. Therefore, the spatial attention map is defined as:

$$M_s(F) = \sigma(f^{1\times 3}([F^s_{avg}, F^s_{max}])) \tag{4.5}$$

where σ refers to the sigmoid function, $f^{1\times 3}$ refers to the filter size used in the convolutional layer. The arrangement of channel-attention and spatial-attention was similar to proposed in the work by (Woo et al., 2018). The third block contained 128 feature maps, therefore the CBAM helped the network to perform efficient feature refinement in this block. The CBAM module was followed by the classification block.

The final block in the network was the classifier block, which contained four dense layers. The number of nodes within the dense layers were 256, 128, and 64, where all the layers had *ReLU* activation and two in the last dense layer (classifier layer) with $sigmoid function$ for classification.

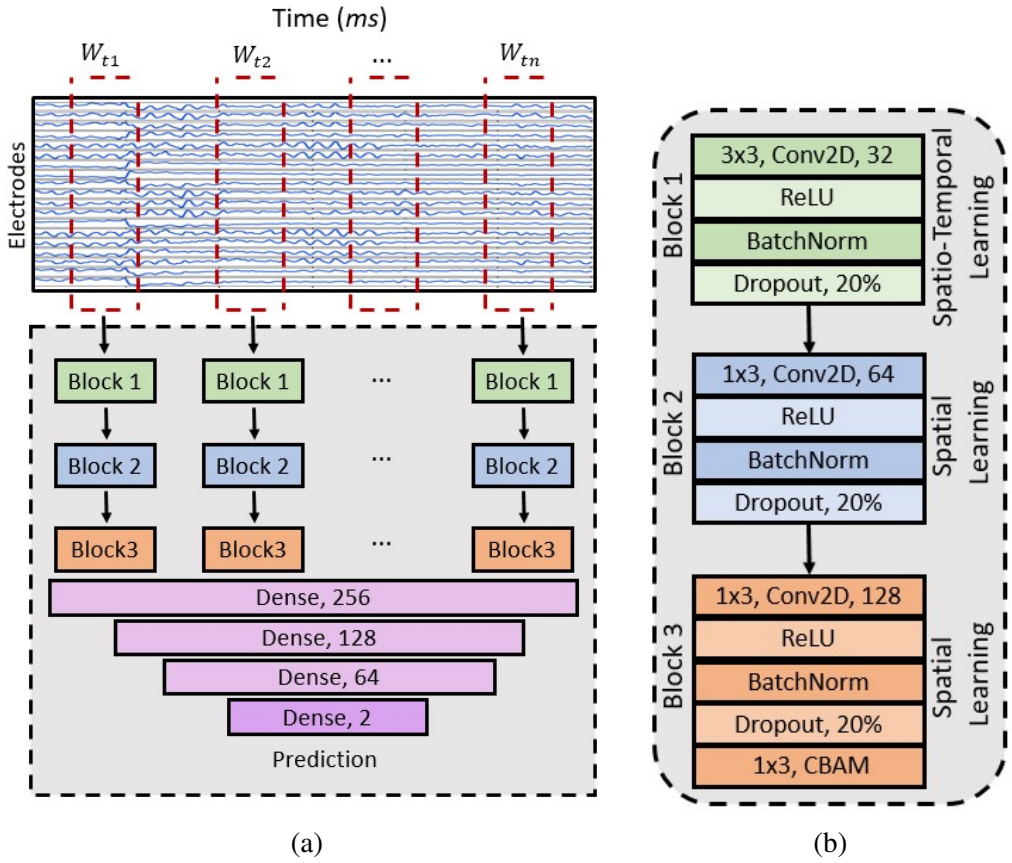


FIGURE 4.10: The architecture of convolutional-attention network. (a) The overall network architecture (b) three blocks used the network architecture. The spatial ordering of electrodes was done with respective to their position number, as shown in figure 4.9.

4.5.3 Design Choices

Different design choices for the network architecture, which would impact the performance of the network were investigated. Further, by varying the design choices (the activation functions, kernel size, and filter size) some insights were developed about the components in the network. Performance of the network was most affected by the filter size. For example, a filter of size 3×3 in the second and third block reduced the accuracy of the network, whereas the filter of size 1×3 led to better performance. This indicated that using convolutional operation to estimate features led to poor approximation of the temporal dynamics of the EEG signals. In addition, the *ReLU* activation gave the best results. Furthermore, the dense layers were implemented without *ReLU*, which reduced the recognition rate. This led to the conclusion that the *ReLU* function made the dense layers more robust in feature extraction towards the end of the network. The analysis tried to be thorough in the evaluation of while designing the network, this was done by evaluating different configurations to see which one achieves the highest performance on the test set. However, this in itself could lead to overfitting, therefore to avoid this only on the first test set in the leave-one-out cross validation method, this was done to avoid overfitting on the test data.

4.5.4 Network Training

The CNN network was trained at a learning rate of 0.0001 to minimize the cross-entropy loss. In order to avoid varying gradient at deeper layers, the network was trained at a slower learning rate (Bashivan et al., 2015). The "He" initialization (He et al., 2015) was used to initialize weights with purpose of avoiding unstable gradient. The network was trained for 120 iterations (epochs), with a mini-batch size of 32.

4.6 Results

Results were obtained from EEG signals for recognition of different cognitive processes in the brain. EEG signals were used from three cognitive tasks, *covert speech, overt speech, and visual imagery*. The aim of this analysis was to distinguish covert speech from overt speech and visual imagery. Each subject had 100 EEG trials for each class (cognitive task). However, because some of the recorded trials had to be eliminated due to excessive noise, some subjects ended up with only 90 trials. Three distinct electrode sets were used to assess the effectiveness of the suggested framework. Three sets of results were acquired in a subject-dependent manner, i.e., the network was trained and evaluated independently for each individual. This analysis was only conducted in a subject-dependent manner.

Three set of results were obtained using different electrode groups. The first set of results was obtained when all the 64 electrodes were used to train and test the network. The second and third sets of results were obtained when network was trained and tested using electrode groups C1 & C2, retrieved from the K-means clustering. The electrodes were arranged in sequential manner i.e., in order of the electrode number starting as shown in figure 4.9. In this experimental setup, the results were obtained using leave-one-out cross validation manner in which 90% of the data was used from training and 10% for testing. Results obtained from all training and testing set were averaged for each subject separately. The three experimental protocols are shown in Table 4.1 and described as follows:

- **All**: EEG signals from all the 64 electrodes were used for training and testing of the network.
- **C1**: EEG signals from electrodes in cluster 1 (parieto-occipital lobe) were used for training and testing the network.
- **C2**: EEG signals from electrodes in cluster 2 (frontal lobe) were used for training and testing the network.

TABLE 4.1: Three evaluation methods: first, when all the electrodes were used for training and testing the network, and second & third, when electrodes in clusters C1 & C2 are used for training and testing.

Exp	Electrodes	Brain Area	Selection Method
All	All	All	-
C1	Cluster 1	Parieto-Occipital Lobe	K-means clustering
C2	Cluster 2	Frontal Lobe	K-means clustering

4.6.1 Classification between EEG signals for covert speech and visual imagery task

TABLE 4.2: Classification accuracy for EEG signals recorded during covert speech and visual imagery. Results for three experiments are shown.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
All	79.8	79.0	91.4	89.0	83.3	85.0	66.7	75.8	89.3	80.2	77.3	94.0	**82.7**
C1	73.0	84.8	90.2	90.7	88.5	80.3	80.1	74.7	87.8	70.0	78.1	97.6	**82.9**
C2	69.3	56.3	91.6	77.6	61.0	80.1	54.3	74.8	79.1	79.4	71.6	62.3	**71.5**

In order to discrimination between the EEG signals produced during covert speech and visual imagery task, the signal was band pass filtered between $4-30Hz$. This range was chosen because the two cognitive tasks were most distinct in these frequencies in the MVPA analysis. The results are presented in Table 4.2, showing the achieved results when the electrodes in cluster 1 (C1) were used in the analysis. The electrodes in cluster 1 performed better when compared to the performance achieved by the network trained and tested using all (64) electrodes. This shows that the proposed method can find

the optimal electrode group, containing the most discriminative information between visual imagery and covert speech tasks.

4.6.2 Classification between EEG signals produced during Covert and Overt speech

TABLE 4.3: Classification accuracy for EEG signals in covert speech and overt speech. Result for three experiments are shown.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
All	78.0	72.0	82.5	83.2	69.0	66.6	75.6	77.2	80.8	75.3	63.1	71.3	**74.5**
C1	82.1	70.1	78.2	80.4	78.3	68.0	76.0	75.0	93.6	68.9	70.1	92.2	**77.7**
C2	65.1	69.1	82.3	71.1	61.2	66.2	57.3	80.0	73.6	69.5	57.3	72.5	**68.7**

In this experiment, three set of results were obtained using different electrode groups. The first set of results was obtained when 64 electrodes were used to train and test the network. The second and third sets were obtained, when network was trained and tested using electrode groups retrieved from the K-means clustering. In this experimental setup, the results were obtained using leave-one-out cross validation, in which 90% of the data was used for training and 10% for testing. Results obtained from all training and testing set were averaged for each subject separately. The results are shown in Table 4.3. Similar to previous experiment, electrode over the Parieto-Occipital lobe (C1) achieve best recognition rate. On the other hand, some subjects performed better when all the electrodes were used for training and testing the network. The average recognition rate over all subjects was lower in comparison to accuracy achieved when distinguishing between covert speech and visual imagery tasks.

4.6.3 Comparison with complete EEG trial length

TABLE 4.4: Classification accuracy, when the network was trained and tested on EEG signals of 3000 *ms*. The experimental results are evaluated only using electrodes in C1. The recognition was performed between EEG signals from CO: covert and overt speech, CV: covert speech and visual imagery.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
CO	90.0	71.7	89.2	80.9	72.7	64.6	90	66.1	88.3	82.1	62.7	79.5	**78.1**
CV	75.1	75.5	91.8	76.8	75.0	75.1	65.8	62.5	74.6	72.9	66.7	89.6	**75.1**

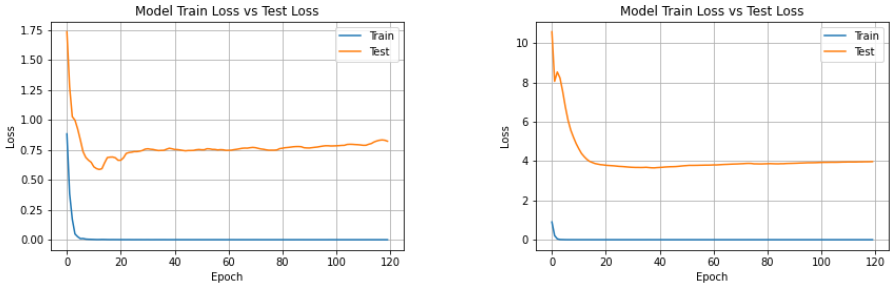
In the previous experiment, a 600*ms* time window was used from the EEG signal. To validate the effectiveness of the time window found by MVPA analysis, our framework was also evaluated on full length EEG signals of 3000*ms*. This experiment used electrodes in cluster 1 (electrodes over parieto-occipital lobe), which performed best in the previous analysis. Results are shown in Table 4.4. As can be seen, the recognition rate achieved by EEG trials of length 3000*ms* is lower, in comparison to analysis which uses 600*ms* time window for training and testing the network. This showed that the most task-discriminative events occur immediately after the stimulus onset, i.e., when the subjects were asked to perform the given task.

4.6.4 Limitation of the Methods used in this Chapter

The performance of the proposed network showed that it can distinguish between imagined speech and other cognitive tasks. However, the proposed network failed to achieve high recognition rate. In addition, evaluating frequency bands based on MVPA have its limitations, as MVPA considered contribution of the frequency band based on the spatial information. The temporal variation in the EEG signals at a particular frequency are disregarded because each time point is processed separately in the MVPA analysis. The performance of the proposed method was evaluated in subject-dependent manner, which

has limited applications in real life as the network will have to be trained directly on the EEG signals of the user. This would result in a BCI system unsuitable for multiple users. However, this approach offers privacy to the user of the BCI system.

Different network configurations were implemented to see which one achieves the highest performance on the test set. This in itself could lead to overfitting on the test set, although the test data was the first set in the leave-one-out cross validation. Therefore, inspecting the training loss curve can provide information potential over-fitting in the network. Therefore, learning curve for training and test sets were used to estimate overfitting. As can be seen from the curve in figure 4.11, the network suffer from overfitting when discriminating between covert speech and visual imagery EEG signals. It is to be noted that the difference in the magnitude of loss between training and test curve is due to number of trials in training and test data.



(a) (b)

FIGURE 4.11: Training (blue) and testing (orange) loss for the convolutional-attention network. (a) covert speech and visual imagery; (b) covert and overt speech.

4.7 Conclusion

The work in this chapter investigated the difference between covert speech and two other cognitive tasks in the brain. The two cognitive tasks were: visual imagery and covert speech, which in the past have been associated with covert speech (Dentico et al., 2014; Schendan & Ganis, 2012). Two analysis were conducted to investigate the difference in EEG signals of cognitive tasks: (1) MVPA and (2) TG.

The analyses showed that the 500ms time window after the stimulus onset to contain most discriminative information, when distinguishing covert speech with two cognitive tasks (visual imagery and overt speech). However, MVPA analysis of EEG signals for each frequency band separately showed the theta band (4-7Hz) to be carrying the most discriminative information when distinguishing between covert speech and visual imagery. Further, the TG matrix for the two cognitive tasks showed oscillatory behavior, which are known to be activated when recalling information from the memory (Fuentemilla et al., 2010). On the other hand, EEG signals for overt and covert speech were most discriminative at the delta band (0.5-3Hz). The TG matrix revealed transient processing of neural activity at the delta band.

Further, classification was performed between two cognitive tasks and covert speech. Recognition between each cognitive task and covert speech was performed separately. Further, the K-means clustering algorithm was used for electrode selection, which provided best results from electrodes covering Parieto-Occipital lobe. In addition, a comparison was made between the performance achieved by EEG signals of 600ms time window and a complete EEG trial (3000ms). The 600ms time window achieved higher results compared to EEG signals of 3000 ms. However, contrary to MVPA best results using CNN-attention network were achieved when a larger frequency band was selected, 4-30Hz for recognition between covert speech & visual imagery and 0.5-30Hz for recognition between covert and overt speech.

4.8 Summary

Covert (imagined) speech is often associated with visual imagery, i.e., thoughts with visual imagination. Similarly, comparison between covert and overt (loud) speech can help understand the underlying neural mechanism of covert speech processing. Therefore, this chapter investigated the difference between temporal dynamics in EEG signals produced by covert speech & visual imagery task and covert & overt speech task. This work examined differences in the neural activity using two different methods. First method was temporal decoding, where a linear classifier was trained and tested on each time point separately, to investigate the temporal window when two cognitive tasks in the brain were linearly separable. The second method was TG, where the classifier was trained on time point t_x in the EEG signal and tested on a different time point t_y. This was done for all the time points in the EEG signal, swapping training and testing time points. This provides a two-dimensional decoding matrix which provided information about the stages of neural processing for a given activities. The results of the analysis showed that 500 *ms* time period after stimulus onset produced the best results, when comparing two activities with covert speech. Covert speech and visual imagery where most separable in the theta band, which showed oscillatory behavior associated with top-down processes (i.e., that is generating from memory). On the other hand, the highest recognition between covert and overt speech was achieved in the delta band showing transient neural activity in the TG matrix. Further, we proposed a framework using the K-means clustering algorithm and the convolutional-attention network for recognition of two cognitive tasks (visual imagery and overt speech) from covert speech. Also results acquired in this analysis validated the discriminatory information presented by the 600 *ms* window found in the first two analysis.

Chapter 5

Imagined Speech Recognition using Dynamic Time Warping

In the first part of this chapter, we extract time-frequency and time domain features using traditional feature extraction method, such as the linear discriminant analysis (LDA) and common spatial patterns (CSP) for recognition of mentally spoken words from EEG signals. The method uses classifiers such as the K-NN and SVM for recognition task. The second half of the chapter introduces a method for measuring similarity between imagined words using the dynamic time warping (DTW) and an electrode fusion technique.

5.1 Introduction

As mentioned in chapter 4, BCI based on imagined speech can sever as mode of communication for locked-in patients. For development of a speech based BCI there is need to recognize imagined speech from EEG signals. Therefore, this chapter focuses on classification of imagined words using EEG signals under binary condition. The chapter aims to develop an optimal classification technique and also extract most informative features in both time-frequency and time domain. In order to achieve this, statistical features were extracted from spectrograms along with eigen values using LDA. In the past, studies have used LDA for feature extraction and dimensionality reduction of EEG signals (Chen et al., 2019; Kołodziej & Majkowski, 2012). The work in (Kołodziej & Majkowski, 2012) aimed at classifying EEG signals for imagined hand movement. To achieve this the author used discrete Fourier transform for feature extraction and LDA for dimensionality reduction of the extracted features. The method was evaluated using K-NN classifier with K=10 under cross-validation scheme achieving above chance level recognition rate with only eight electrodes. The work in (Chen et al., 2019) proposed an algorithm for features extraction from EEG signals for emotion recognition. Their proposed method used fusion of features from LDA and differential entropy. The authors tested the method on three emotion-class datasets using five classifiers: K-NN, logistic regression (LR), (SVM), random forests (RF), and multi-layer perceptron (MLP). The method achieved an average accuracy of 68%, higher than the previous method. The focus of the work in (Wester, 2006) was to recognize unspoken speech from EEG signals. The author extracted short time Fourier transform (STFT) and delta-delta features from the raw EEG signal and LDA was used to reduce the dimensionality of these features. In EEG studies, LDA is also used as a classifier. The study (Chi et al., 2011), performed classification between overtly spoken phonemes using Naive Bayes and LDA classifiers. The performance of the LDA classifiers was reported to be superior. The work in (Alsaleh, 2019), focused towards recognition of imagined speech using EEG signals. The author extracted time domain and spatio-spectral features such as mean, standard deviation, and components using common spatial pattern. These features were classified using LDA classifier along with SVM and random forest.

Further, spatial features are also extracted using CSP, which is known to perform well with EEG data. Results from two different methods have been proposed for recognition of imagined speech of three action words under binary classification task. To extract important time-frequency components we used

LDA and spatio-temporal features were extracted using CSP. Further, limitations of the LDA and CSP are discussed, along with certain conclusions drawn from the analysis. In the Second part of this chapter, an optimal similarity matching technique based on DTW was implemented for recognition of imagined words. This was done to solve the limitation caused by temporal variation in EEG signals. Further, to make the similarity matching method more effective, an electrode fusion method was introduced. The contributions of this chapters are as follows:

- Analysis with the LDA and CSP for feature extraction and classification, which were not effective in recognition of imagined words.
- A similarity matching technique for recognition of imagined words (verbs) from EEG signals using DTW. The results with DTW also show the effectiveness of non-linear methods in dealing with the dynamic behavior of EEG signals.
- An electrode fusion method used after DTW to combine distances from multiple electrodes, which leads to improved recognition of imagined words.

5.2 EEG Dataset

As the focus of this chapter is to recognize imagined speech, therefore, we used EEG data acquired from covert speech task. EEG signals from 12 subjects were used to perform the analysis, contaminated trials or noisy data was rejected. After artifact rejection and filtering, some subjects were left with less than 64 electrodes, therefore, the EEG signals for those subjects wer not included in order to avoid variation in parameter when calculating spatial patterns using the CSP algorithm. The EEG signals used in the experiment were recorded for mentally spoken "*action words*" which represented action or activity. The words were: "*Run*", "*Swim*", and "*Write*". EEG signals for these words were randomly selected from the list of words in chapter 3. Each word had ten trials in total per subject.

5.3 Linear Discriminant Analysis (LDA)

The LDA is a classification and dimensionality reduction technique, which offers an advantage over some other dimensionality reduction techniques. LDA increases the ratio of within class to between class variance (Balakrishnama & Ganapathiraju, 1998) also known as Fischer ratio (5.2). LDA was chosen over other methods such as the principal component analysis (PCA), this is because PCA changes the feature vectors spatially, whereas the LDA does not changes the spatial information of the features, rather it creates more class separability between the features (Balakrishnama & Ganapathiraju, 1998). Another motivation factor for using LDA was that the direction of separation in terms of frequencies projected by the eigen vectors could potentially be useful in classifying the underlying concept.

It is a supervised dimensionality reduction technique which project the given data to a linear subspace which maximize the separation between classes by using the projection direction. Data separation can be measured as:

$$F = \frac{tr(S_m)}{tr(S_w)} \tag{5.1}$$

where F is the separation coefficient, S_m is the between class scatter matrices, and S_w is within class scatter matrices. For C number of classes with each containing N samples x_i, the with-in class scatter matrices for one class c can be estimated as:

$$S^c = \sum_{i=1}^{N} (x_i^c - m^c)(x_i^c - m^c)^T \tag{5.2}$$

where m^c is the sample mean in x_i for c^{th} class. Therefore, with-in class scatter for all the classes can be measured as:

$$S_w = \sum_{i=1}^{C} \frac{n_i}{N} S^i \tag{5.3}$$

where n_i is the number of x_i samples in each class and N is the sample size. Similarly, between class scatter matrices for a given class c can be calculated as:

$$S_B^c = \sum_{i=1}^{C} (m_i^c - m^c)(m_i^c - m^c)^T \tag{5.4}$$

where m_i is the sample mean of all the samples x_i for i^{th} class and m is the overall mean across all the samples x_i and classes. Hence, the between class scatter metrics can be calculated as:

$$S_m = \sum_{i=1}^{C} \frac{n_i}{N} S_B^i \tag{5.5}$$

where n_i is the sample mean of x_i each class and N is the total number of samples. The projection direction providing optimum class separation are given by the eigen vectors with the highest eigen values of scatter matrix S (Kołodziej & Majkowski, 2012):

$$S = S_w^{-1} S_m \tag{5.6}$$

Eigen vectors from the input data are collected into the S scatter matrix. The scatter matrix S is not symmetric, this is resolved by using generalized eigenvalue problem (Kołodziej & Majkowski, 2012). The new transformed dimensional samples in the new subspace are obtained as:

$$Y = x^T W \tag{5.7}$$

where W is the weight matrix containing eigen d eigen vectors $(e_1, e_2, ..., e_d)$ Each eigen vector have its own eigen value which provide information about its length and magnitude. Eigen vectors are sorted in accordance with the eigen values i.e., the vector with the highest value is the first and most information containing component. Y is the transformed dimensional samples in the new subspace (the projecting the data in the lower dimension).

5.4 Feature Extraction

In this experiment spectro-temporal features and spatial features were investigated. Statistical features such as mean, median, standard deviation, variance, and entropy were obtained from windows of spectrogram. Further, eigen values representing the most important time-frequency information was evaluated using the LDA. In addition, energy for each time window was calculated. Spatial features were computed using CSP (Koles et al., 1990). The description of feature extraction is provided in the following subsections.

5.4.1 Spectro-Temporal Features

As mentioned in chapter 3, raw EEG signals were converted into spectrograms. Therefore, spectro-temporal input was used at the feature extraction stage. In this analysis, the spectrograms were divided into four overlapping windows: $t_{w1}, t_{w2}, t_{w3}, t_{w4}$. This was done to capture time-frequency information in a smaller time frame, because extracting single feature value for whole spectrogram might not be able to efficiently capture spectro-temporal dynamics of underlying brain activity.

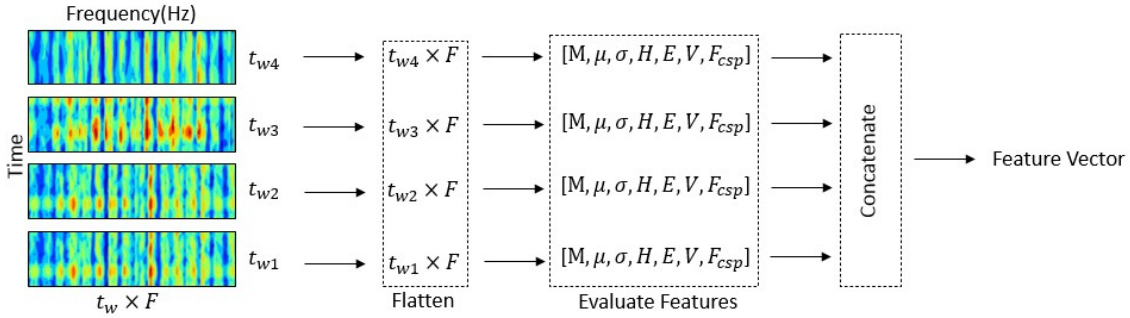


FIGURE 5.1: Spectrogram separated into four overlapping window. Each window was of dimension $T \times F$ where $T = 26$ and $F = 86$. Features were extracted from each window separately.

Each window was converted from two dimension of $(T \times F)$ time and frequency to one dimensional feature vector X of length N. Seven features were calculated from each window for each electrode, the features were:

- Energy:

$$E = \sum_{i=1}^{N} |X_i|^2 \tag{5.8}$$

- Mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (X_i) \tag{5.9}$$

- Standard Deviation (STD):

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} \tag{5.10}$$

- Entropy :

$$H = -\sum_{i=1}^{n} p(X_i) \log_2 p(X_i) \tag{5.11}$$

  Where H is entropy. p is the probability of observing i^{th} value of the bin data t_{wn}

- Median:

$$M = \begin{cases} X_{n\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}\left(X_{n\frac{n}{2}} + X_{\frac{n}{2}+1}\right) & n \text{ even} \end{cases} \tag{5.12}$$

- Variance:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \tag{5.13}$$

- Root mean square (RMS):

$$X_{RMS} = \sqrt{\frac{1}{N}\sum_i X_i^2} \tag{5.14}$$

- LDA components were calculated using (5.7). In case of LDA, the number of components for new feature set is one-less the number of classes $C-1$. The eigen vector with highest eigen value bears the most information.

Figure 5.1 shows the feature extraction method. First, the spectrogram was divided into four overlapping windows of equal length with an overlap of six time points. These windows were transformed into vectors from which features were calculated and at the later stage feature vectors from all the windows were combined into a global feature vector.

5.4.2 Spatio-Temporal Features

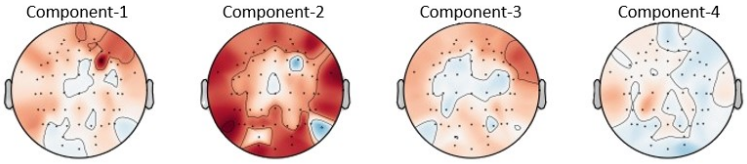


FIGURE 5.2: CSP components extracted from the EEG signals.

Along with the spectro-temporal features, the spatio-temporal features were also calculated using CSP algorithm (Koles et al., 1990). The CSP features were extracted in time domain and have been very successful in recognition of motor imagery tasks at lower frequencies (<50Hz) (Blankertz et al., 2007). Therefore, CSP algorithm was applied to bandpass filtered data between $4-45Hz$. The idea of using spatial filters in the CSP algorithm is to evaluate spatio-temporal features that can optimally discriminate different classes. Spatial filters can be used to measure the variations in the potential of EEG signals during mental activity. Let the data $X \in \mathrm{R}^{E\times T}$ with E electrodes and T time points. At a single time point in the signal (data) is defined as $x(t)$, therefore X is defined as:

$$X = [x(t), x(t+1), ..., x(t+T-1)] \tag{5.15}$$

The decomposition of the signal along the sensor dimension to CSP component matrix is evaluated using the following transformation:

$$x_{csp}(t) = W^T x(t) \tag{5.16}$$

where W is known as de-mixing or the projection matrix of dimensions $E \times E$ and rows of x_{csp} contains CSP components. The de-mixing or projection matrix is computed by simultaneous diagonalization of the covariance matrices from the two conditions (classes) (Ramoser et al., 2000). The features were calculated using (5.14), but only a defined number (n) of spatial filters were used for feature extraction, in this study $n = 4$. For signal $x_{csp}(t)(t = 1, ..., n)$, the feature vector was calculated as:

$$F = \log\left(\frac{var(x_{csp}(t))}{\sum_{t=1}^{n} var(x_{csp}(t))}\right) \tag{5.17}$$

5.4.3 Classification

Three classifiers were used for the classification of imagined speech EEG signal. All the three classifiers were trained and tested on the extracted spatio-temporal and spectro-temporal features. The classifiers used were SVM, K-NN, and decision tree (DT).

Support vectors are based on the predictors that are closest to the decision boundary creating separation among different classes (Subasi & Gursoy, 2010). The decision boundary does not only separate classes, but also have clear separation from closest predictor value, which have maximum impact on the decision made by the classifier. The SVM can be optimized or designed to suit input features, SVM can perform hard margin and soft-margin classification. The hard margin classification works well with linearly separable data. This is not suitable for classification of EEG dataset, which is non-linear and non-stationary in nature. However, if soft-margin classification is used, then hyper-plane can be much more flexible, it can be adjusted accordingly. To achieve better performance, the SVM classifier was used with polynomial regression. 3^{th} order polynomial regression was chosen as it achieved results.

A *K*-NN classifier is non-parametric, which means that the parameters are evaluated based on training data and that no prior assumption about the data distribution is made to create the model (Abu Alfeilat et al., 2019). The prediction is made based on a vote of similar samples in the nearest neighbor closest to input sample. *K*-NN with euclidean distance has been ineffective in several BCI experiments because of its sensitivity to high-dimensionality of the data (Siuly et al., 2016), however it performs efficiently with low dimensional features. In this work correlation distance was used as a distance measure in the *K*-NN algorithm.

A Decision Tree is a non-parametric algorithm, used for classification and regression. It divides the dataset into smaller subsets using binary partitioning while making several simpler decisions. The structure of a DT contains nodes, branches, and leaves nodes. Decision Trees are good for mapping non-linear trends in the data which can be effective in analyzing EEG signals (Aydemir & Kayikcioglu, 2014; Guan et al., 2019).

5.5 Results

The proposed features were extracted only on EEG signals acquired during the covert speech task, the dataset contained *action words*: "*Run*", "*Swim*", and "*Write*". Binary classification was performed for three word pairs, "*Run and Swim*", "*Run and Write*", and "*Swim and Write*". Classification was performed in subject-dependent manner i.e., recognition of imagined speech from EEG signals for each subject was done separately. For experimental, evaluation of two groups of features (spectro-temporal and spatio-temporal) was done in two different manners. In the first method, the features were used for training and testing the classifier separately and the results were evaluated. In the second method, both the features were combined for training and testing the classifiers. Most subjects had 10 trials for each word, however some subjects had less trials after artifact reduction. Therefore, cross-validation was performed for eight times. The EEG dataset was divided into 90% training and 10% testing data and classification was performed using leave-one-out (LOO) cross validation method.

5.5.1 Classification using Spectro-Temporal Features

The spectrograms were sliced into four spectro-temporal windows and from each window seven spectro-temporal features were extracted, as described in section 5.4.1. Features were extracted from each electrode separately and later the feature vectors from each electrode were concatenated together. These features were split into 90% training and 10% testing data, which were normalized to be on the same scale with zero mean and unit standard deviation. Mean and standard deviation of testing features were normalized to training features.

Three different sets of results were obtained for 12 subjects using three classifiers, SVM, K-NN, and Decision Tree. The results are shown in Table 5.1. Average accuracy achieved using each classifier for three binary pairs were 53.7% (SVM), 53.4% (*K*-NN), and 52% (DT). The average accuracy obtained by all the three classifiers was above chance level, however all classifiers failed to achieve high recognition rate. Further, the results between subjects varied, for example *K*-NN for word pair "Run, Swim" achieves high accuracy for subject 7, whereas it failed to achieve chance level accuracy for subject 2. On the other hand, DT classifier performs better on subject 2 than most of the the other subjects. Further, variation

TABLE 5.1: Classification accuracy for EEG signals of three word pairs using the spectro-temporal features.

Subject	(Run, Swim)			(Run, Write)			(Swim, Write)		
	K-NN	SVM	DT	*K*-NN	SVM	DT	*K*-NN	SVM	DT
1	62.5	56.2	68.7	56.2	56.2	68.7	43.7	62.0	43.7
2	37.5	56.2	75.0	62.5	25.0	56.2	50.0	50.0	37.5
3	62.5	56.2	43.7	68.7	37.5	68.7	56.2	56.2	75.0
4	62.5	50.0	56.2	43.7	43.7	43.7	50.0	37.5	43.7
5	62.5	56.2	37.5	43.7	50.0	62.5	56.2	37.5	62.5
6	56.2	50.0	43.7	37.5	43.7	56.2	56.2	68.7	50.0
7	87.5	50.0	56.2	43.7	50.0	50.0	50.0	93.7	62.5
8	50.0	37.5	50.0	43.7	62.5	43.7	43.7	68.7	50.0
9	62.5	50.0	37.5	56.2	50.0	50.0	43.7	43.7	43.7
10	37.5	56.2	56.2	62.5	56.2	68.7	50.0	56.2	37.5
11	56.2	43.7	43.7	56.2	56.2	50.0	43.7	56.2	37.5
12	56.2	68.6	43.7	50.0	50.0	62.4	62.5	93.7	37.5
Average	**57.8**	**52.6**	**51.0**	**52.0**	**48.4**	**56.7**	**50.4**	**60.3**	**48.4**

in accuracy is also observed when comparing different word pairs. The SVM generalized well for word pair "Swim, Write".

5.5.2 Classification using Spatio-Temporal features

The spatial feature extracted using the CSP algorithm were evaluated in similar manner as the spectro-temporal features. Three classifiers were trained and tested using LOO cross-validation. The accuracy achieved by the CSP features was tested by varying the number of components, best results were achieved by using four spatial components. The training and test feature were normalized to same scale with zero mean and unit standard deviation. The average accuracy for three binary pairs with spatial features is lower than spectro-temporal features. The results are shown in Table 5.2. The average accuracy for the three classifiers were; *K*-NN: 49.2%, SVM: 53.5%, DT: 51.2%.

TABLE 5.2: Classification accuracy using the spatio-temporal features extracted using CSP.

Subject	(Run, Swim)			(Run, Write)			(Swim, Write)		
	K-NN	SVM	DT	*K*-NN	SVM	DT	*K*-NN	SVM	DT
1	56.3	68.7	50.0	25.0	56.3	43.7	56.2	43.7	56.3
2	75.0	50.0	50.0	56.3	50.0	43.7	43.7	31.3	37.5
3	62.5	43.7	25.0	50.0	56.3	56.2	62.5	50.0	50.0
4	62.5	62.5	75.0	62.5	68.7	56.2	75.0	62.5	56.3
5	43.7	43.7	56.2	43.7	56.3	50.0	18.7	62.5	56.3
6	25.0	50.0	50.0	59.0	62.5	68.7	62.5	56.3	50.0
7	37.5	56.3	37.5	25.0	50.0	43.7	43.7	31.5	25.0
8	56.3	43.7	50.0	37.5	62.5	56.3	50.0	56.2	56.3
9	68.7	75.0	50.0	50.0	62.5	56.3	50.0	31.5	25.0
10	43.7	56.3	56.2	43.7	56.3	56.3	62.5	81.3	81.2
11	62.5	43.7	50.0	50.0	56.3	62.5	43.7	43.7	50.0
12	43.7	43.7	56.2	31.3	75.0	62.5	62.5	43.7	56.3
Average	**52.4**	**51.7**	**50.5**	**42.8**	**59.4**	**54.6**	**52.5**	**49.5**	**48.5**

5.5.3 Combining Spectro-Temporal and Spatio-Temporal Features

Further, the performance of combined features was investigated from spectro-temporal and spatio-temporal domain. We trained and validated three classifiers using LOO cross-validation. Four CSP components were used in the spatial features. Both the training and test features were normalised to the same scale, with a mean of zero and a standard deviation of one unit. The results are shown in Table 5.3. The average accuracy for the three classifiers were; *K*-NN: 51.8%, SVM: 52.2%, DT: 52.2%.

TABLE 5.3: Classification accuracy achieved by combining the spatio-temporal and spectro-temporal features.

Subject	(Run, Swim)			(Run, Write)			(Swim, Write)		
	K-NN	SVM	DT	*K*-NN	SVM	DT	*K*-NN	SVM	DT
1	50.0	68.7	50.0	50.0	43.7	43.7	56.2	37.5	56.3
2	56.2	50.0	50.0	62.5	56.2	43.7	43.7	50.0	37.5
3	87.5	31.2	25.0	62.5	37.5	56.2	62.5	43.7	50.0
4	50.0	62.5	75.0	43.7	68.7	56.2	75.0	62.5	56.3
5	31.2	56.2	56.2	50.0	62.5	50.0	18.7	50.0	56.3
6	56.2	56.2	50.0	43.7	43.7	68.7	62.5	31.2	50.0
7	50.0	50.0	37.5	43.7	56.2	43.7	43.7	50.0	25.0
8	43.0	62.5	50.0	62.5	50.0	56.3	50.0	25.0	56.3
9	68.7	50.0	50.0	37.5	50.0	56.3	50.0	81.2	25.0
10	25.0	62.5	56.2	37.5	56.2	56.3	62.5	50.0	81.2
11	68.7	43.7	50.0	56.5	68.7	62.5	43.7	37.5	50.0
12	43.7	56.2	56.2	56.2	56.2	62.5	62.5	62.5	56.3
Average	**52.4**	**54.1**	**50.5**	**50.5**	**54.1**	**54.6**	**52.5**	**48.4**	**48.5**

5.5.4 Limitations of the Features

As can be seen from the results, the statistical, eigen, and spatial features are not effective enough to capture distinctive information from imagined speech. A reason for LDA and statistical features not achieving below chance level (55%) accuracy could be due to the fact that task specific events between spectrograms cannot be calculated directly. As events (frequency columns of the spectrogram) caused by same concepts may vary temporally in the compared spectrogram i.e., each trial. Further, spatial features captured using CSP algorithm achieved even lower accuracy which suggests that event during imagined speech in temporal domain are not consistent and alone cannot help distinguish imagined speech from the EEG signals. Further, some of the statistical features such as mean, median, and standard deviation contained correlated information which limits the representation of the data, subsequently reducing the performance of the overall system.

5.6 Similarity Matching with Dynamic Time Warping (DTW)

An important conclusion can be drawn from the analysis with spectro-temporal features is that the similarities between spectrograms cannot be evaluated directly due to inter-trial variations, which lead to the events varying temporally in the compared spectrograms. In order to solve this problem, the events in the given spectrograms need to be aligned, in other words temporal alignment of EEG features could be important for improving the discrimination between different mentally spoken words (classes). DTW is a non-linear alignment technique (Sakoe & Chiba, 1978), based on dynamic programming, famously used in speech recognition to reduce the effects of time delay and distortion between two time series. Therefore, DTW can be used to compensate for the differences in time of occurrence of these events and can serve as a measure of dissimilarity between two trials.

DTW has been used successfully in applications such as speech and gait recognition (Ismail et al., 2020; Rong et al., 2007). Also, DTW has been used for processing of brain signals. Previous studies have used DTW to align the high gamma ECOG activity with audio output and the aligned signals were used to create a template for each sentence (Zhang et al., 2012). For evaluation, correlation between test and the template was calculated followed by classification using SVM. Further, DTW has been proposed for estimating ERP, which help overcome the limitations faced by standard methods (Zoumpoulaki et al., 2015). The author tested their method for single participant and at group level, which provided results that outperformed the existing methods. In the past, improved recognition rate has been achieved by aligning the brain signals (ECOG) produced by imagined and overt speech using DTW (Martin et al., 2014). This is because time warping of two different signal synchronizes associated events. Similarly,

the research in (Martin et al., 2016) used DTW in SVM kernel function to form a non-linear alignment technique. ECOG signals from different trials were aligned using DTW.

5.6.1 Dynamic Time Warping

DTW is an non-linear technique used for measuring similarity between two time series or vectors and it calculates the best alignment path among them. The method uses non-linear warping function to estimate the time differences between two vectors or time series. Let the two vectors be:

$$\begin{aligned} A &= a_1, a_2, \cdots, a_s, \cdots, a_n \\ B &= b_1, b_2, \cdots, b_s, \cdots, b_m \end{aligned} \tag{5.18}$$

The distance between a_s and b_s is estimated using a distance measure (e.g., correlation distance), and a distance metric $D(n,m)$ is calculated which stores the distance between point a_s and b_s in the two vectors (time series). Next, a set of elements are mapped through the distance metric D which defines minimum distance between two vectors A and B by finding the warping path given as:

$$P = p_1, p_2, p_3, \cdots, p_s, \cdots, p_k$$

where $p_{s=(a_s,b_s)}$ and P is also known as the warping function. The distance between A and B is given by:

$$DTW(A,B) = min_P[\frac{\sum_{s=1}^{k} d(p_s) \cdot w_s}{\sum_{s=1}^{k} w_s}] \tag{5.19}$$

where $d(p_s)$ is the distance between a_s and b_s, and w_s is weighting coefficient. The warping path needs to satisfy to following constraints:

1. Boundary conditions: The warping path should start and end at diagonal opposite corner cells of the distance metric. $p_1 = (1,1)$ and $p_k = (n,m)$.
2. Continuity: This restricts the warping path to have one step at a time. Given $p_s = (x,y)$ then $p_{s-1} = (x',y')$, where $x - x' \leq 1$ and $y - y' \leq 1$.
3. Monotonicity: This constraints the points in warping path to go back in time index. $p_k = (x,y)$ then $w_{k-1} = (x',y')$ where $x - x' \geq 0$ and $y - y' \geq 0$.
4. Slope Constraint: This constraint propose that the slop of the warping path should neither be too gentle nor too restricted in order to avoid unrealistic alignment.

The path is find using dynamic programming. The distance between two points $a_s \in A$ and $b_s \in B$ can be estimated as:

$$D_c(a_s,b_s) = minD_c(as-1,b_s-1), D_c(as-1,b_s), D_c(as,b_s-1) + d(a_s,b_s) \tag{5.20}$$

where D_c is the cumulative distance and $D_c(a_1,b_1) = d(a_1,b_1)$. The final distance between two vectors or time series is equal to the distance at the end of the optimal path:

$$DTW(A,B) = D_c(a_n,b_m) \tag{5.21}$$

5.7 Similarity Matching of Spectro-Temporal Features using DTW

DTW is performed on the sequence of frequency vectors (columns of spectrogram) in the spectrogram , $S_x : x_1, x_2, ..., x_t, ...X_T$. Spectrograms are compared based on the dissimilarities between the frequency vectors, these dissimilarities are measured using a distance metric. To calculate the distance metric,

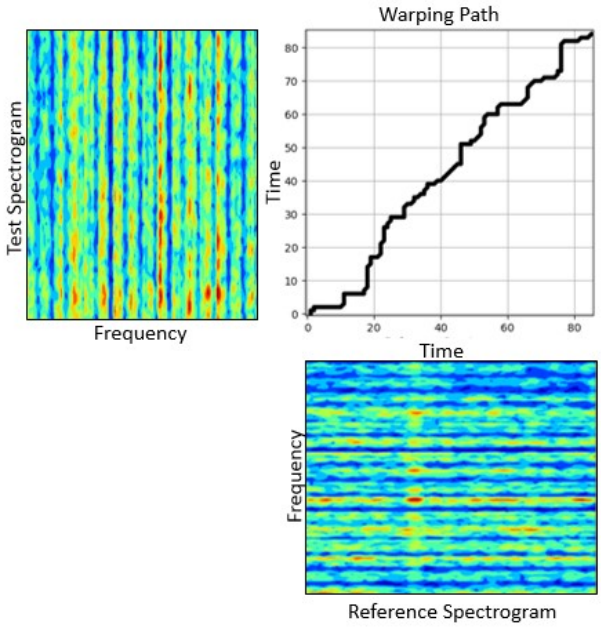


FIGURE 5.3: Warping path between reference spectrogram for mentally spoken word "write" with test spectrogram for mentally spoken word "write" using correlation distance.

correlation and cosine distance was used instead of using Euclidean distance, this is because spectral power in the spectrograms is expressed in decibels and Euclidean distance based on decibels values did not provide any useful conclusions. Correlation distance (Borgatti, 2019) between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$corr(\mathbf{x},\mathbf{y}) = \frac{\frac{1}{n}\sum_i x_i y_i - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{5.22}$$

where x and y are the reference and test spectrograms, μ_x, μ_y and σ_x, σ_y are the mean and standard deviation of x and y respectively. Correlation distance does not assume linear relationship between two variables. Similarly, the cosine distance is defined as:

$$cosine(\mathbf{x},\mathbf{y}) = 1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}} \tag{5.23}$$

DTW used to calculate the distance between the spectrogram of a test trial and all reference spectrograms (of all the classes). Equation (5.19) shows the time axis of the reference and test spectrograms, mapped using a warping function, a trellis matrix W consisting of distances δ between two spectrograms calculated using correlation were used to find the warping function. The path (shown in figure-5.3) that yields the minimum distance is the warping function F. Therefore, the minimum cumulative distance is:

$$I = min_m[D_m] \tag{5.24}$$

5.7.1 Electrode Fusion

The distance between trials was calculated using DTW on an electrode-by-electrode basis. This distance was calculated for all pairs of test and reference trials (spectrograms). Since we have M classes in total

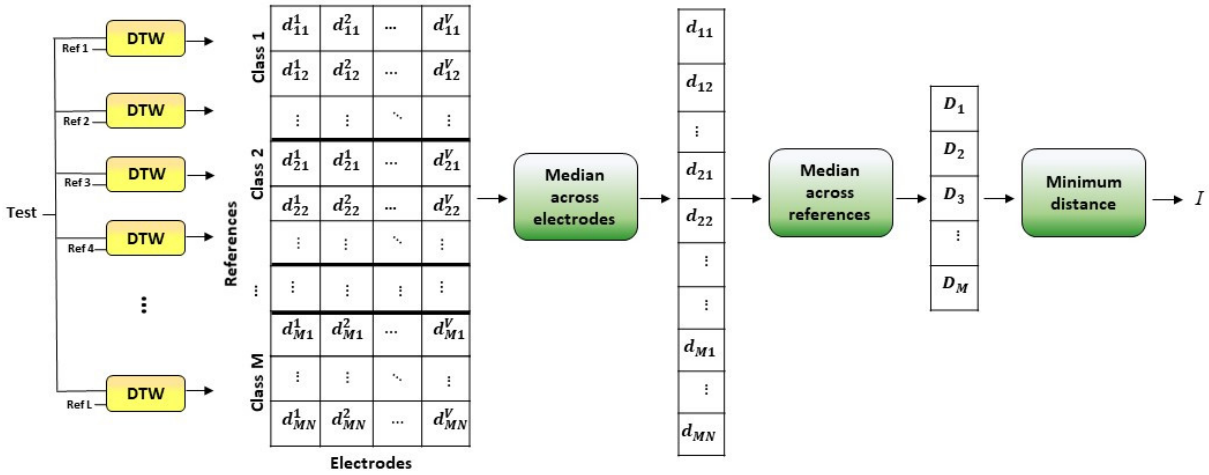


FIGURE 5.4: Block diagram of the proposed system. Each reference and test signal comprises of 64 channels (from 64 electrodes).

and N reference trials for each class, the total number of reference trials are $L = M \times N$. As the number of electrodes is 64, the total number of distances calculated is $L \times$ 64.

For classification based on the calculated distances, a fusion technique was applied. In this technique, the distance d_{mn} between the test trial and the n^{th} reference trial from the m^{th} class is given as the median of the 64 distances calculated for all available electrodes. Therefore,

$$d_{mn} = median(d^1_{mn}, d^2_{mn}, d^3_{mn}, ..., d^e_{mn}) \tag{5.25}$$

where d^i_{mn} is the electrode-by-electrode distance and i denotes the electrode index. Median is more suitable than the mean once the distance between the test and all reference trials has been calculated, because it rejects the noise in some of the electrodes. The trial-to-class difference is subsequently calculated as:

$$\begin{aligned} D_m = median(d_{m1}, d_{m2}, ..., d_{mN}) \\ m = 1, 2, 3..., M \end{aligned} \tag{5.26}$$

where m is the class index and M is the number of classes. Each test trial was classified by determining the class label that yields the minimum distance D_m

$$I = min_m[D_m] \tag{5.27}$$

5.8 Recognition

For the experimental evaluation of our method, we used *Action words*: "*Run*", "*Swim*", and "*Write*". The results were estimated using three word pairs from the database "*Run and Swim*", "*Run and Write*", and "*Swim and Write*" for binary classification. In general, 10 trials for each class were recorded from each subject, however some subjects were left with only 9 trials as a result of artifact rejection. This study focuses on subject-dependent classification, therefore, only ten trials for each class could be used. We evaluated the system by splitting the data in two different manners. In the first evaluation, 50-50 split was used to create test and reference spectrogram, whereas the second evaluation was conducted using leave-one-out cross validation and the results from each trial were averaged for each subject separately. As each electrode provides signal of different amplitude range, the reference signal from all the electrodes were

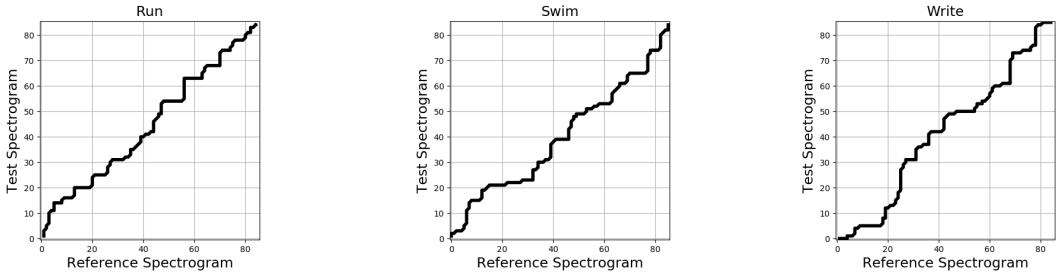


FIGURE 5.5: Warping path for the temporal alignment between a reference spectrogram and a test spectrogram for the three imagined words in our experiments.

z-scored (Kokoska & Zwillinger, 2000) in the time domain before performing time-frequency transformation. This improved the effectiveness of fusion when distances from different electrodes were calculated. Test data was also z-scored with respect to training data.

5.8.1 Classification of Imagined Speech

TABLE 5.4: Classification accuracy on *Action words* in 50-50 split.

Words	1	2	3	4	5	6	7	8	9	10	11	12	Average
("Run", "Swim")	70	75	87.5	50	62.5	70	60	70	80	50	62.5	62.5	**66.7**
("Run", "Write")	80	50	75	50	50	90	70	60	60	60	62.5	62.5	**64.2**
("Swim", "Write")	70	75	100	90	100	70	40	50	90	60	37.5	75	**71.4**

TABLE 5.5: Average accuracy of 12 subjects using correlation and cosine distance in DTW on *Action words* in 50-50 split.

Words	Cosine	Correlation
("Run", "Swim")	60.0	66.7
("Run", "Write")	63.4	64.2
("Swim", "Write")	64.2	71.4
Average (%)	**62.3**	**67.4**

Four sets of results were obtained for 12 subjects, covert speech. DTW of time-frequency signal measured the level of energy in the given frequency band for the reference and test spectrograms, the frequency range of $6-330Hz$ was used in this analysis. Mean of test spectrogram was normalized according to reference spectrogram to bring values of test and reference data to same order, small differences in features can influence distance calculation. Each reference and test signal comprises of 64 channels (from 64 electrodes) and DTW is applied on an electrode-by-electrode basis. The purpose of this work was to recognize unspoken words by means of EEG spectrogram, therefore we tried two distance measures in our DTW method cosine and correlation distances. The comparison is shown in Table 5.5.

DTW was applied under normal constraints (Sakoe & Chiba, 1978). Classification was initially performed using the minimum distance for each electrode, although the combined results were not good for all electrodes. To improve recognition performance, the proposed fusion techniques, presented in the section 5.7.1 was used to combine distances from different electrodes and improve overall classification performance. First the method was evaluated using 50-50 train-test data split test, where the final distance matrix contains five distances from each class for an electrode. Distances were evaluated for each electrode separately and were later combined using fusion method. Results are shown in Table 5.4.

TABLE 5.6: Classification accuracy on *Action words* in leave-one-out cross validation manner.

Words	1	2	3	4	5	6	7	8	9	10	11	12	Average
(“Run”, “Swim”)	75	68.7	50	56.2	81.2	50	50	68.7	56.2	50	75	56.2	**61.4**
(“Run”, “Write”)	75	62.5	68.7	50	68.7	37.5	50	62.5	62.5	56.2	75	56.2	**60.4**
(“Swim”, “Write”)	50	56.2	62.5	75	56.2	56.2	37.5	81.2	75	68.7	43.7	50	**59.3**

From the results it can be inferred that the trials of an imagined word have similarity in frequency patterns, which can be exploited for recognizing covert speech from EEG signals. As correlation distance performed the best, therefore, for rest of analysis in this chapter we used correlation distance.

The performance of the proposed method was better in comparison to the previous method with LDA and CSP features. However, DTW and electrode fusion failed to achieve high recognition rate in subject-dependent manner. Therefore, the method was not evaluated in subject-independent manner, where the recognition task become more challenging due to inter-subject variability (Bashivan et al., 2015).

5.8.2 Region Based Classification

In the above experiment, distances from all the electrodes were used to calculate the dissimilarities between the test and reference spectrograms. As can be seen in (5.26), median is calculated using all the electrodes, however only the electrodes chosen by calculating median are used as the final distance value.

TABLE 5.7: Electrodes in different brain areas.

Area 1:	F5, F7, FT7, FC5, T7, TP7, P3
Area 2:	F3, F1, FZ, F2, F4, F6, FC1, FCZ, FC2, FC4, FC6, FP1, FPZ, FP2, AF3, AF4
Area 3:	P3, P1, PZ, P2, P4, PO7, PO5, PO3, POZ, PO4, PO6, PO8, O1, OZ, O2, TP7

Therefore, we investigated the electrodes that were used in the distance calculation to make correct prediction of the test input. Electrodes for all the trials and subjects were investigated, total of 468 instances were estimated for electrode selection 288 in leave-one-trial out manner and 180 for 50-50 split. Electrodes that were repeatedly ($\geq$ 10 times) used by the fusion calculation across 64 channels were selected as important electrodes. These electrodes were mainly spread over three main brain areas. The three brain areas were Broca's area, Wernicke's area, and temporal lobe as Area 1, Frontal lobe as Area 2 and Occipital, superior parietal lobe and superamarginal gyrus as Area 3. Electrodes from different regions of the brain can provide crucial information about the areas playing important role in though recognition. Electrodes from frontal part of the brain could suggest decision making aspect of the brain. Most of the areas cover the left hemisphere and middle part of the brain which have been known to play an important role in language processing and production (Al-Fahoum & Al-Fraihat, 2014; Flinker et al., 2015; Pei et al., 2012).

Further, results were calculated for three brain areas separately using DTW. Results in figure 5.6 shows area 3 that is electrodes covering Wernicke’s area, Superamarginal gyrus, Occipital lobe and Superiorperital lobe achieved best recognition with 66.2% accuracy compared to the other two regions. These results are supported by some of the previous studies, it is known that a silent reading task involves Parietal lobe (Petsche et al., 1992) and Wernicke 's area is involved in speech production and comprehension (Pei et al., 2012). Further, speech in mind can be conceptually considered an impression or visual imagery of sentences. This has been proven in (Suppes et al., 1999), which showed that, on certain occasions, brain responses to images resembled responses to the verbal representation of those images.

Another important fact is that Occipital, Parietal, and Temporal lobes are known to play an important role in word processing, when written words are presented visually (von Stein et al., 1999). As visually

FIGURE 5.6: Average classification accuracy for three word pair from electrode groups from three brain areas.

perceived information flows from occipital lobe to the parieto-frontal region of the brain (Dentico et al., 2014), electrodes from those areas, therefore conclusion can be drawn that most useful information can be extracted from these regions of the brain. Specifically, signals from Wernicke's area, Superamarginal gyrus, Occipital lobe, and Superiorparietal lobe have recognition rate above chance rate with only 16 electrodes. The respective head model is shown in figure 5.7 (the image was taken from (Torres, 2017) and edited), these regions have been known to play role in speech processing, (De Benedictis et al., 2014; Pei et al., 2011). Signals from all the electrodes and classes (imagined words) were normalized so that they have zero mean and unit standard deviation.

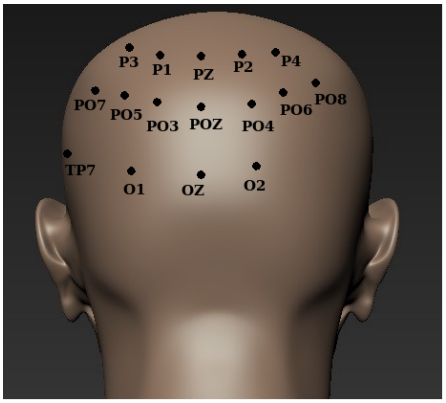


FIGURE 5.7: Electrodes covering Area 3, Wernicke's area, Superamarginal gyrus, Occipital lobe and Superiorperital lobe.

An alternation to (5.26) was made by taken the mean across the regions rather than the median, as it was assumed to have given better results, however, it did not increase the recognition rate. Although, electrodes covering Wernicke's area, Superamarginal gyrus, Occipital lobe, and Superiorperital lobe provides sufficient information for recognizing imagined speech. However, the recognition rate with area 3 was low compared with accuracy achieved when all electrodes were used for fusion.

5.8.3 Frequency based classification

An important part of the experiment was to investigate the frequency range that play an important role in better recognition of imagined speech. Several frequency bands of interest were investigated, frequencies

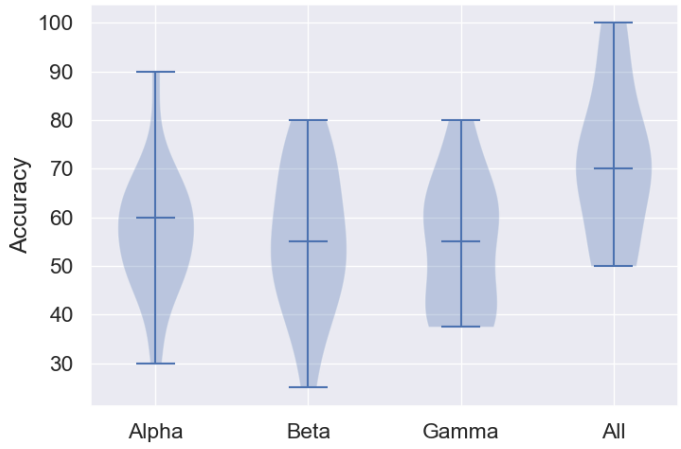


FIGURE 5.8: Average classification accuracy for three word pair for different frequency bands. As can be seen best results are achieved when all the frequency bands are used together. Whereas, gamma band performs better than alpha and beta.

containing information from alpha $(6-12Hz)$, beta $(13-30Hz)$, and gamma $(31-330Hz)$ were investigated. Raw EEG signals were filtered using FIR (zero-phase) bandpass filter with Hamming window to extract the frequency band of interest. DTW and electrode fusion were applied to the spectrograms from each band separately. The results are shown in figure 5.8. As shown, the best results were achieved when all the frequency bands were used together. In other words, our system performed best when more frequency information was present. These results are in consistency with finding in a recent study (Dash et al., 2021). Where best performance was obtained when all the frequency bands were used together.

5.8.4 Limitations of the Proposed Method using DTW

The proposed method using DTW achieved better accuracy in recognition of imagined words from EEG signals compared to the CSP and LDA method. However, the proposed methodology have limitations; for example, DTW is computationally expensive and slow to be implemented in real time.

5.9 Conclusion

In this chapter, linear and non-linear machine learning methods were evaluated for feature extraction and classification of imagined words (covertly spoken words). First, the performance of spectro-temporal and spatio-temporal features were evaluated using LDA, statistical features, and CSP. The performances of the features were evaluated separately and when combined together. For classification, three different classifiers were used, SVM, KNN, and DT. Where SVM achieved highest average (60.3%) accuracy with a best of 93.7% recognition rate for a subject and word pair (swim, write). Spectro-temporal features performed slightly better than spatio-temporal features. This showed that the importance of spectral information in the input for achieving better recognition rate for imagined words. However, average accuracy achieved by combination of different feature set was around chance level, failing to achieve high recognition rate. Low performance with feature extracted using LDA and CSP indicated to the unaligned inter-trial events of the imagined words. Further, lack of muscle activity in covert speech task made it even more difficult to capture discriminative events, especially in spatio-temporal patterns. This analysis demonstrated that linear methods such as the LDA are ineffective in recognition of imagined speech.

To overcome this problem of unaligned inter-trial events in EEG signals, we used DTW for similarity matching of imagined words from EEG signals. The DTW achieved better results in comparison to

previous techniques for feature extraction. This suggests that by overcoming the problem of inter-trial variation we can improve the accuracy for recognition of imagined speech. From DTW experiment, two sets of results were obtained, first, using 50-50 train-test split and other using leave-one-out cross validation. It was found that performance of the method with 50-50 split was better, whereas the performance of the proposed system slightly reduced when evaluated under leave-one-out manner. This showed the limitation of proposed methods in dealing with spectro-temporal variations in larger training-testing data. Further, using the proposed electrode fusion technique we found three regions of interest: Area 1: Broca's area, Wernicks's area, and temporal lobe, Area 2: Frontal lobe, and Area 3: Wernicke's area, Superamarginal gyrus, Occipital lobe and Superiorperital lobe. When the performance of three area's was evaluated separately, Area 3 achieved highest recognition. Finally, the analysis with different frequency bands did not provide any useful information and the best recognition was achieved with all frequencies combined together. The improvement in recognition rate using DTW showed that linear methods applied in the previous experiment were unable to capture meaning information from EEG signals due to inter-trial variability.

5.10 Summary

Recognition of imagined speech may be the most practical brain-computer interface for individuals with locked-in syndrome and speech disabilities. As a result, studies have investigated various techniques for recognition of imagined speech from EEG signals. In this chapter, linear and non-linear methods were investigated for recognition of imagined words. The LDA and statistical methods were used to extract time-frequency features from spectrogram of EEG signals. Also, spatio-temporal features were extracted using CSP algorithm. For classification, three classifiers were used: SVM, KNN, and DT. The evaluation involved classification of three pair of imagined words, recorded from 12 subjects. Three set of features were investigated for recognition of imagined speech (1) time-frequency features, (2) spatio-temporal features, and (3) combined. The classification results achieved from all the features and classifiers were around chance level.

Therefore, a non-liner method was evaluated using DTW for measuring similarity between imagined words from EEG signals. In addition, this chapter proposed a simple yet effective method for combining distances from multiple electrodes. The proposed system was evaluated using two different train-test splits: (1) 50-50 split, (2) leave-one-out cross validation. The results achieved in both the splits were above chance level and outperformed feature extraction methods used earlier in the chapter. Further, this chapter presented an investigation of the electrodes from different brain areas that play an important role in language production and processing.

Chapter 6

Classification of Imagined Words using Convolutional Neural Network

In the first part of this chapter, classification was performed between covert speech (words) and non-speech activity from EEG signals. In addition, recognition of covertly spoken words was performed. In order to discriminate between the different tasks, a convolutional neural network (CNN) was used to learn spatio-temporal features. The proposed CNN architecture performed well in recognition of imagined speech vs non-speech (silence) tasks. However, the performance of the CNN with spatio-temporal feature was low in recognizing imagined words. Therefore, in the second part of this chapter we proposed a CNN-attention network that learns spectro-temporal features for recognition of imagined words. In addition, an electrode selection method is proposed for choosing the electrodes with most task-discriminative information.

6.1 Motivation

A thought-to-text BCI system should be able to discriminate brain signals (EEG signals) produced during imagined speech and non-speech tasks such as visual perception and no-activity (i.e., relaxing state EEG signals). A limited number of studies have performed classification between imagined speech and non-speech activity from EEG signals. Some of those studies compared vowel imagery with a control state (no-activity) (DaSalla et al., 2009b; Yoshimura et al., 2011), whereas others compared EEG signals produced during mental repetition of words with resting state EEG signals (Sereshkeh et al., 2017). A more recent work compared imagined words and syllables with visual attention and relaxing (resting state) tasks using EEG signals (Alsaleh, 2019). In the first part of this chapter, we performed classification between imagined speech and non-speech activity.

The main focus of this chapter is recognition of imagined words. Therefore, we performed classification between two imagined words. However, recognizing words (speech) from EEG signals is difficult because EEG signals suffer variation in time, frequency, and electrodes information. During normal speech, humans can produce a word in about 0.33-0.5 seconds (Alsaleh, 2019), thus detecting event related to a given word becomes difficult, considering such short time of word processing in the brain. Moreover, the temporal occurrence of an event produced by covert speech does not involve motor movement and suffers from inter-trial variations, making it even more complex to detect. Furthermore, EEG signals have multiple electrodes, which can lead to increase in processing time and resources. In addition, not all the electrodes contribute equally to recognition of a given activity (word). Therefore, in order to design a text-to-speech BCI it is important to construct a method that can choose electrodes with task discriminative information and a method for pattern learning that is robust to variations in the EEG signals.

Many techniques have been proposed for feature extraction such as the Fourier transform, auto-regressive (AR) modeling, eigen-analysis, CSP (Al-Fahoum & Al-Fraihat, 2014; DaSalla et al., 2009a), and the wavelet transform. However, most of these techniques suffer from limitations and are invariant to deformations in EEG signals. Although, the DTW-fusion method in the Chapter 5 achieved decent

results, it did not take into account the class relevant components present in the time-frequency information. In other words, certain frequency components which did not contribute to class discrimination were also included in measuring similarity between different EEG signals of mentally spoken words.

To avoid these problems, this chapter used deep learning technique for feature learning and classification. Deep learning has been successful in many fields and have also achieved state-of-the-art results in recognition of motor movement and motor imagery from EEG signals (Bashivan et al., 2015; Donahue et al., 2015; Tabar & Halici, 2016). Further, deep learning techniques such as attention and self-attention mechanism accomplish the state-of-the-art results in dealing with sequential data (Bahdanau et al., 2014; Vaswani et al., 2017).

This chapter propose an application of convolutional network for recognizing covert speech and non-speech activity from EEG signals. Further, in order to achieve better recognition of imagined words, an electrode selection method and CNN-attention architecture are proposed for learning spectro-temporal patterns from EEG spectrograms. The following are the contributions of this chapter:

- An electrode selection method, which select electrodes based on mean power calculated from spectrograms. The proposed method helps in reducing the dimensionality of the EEG data and reducing the training time of the network.
- Application of CNN-attention network to recognize imagined words by learning spectro-temporal components from the input spectrograms. The performance of the proposed network is tested against a baseline CNN-LSTM-attention network. The combination of the proposed electrode selection technique and CNN-Attention on 12 participants, achieved high accuracy in recognition of covertly spoken words.
- Classification of imagined speech and non-speech activity from EEG signals using convolutional neural networks (CNNs) and comparison with previous work.

6.2 EEG Data-sets

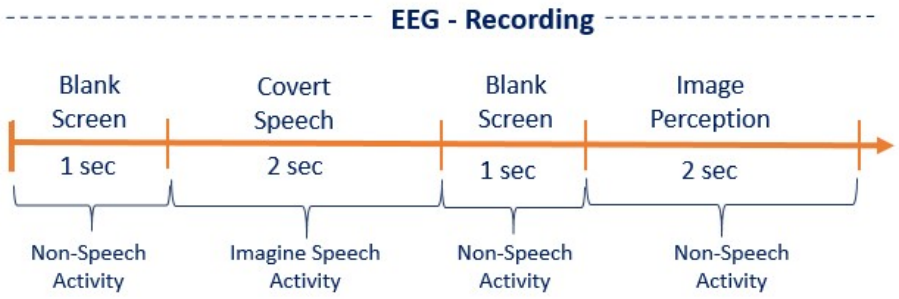


FIGURE 6.1: The sequence of single trial EEG recording. In total, 100 trials were recorded for each subject.

The focus of the first part of this chapter is to discriminate between imagined speech and non-speech tasks. Therefore, for experimental evaluation, we used EEG data acquired from imagined speech and two non-speech tasks. EEG signals from 12 subjects were used to perform the analysis, contaminated trials or noisy data were rejected. Each word had ten trials in total for each subject. However, after artifact rejection and filtering some subjects had only nine trials for each class. For evaluation of electrode selection method and the CNN-attention network proposed in this chapter, EEG signals were used from three mentally spoken words:"*Run*", "*Swim*", and "*Write*". Further, to discriminate between imagined and non-speech activity, we used EEG dataset from three tasks:

- **Imagined Speech**: Words were presented in random order on a computer screen for 2 seconds and subject was asked to mentally read the word once as soon as it appeared.

- **Visual Perception**: Subjects were presented with an image for 2 seconds, the image was associated with word presented in imagined speech task. The participants were instructed to watch the presented image.
- **Resting State**: The participants were presented with blank screen and were asked to perform no mental or physical activity. This is also known as the relaxing state task which lasted 1 second.

The recording procedure and time of each task is shown in figure 6.1. Where EEG signals recorded during the "*Visual perception*" and "*Resting State*" tasks were used as non-speech activity in the first part of this chapter. Detailed information about the recording procedure is provided in chapter 3.

6.3 Discrimination between of Imagined speech and Non-speech from EEG signals

The aim of this section is to discriminate between imagined words (imagined speech) and non-speech activity using EEG signals. We used the CNN to perform feature learning and classification on raw EEG data. The CNN offer a desirable attribute for end-to-end learning without prior feature extraction, which has been exploited in variety of past applications (Schirrmeister et al., 2017). This property of CNN is most useful for designing a BCI system.

However, CNNs have been most effectively used for recognition tasks involving images,i.e., a two dimensional input. In comparison, EEG signals are dynamic time series derived from multiple channels on the scalp. To deal with this issue, we used EEG data as a two dimensional matrix of size $T \times C$, where C is the number of electrodes (channels) and T refers to the time points in the EEG signals. Further, to deal with the dynamics in the EEG signals, we divided T time points into smaller window frames W_{tn}, where n is the number of windows. The network contains n parallel CNNs to process each window separately. Where each frame is a two dimensional matrix (similar to an image) of size $W_t \times C$. Each window, W_t is an input to a separate CNN for local feature learning. These local features from each window are combined at a latter stage to create global features. Processing of the input is shown in figure 6.2.

6.3.1 Network Architecture

The architecture of the model used in this experiment had three main components: (1) windowed CNN which performs spatio-temporal feature learning on multiple time windows in parallel, (2) combination of global average pooling and dense layer for learning distinct representation of features, and (3) dense layers for extraction of high level features. The network architecture is shown in figure 6.2.

TABLE 6.1: The parameters used in the architecture of the network. GAP: global average pooling; K/N/DR: kernel/neurons/dropout rate.

Layer	K/N/DR	Filter	Stride	Activation
CNN (Block 1)	64	3 ×3	2 ×2	*ReLU*
Dropout (Block 1)	20%	-	-	-
CNN (Block 2)	128	3 ×3	2 ×2	*ReLU*
Dropout (Block 2)	20%	-	-	-
GAP-Dense (Block 3)	128	-	-	*ReLU*
Dense (Block 3)	128	-	-	*ReLU*
Dense	64	-	-	*ReLU*
Dense	2	-	-	*Sigmoid*

To obtain short-term features, the input was partitioned into numerous windows W_{tn} and used n parallel CNNs to extract spatio-temporal features from each window W_t. The network consist of two convolutional blocks, with each block containing single convolutional layer, batch-normalization, and

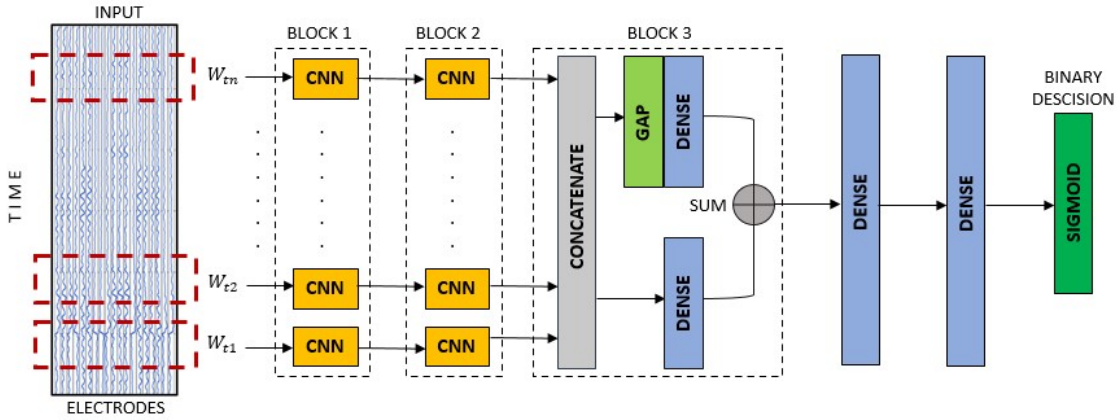


FIGURE 6.2: Architecture of the convolutional network used for recognition between imagined speech and non-speech activity. The spatial ordering of the electrodes was in accordance with their position in the cap, as shown in figure 4.9.

dropout layer. Where the convolutional layer in the first block filtered the data using 64 kernels with a receptive field of size 3×3 and stride of size 2×2. The convolutional layers in the second block have 128 kernels of size 3×3 applied with a stride of size 2×2. where the number of kernels is $k = 64$ or 128 in (2.1). For convolution we used "same" (zero) padding in order to preserve the spatial resolution of the input. The output features from the parallel CNNs were concatenated and fed to block 3, which contains two separate layers: a dense layer and 1 dimensional global average pooling layer followed by dense layer. Reason for using two separate layers was to obtain distinct feature representations. One dimensional global average pooling was used to combine spatial activity in the input. Dense layers in block 3 had 128 nodes and the *ReLU* activation. Further, the two feature representations were added together, and the summed features were passed to a dense layer with 64 nodes and the *ReLU* activation. In the classification layer, the number of nodes was two (classifier layer) with the *sigmoid* function (2.5) for classification.

6.3.2 Training

The network was trained for 120 epochs, with mini-batch gradient descent of size 32. Cross entropy loss was minimized using the Adam optimization algorithm (Kingma & Ba, 2014) with learning rate of 0.001. In order to avoid overfitting, the network used batch-normalization as well as dropout regularization (Srivastava et al., 2014) with dropout rate of 20%. Dropout enables robust feature learning by setting output from some hidden neurons to zero for an input. Therefore, with each input, the network samples a different architecture, but weights are shared among all these architectures (Krizhevsky et al., 2012).

6.3.3 Design Choices

Different design choices were investigated for the network architecture which would impact its performance. By varying the design choices, the activation functions, the kernel size, and the filter size, we tried to develop some insights into the network. Performance of the network was most affected by the activation function. For example, the *ReLU* gave the best results. Furthermore, dense layers were implemented without the *ReLU* activation, which resulted in reduced recognition rate. This leads to the conclusion that the non-linearity makes dense layers more robust in feature extraction towards the end of the network. In addition, the network performance was assessed using filters of different sizes, among which the 3×3 filter performed best. This is because in a small spatio-temporal window, a larger filter (i.e., 7×7 or 9×9) can over-look important features and skip essential details. Further, the network's performance was also evaluated by removing the 1D global average pooling layer, which led to reduced

overall performance. However, the network performed even worst when the parallel dense layer was removed in block 3. The parameters used in the network architecture are shown in Table 6.1. The network performance reduced by increasing the depth i.e., the convolutional, dense and/or pooling layers in the network, with best results being achieved by two convolutional layers.

6.4 Results

In order to discriminate between imagined speech (mentally spoken words) with non-speech task, EEG signals were acquired from three activities: *imagined speech*, *visual perception*, and *black screen/resting state*. Where resting state and visual perception task were regarded as non-speech activity. The imagined speech activity contained EEG signals produced during mentally spoken words. The proposed network was evaluated on 12 subjects, where each subject performed 100 trials for each class. However, some subjects ended up with less trials due to the removal of contaminated trials. Two different experiments were conducted to discriminate imagined speech with non-speech activity. In the first experimental protocol, imagined speech activity was discriminated against visual perception, whereas the second evaluation was performed between imagined speech and resting state (blank screen). The experimental evaluation was performed in subject-dependent manner. The results were evaluated using leave 10% out cross-validation method for each subject, where the network was trained on 90% of the dataset and tested on 10% of the dataset. The results obtained by varying training and test set were averaged.

6.4.1 Classification between Imagined Words and Visual Perception

TABLE 6.2: Discriminating imagined speech with visual perception for different time windows in the EEG signals. The network was evaluated on EEG signals bandpass filtered between 1-80Hz frequency range.

Subject	0-500ms	500-100ms	1000-1500ms	1500-2000ms
1	92.3	53.8	52.7	54.3
2	68.3	48.8	54.6	51.1
3	84.8	50.6	64.7	63.1
4	71.3	52.5	50.5	43.1
5	87.3	49.5	45.5	53.2
6	74.0	51.0	56.4	48.2
7	86.8	56.9	48.2	51.2
8	84.5	57.9	50.4	51.6
9	86.2	54.2	55.0	58.8
10	82.3	52.1	53.5	56.3
11	81.6	60.1	57.1	52.8
12	91.3	57.7	51.8	48.8
Average	**82.5**	**53.7**	**53.3**	**52.8**

In the first evaluation step, EEG signals were band pass filtered between 1-80Hz using a zero phase, non-causal FIR filter. Discrimination of imagined words with visual perception (non-speech activity) was performed on four separate windows each of 500ms, obtained by splitting the 2000ms trial. The network was evaluated on each window separately to investigate the time window which provides best recognition rate. Therefore, the size of each window used for this analysis was $T \times C$, where $C = 64$ is the number of channels and $T = 500$ is the number of time points in the EEG window. Therefore, the input to the network was $n = 10$ small windows frames of size $W_t \times C$, with $W_t = 50$ and n is the number of parallel CNNs used in the network. The input was a multi-dimensional matrix of size $n \times W_t \times C$. The results are shown in Table 6.2. As shown, the highest recognition rate was achieved by the 0-500ms time window.

Further, the networks performance on different frequency range was also investigated. This evaluation used the 0-500ms time window, which performed best in the earlier evaluation. First, the network was evaluated for a broader frequency range, $1 - 80Hz$. In the second evaluation, where high frequency

TABLE 6.3: Evaluation of the network's performance for discriminating imagined speech and visual perception using different frequency ranges. The performance was evaluated using 0-500*ms* time window.

Range	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
1-80Hz	92.3	68.3	84.8	71.3	87.6	74	86.8	84.5	86.2	82.3	81.6	91.3	**82.5**
1-30Hz	91.7	68.5	88	69.9	89.9	77.5	87	86	83.4	82.5	81.8	90.9	**83**
4-30Hz	93.1	72.3	91.8	78.5	75	80.4	88.5	81.6	84.8	75.9	86.2	95.8	**83.6**

information 31-80Hz was removed which led to slight improvement in the recognition rate. Another analysis included evaluation of EEG signals by removing low frequency information 1-4Hz which led to further increase in the recognition rate. The results are shown in Table 6.3. The best recognition rate of 83.6% was achieved using EEG signals with frequencies between 4-30Hz. Although, performance for other two frequency range were also comparable.

6.4.2 Classification between Imagined Speech and Resting State

TABLE 6.4: Classification accuracy for recognition of imagined speech and resting state.

Range	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
1-80Hz	90.1	83	88.6	84.4	82.6	83	86.9	94.7	84.3	81.1	95	95.7	**87.4**
1-30Hz	89.5	80.8	91.6	84.7	84.7	84.3	86.9	92.2	84.9	80.4	95.6	95.5	**87.6**
4-30Hz	88.6	66.1	86.4	88.8	74	86.2	89.4	90.2	83.2	96.4	72.6	86.9	**84**

To classify between imagined speech and non-speech activity (blank screen), EEG signals of 0-500*ms* time window were used for training and testing the network. This window was chosen based on the results obtained from classification between imagined speech and visual imagery task. In the first evaluation step, EEG signals contained frequency range between 1-80Hz. In the second evaluation, high frequencies (gamma band, $31-80Hz$) were removed from EEG signals and the network was evaluated on EEG signals containing frequencies between 1-30Hz. For the imagined speech class, the network was trained on EEG data from $N-1$ words and N^{th} word was used for testing. In other words, the network was trained on EEG signals from nine mentally spoken words and tested on the 10^{th} word. The results are shown in Table 6.4. As can be seen, EEG signals containing 1-30Hz frequencies perform best.

Comparing the results in Table 6.3 and 6.4 shows that classifying imagined speech (words) vs no-activity is much easier than imagined speech vs visual imagery. Further, the results suggests that initial time window of 0-500*ms* are provide the most task discriminative information. Although, 1-30Hz performed well for both the recognition tasks, but was not the best performing frequency range for distinguishing visual perception and imagined words.

TABLE 6.5: Comparison with other methods for classification of imagined speech and no-activity.

Method	Subjects	Trials per Subject	Channels	Accuracy
(Sereshkeh et al., 2017)	12	60 per class	64	75.7%
Proposed	12	100 per class	64	**87.6%**

Comparison: Results obtained by the proposed method were also compared with previous studies which performed classification between imagined speech and resting state (no activity). The comparison is shown in Table 6.5. Although, a direct comparison cannot be made because EEG dataset used in past studies were not publicly available. The results achieved by the proposed architecture significantly out-performed previously proposed methods.

6.4.3 Recognition of Imagined words

In this section, classification was performed between two imagined words from EEG signals. EEG signals were used for three imagined words: *"Run"*, *"Swim"*, and *"Write"*. Binary classification was performed between two words and therefore we obtained results for three word pairs, shown in Table 6.6. Each subject had 10 trials for each word, we evaluated the network performance for three frequency range using 0-500*ms* time window. The results are shown in Table 6.6. The average accuracy achieved by the network is just below the chance level, where the chance level is considered to be 55.0%. This shows that spatio-temporal features extracted by the CNN were not robust enough to recognize imagined words from EEG signals.

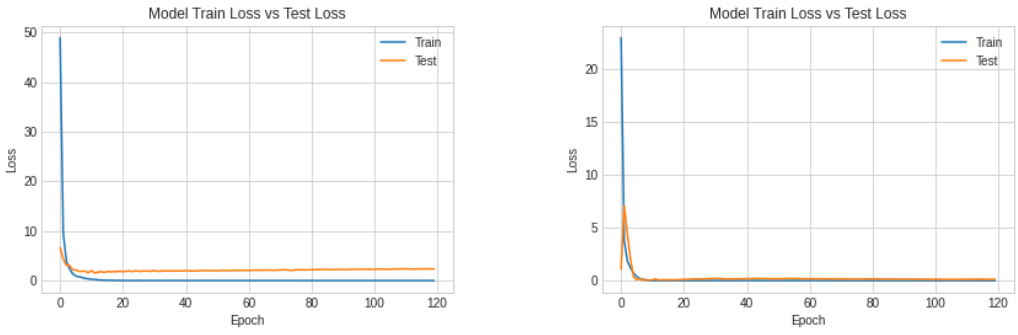
TABLE 6.6: Classification accuracy for recognition of imagined words in subject dependent manner.

Range	("Run", "Swim")	("Run", "Write")	("Swim", "Write")	Average
1-80Hz	55.4	47.1	52.0	**51.6**
1-30Hz	56.2	49.7	51.9	**52.6**
4-30Hz	53.7	48.5	50.9	**51.0**

6.4.4 Limitations of Spatio-Temporal Information and CNN for Imagined Word Recognition.

From the above analysis it can concluded that the method of extracting spatio-temporal features with the CNN is effective in classification between imagined speech (words) and non-speech activity. However, spatio-temporal features extracted by the CNN are not robust enough to recognize the imagined words from EEG signals. Poor performance of the network could be a result of limited training data. Further, the CNN networks are known to learn spatial patterns and cannot effectively learn temporal dependencies. In addition, the covert nature of imagined speech activity does not involve muscle movement and makes the spatio-temporal pattern in EEG signals more difficult to learn. Lack of spectral information can also be associated with low recognition rate for imagined words, because EEG signals in the time domain can sufferer from several limitations due to its non-stationary nature. Therefore, in the next part of this chapter we use time-frequency features in recognition of imagined words.

Further, the proposed CNN network suffers from overfitting, when differentiating imagined speech and visual perception task, which is evident from the learning curve presented in the figure 6.3. This shows that the network performs well on the training data however, fails to achieve high performance on test data.



(a) (b)

FIGURE 6.3: The training (blue) and test (orange) curve for the CNN network for (a) imagined speech and visual perception; (b) imagined speech and resting state. The training curve (a) imagined speech and visual perception indicate overfitting.

6.5 Spectro-Temporal Feature learning using a CNN-Attention Network

This section, propose a framework for recognizing imagined words from EEG signals using a *mean power-based electrode selection technique* and *CNN-Attention network* shown in figure 6.4. The framework contains two main components; a channel selection method which provides with electrodes containing the most task discriminative information and a *CNN-Attention* network. The EEG signals for each word were converted into spectrograms. This was done by performing STFT using windowing function on EEG signals for each electrode, STFT performed using windowing avoids the assumption of stationarity in an EEG signal (Cohen, 2014). Spectrograms from all electrodes were used to select most informative electrodes which would reduce the dimensionality of the data and increase recognition rate. The spectrograms from multiple electrodes presents time, frequency, and spatial information to the network for optimal feature learning.

An application of the CNN-attention network is proposed to learn spectro-temporal patterns from the input spectrograms. The network treats the input spectrograms as a time-varying information, where CNNs extract important frequency components from each time point separately. Incorporating CNNs at initial stage has the advantage of reducing the input size. Further, the use of CNN allows our system to handle correlations between neighboring frequencies. The proposed network used a dense layer after the CNN block to perform dimensionality reduction on the spatio-spectral features extracted from the CNN blocks. The dense layers had the *ELU* activation to learn non-linear trends in the extracted features. Further, the network used attention mechanism to focus on the most informative temporal-features in the output produced by the CNN-dense layers. The self-attention mechanism also helps the network to learn sequential relationship between the extracted features. This makes the network capable of learning important components at each time point. The proposed techniques have some interesting properties:

- The electrode selection method provides electrodes that helps in achieving high recognition rate and reduces the dimensionality of the EEG data. This also helps in reducing the training and testing time of the network.
- The proposed CNN-attention architecture learn important spectral components at each time point separately from the input spectrograms using the CNNs. Therefore, treating it as a time varying input.
- The proposed architecture uses the attention mechanism rather than using the LSTM to learn the temporal dynamics in the spectrogram. This avoids the problems suffered by LSTM such as the memory limitations and slow learning rate (Monesi et al., 2020).

By using the electrode selection method and the CNN-attention network for feature learning and classification, we present a framework which is more powerfully in recognition of imagined speech from EEG signals. The method is evaluated on our EEG dataset for imagined speech.

6.5.1 Electrode Selection using Mean Power

EEG signals are captured by multiple electrodes from different sites on the scalp which increases the dimensionality of data. The EEG signals from large number of electrodes requires more resources and processing time. Further, not all electrodes add to high recognition rate. Therefore, it is essential to reduce the dimensionality of data by using an appropriate method for selecting electrodes. In the EEG studies, selection of electrodes is a challenging task. Aggregating features calculated from temporal and spectral domain for each individual electrodes into a vector is a standard procedure in EEG studies; however this method is not very effective (Bashivan et al., 2015). Instead, we propose an electrode selection method. The proposed electrode selection method (Algorithm 1) takes spectrograms as input from all the electrodes and choose top-K electrodes that are most informative to the recognition of imagined words, where K is a pre-defined value.

Algorithm 1 Select Top-K Electrodes

Requires: A matrix $X \in \mathbb{R}^{n\times C\times T\times F}$, where $T \times F$ are dimensions of a spectrogram $x_{t,f}$ belonging to an electrode C of the n^{th} trial. K the no of electrodes to be selected, and m_j the vector length.

Calculate mean across T for each spectrogram
create an empty array X_f of size $n \times C$
LOOP through all trials n and electrodes C
for $n \in 1,...,n$ **do**
  for $C \in 1,...,C$ **do**
    x_f = Mean($S_{t,f}$, axis=0) {Mean across time axis}
    Create vector of varying length from the frequency vector x_f.
    $x_{f'_1}, x_{f'_j},..., x_{f'_j} = x_{f-n_1}, x_{f-n_2},..., x_{f-n_j}$
    {where length of $x_{f'_j} > x_{f'_2} > x_{f'_1}$}
    $m_1, m_2,..., m_j$ = Mean($x_{f'_1}$), Mean($x_{f'_2}$),..., Mean($x_{f'_j}$)
    if ($m_1 > 0$ **or** $m_2 > 0$ **or** $m_j > 0$) **then**
      $X_f(n,C) = 1$ {C^{th} position in $X_f(n,C)$ is 1}
    else
      $X_f(n,C) = 0$ {C^{th} position in $X_f(n,C)$ is 0}
    end if
  end for
end for
Create an array containing the number of times an electrode C was 1
L = Sum($X_f(n,C)$, axis=0) {sum across n trials}
L = Argsort(L) {Arrange indices of values in descending order}
$topK = L(1:K)$ {Retrieve first K electrodes from the list L}
return $topK$

Given the training data $X \in \mathbb{R}^{n\times T\times F\times C}$ as input, where n is the number of training trials, T is the number of time points in the spectrogram, F is the number of frequency points in the spectrogram, and C is the number of electrodes. In the first step, the spectrograms are averaged along the time axis. Mean over time is calculated as it provides with a baseline normalized power estimate, which helps avoid the 1/f phenomenon. This is done for all the electrodes and trials separately. This provides with a frequency vector for each electrode for a given trial, defined as:

$$x_f = \frac{1}{N_T}\sum_{t}^{T} x_{t,f} \tag{6.1}$$

where $x_{t,f}$ is the input spectrogram for an electrode of a given trial and N_T is the total number of time points in the spectrogram. The training data X is transformed to $X_f \in \mathbb{R}^{n\times x_f\times C}$. Further, for each electrode in a trial, x_f is divided into overlapping windows of different size, where single window is given as:

$$x_f = x_1,...,x_N \tag{6.2}$$

where N is the length of frequency vector x_f, and a divided window is given as:

$$x_{f'} = x_{f-n} \tag{6.3}$$

where $f-n$ is the length of the new vector $x_{f'}$. Therefore, a new matrix $X_{f'} \in \mathbb{R}^{n\times x_{f'}\times C}$ is obtained. Each frequency vector is divided into smaller overlapping window to provide power representations at different frequency bands. In this analysis, parameter n in (6.3) was varied j=3 times using different

values of n, which provided three $x_{f'}$ of varying lengths: $x_{f'_1}$, $x_{f'_2}$, $x_{f'_j}$. Mean for single $x_{f'}$ is calculated as:

$$m = \frac{1}{N_{f'}} \sum_{f'=1}^{F'} x_{f'} \tag{6.4}$$

(6.4) is used to calculate j mean values; m_1, m_2, and m_j from $x_{f'_1}$, $x_{f'_2}$, $x_{f'_j}$ for a single electrode. A matrix X_f is estimated of size $n \times C$, an electrode is added to the matrix if $m_j > 0$ for an electrode of a given trial (Algorithm 1, step 8). The C^{th} position of the matrix had a 1 or 0 value, where 1 refers to a selected electrode and 0 refers to a rejected electrode. The matrix $X_f(n,C)$ was amended along the rows, i.e., n axis. The n rows were added to calculate an array, from which indices of K highest values were retrieved. These K electrodes were estimated to provide the most discriminative information about imagined words. The electrode selection was performed using training data and same electrodes were selected in the test data for classification purposes.

6.5.2 Architecture of CNN-Attention Network

The proposed network architecture has two CNN blocks and a dense block followed by the self-attention mechanism and the dense layers, where the last dense layer uses the *sigmoid* function for the binary classification task. The network architecture is shown in figure 6.4 and the parameters are presented in Table 6.7. Each block has T parallel 1-dimensional (1-D) CNN layers and batch normalization layers, where T is the number of time points in the spectrogram. The feature learning was performed by 1-D convolution at each time point (frequency vectors) in the spectrogram. Both the blocks contained a single CNN layer. The convolutional layer in the first block filters the data using 64 kernels with a receptive field of size 3 and a stride of size 2. This process can capture high-level features from the frequency vectors of the spectrogram. The convolutional layers in the second block have 128 kernels of size 3, applied with a stride of size 2.

TABLE 6.7: The parameters used in the CNN-attention network. K/N/DR: kernel/neurons/dropout rate.

Layer	K/N/DR	Filter	Stride	Activation
CNN (Block 1)	64	3	2	*ELU*
Dropout (Block 1)	20%	-	-	-
CNN (Block 2)	128	3	2	*ELU*
Dropout (Block 2)	20%	-	-	-
Dense (Block 3)	256	-	-	*ELU*
Attention	-	-	-	*tanh*, $Softmax$
Dense	128	-	-	*ELU*
Dense	64	-	-	*ELU*
Dense	2	-	-	*Sigmoid*

Deep learning models can benefit from strategically designed layers endowed with non-linearities (Donahue et al., 2015). Therefore, to learn non-linear patterns from the EEG spectrograms, two activation functions the sigmoid and the exponential linear unit (*ELU*) (Clevert et al., 2015) were used in the network. The CNN and dense layers had the *ELU* function. The *ELU* was considered over the *ReLU* function because *ReLU* suffers from the dying ReLU problem (Lu et al., 2019), whereas the classification layer had the *sigmoid* function 2.5. Batch-normalization was used in every block, which helped speed up the learning process by centering the data (Clevert et al., 2015).

6.5.3 Using Self-Attention Mechanism for learning Temporal Dynamics

It is known that the event associated to a particular task last few milliseconds in the brain (Butts et al., 2007). Therefore, not all the time points in the EEG signals are informative for detecting imagined

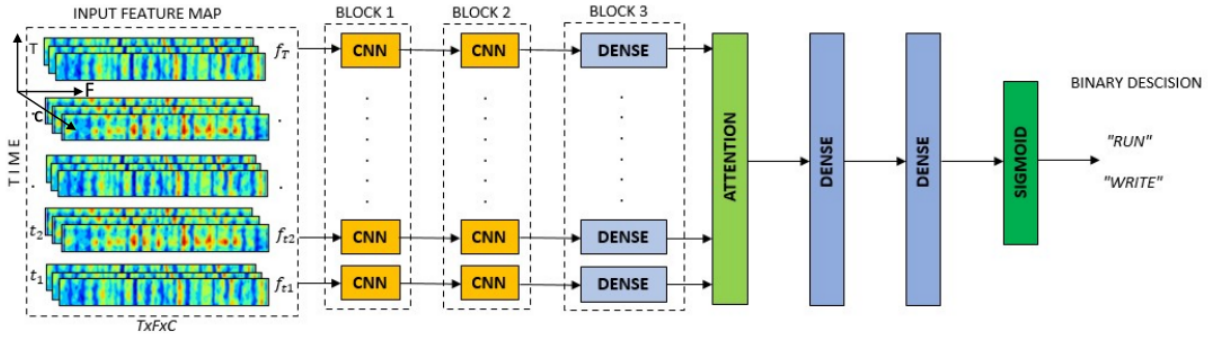


FIGURE 6.4: The architecture of proposed CNN-Attention network. Input is a multi-dimensional tensor in form of $T \times F \times C$, i.e, spectrogram from different electrodes combined to form three-dimensional input. The spatio-spectral of size $F \times C$ ($F = 86$) content at each time point is processed separately by T =86 parallel one-dimensional CNNs. Frequency features are extracted using a chain of two blocks containing the CNNs and batch-normalization layers. The output features from the CNN block as fed to the dense layers used for dimensionality reduction and then to attention layer for learning temporal patterns.

speech. Therefore, we used *self-attention mechanism* to focus on the most informative temporal-features in the output produced by the CNN-dense layers. To implement the attention in our network, we used the self-attention layer, where the output from parallel dense layers were used as input to attention layer to create a more informative global feature map g. The resultant global feature map has more weights assigned to discriminative information which contributes to classification of imagined words. The global feature vector produced by the attention layer. The self-alignment layer is trained using back-propagation algorithm and learn important time points in the feature-map using the gradient of the cost function (Bahdanau et al., 2014). The self-attention layer was followed by two dense layers with 128 and 64 neurons, respectively. Dense layers used the *ELU* as activation function, which made the network capable of non-linear features passed from previous layers. The last layer used the sigmoid function for binary classification.

6.5.4 Network Training

Both the networks were implemented in *Keras* (Chollet et al., 2015) with *Tensorflow* backend (Abadi et al., 2016). For subject dependent evaluation, the network was trained with the Adam optimization algorithm for 200 epochs, and mini-batch gradient descent of size 5. Whereas for subject independent evaluation, the network was trained for 200 epochs with mini-batch gradient descent of size 32. For both evaluations, the cross-entropy loss was minimized with learning rate of 0.0001 (Kingma & Ba, 2014). Due to weight sharing in convolutional networks, the gradient at different layers can vary widely (Bashivan et al., 2015). Therefore, a slower learning rate was used for training the network. Further, in order to avoid the problem of unstable gradient (Simonyan & Zisserman, 2014), we used *He* weights initialization method (He et al., 2015).

6.6 Classification of Imagined Words

For the experimental evaluation of our method, EEG signals were used from the dataset presented in chapter 3. EEG signals were used for three words, i.e., "*Run*", "*Swim*", and "*Write*". Three pairs of words were used for classification purposes. Raw EEG signals were transformed into spectrograms and used as input to the CNN-attention network. The spectrograms were converted into multi-dimensional input of dimensions $T \times F \times C$, where $T = 86$ is the number of time points in the input, $F = 86$ is the number of frequency points, and C is the number of electrodes. To evaluate the effectiveness of the

proposed framework we used three evaluation methods presented in Table 6.10. Three approaches are described as follows:

1. In the first approach, the results were obtained in subject-independent manner. Where data from all but one subject was used for training the network and was tested on a different subject which is not part of the training data.

2. The second approach was also conducted in subject-independent manner. However, 90% of the data from all the subjects were used for training and 10% was used for testing in cross-validation manner. Results obtained for different training and testing sets were later averaged.

3. The third approach for calculating results was subject-dependent, the network was trained and test on the data from same subject. Results for each subject were evaluated separately. Further, for subject dependent classification, we also evaluated different values of C, i.e., number of electrodes on our proposed method. The three evaluation methods have been shown in Table 6.10.

TABLE 6.8: Three evaluation method: leave trial out (LTO) is done on subject-by-subject basis, i.e., training and testing take place using different data from the same subject; leave subject out (LSO) and subject-independent leave trial out (SI-LTO) are subject-independent experiments.

Exp	Training		Testing	
	Subjects	Trials	Subjects	Trials
LSO	All but one	All	One	All
SI-LTO	All	90%	All	10%
LTO	-	90%	-	10%

6.6.1 Subject-Independent Evaluation

The first experimental evaluation was conducted in subject-independent manner, where EEG data from all the subjects were used in recognition of imagined words. For this experiment, 108 EEG trials were available for each class. In order to perform subject-independent experiment, raw EEG data from all the subjects was z-scored, this was done in order to remove variations in EEG signals recorded from different subjects. In the subject-independent study, the network was trained using mini-batch size of 32. Recognition of imagined words in a subject independent study is difficult because EEG signals vary from subject-to-subject. Therefore, the subject independent evaluation was conducted in two different approaches, discussed as follows:

Leave Subject Out Cross Validation (LSO)

In this evaluation technique, the network was trained on all but one subject and tested on EEG data from the left-out subject i.e., 11 subjects were used to train the network and tested on EEG data from a different subject. This method was repeated several times in a leave-one-subject-out (LSO) cross-validation manner. The subject independent results for three-word category are presented in Table 6.9 and average accuracy achieved for selected C is shown in Table 6.10. The proposed system was tested for different C values i.e., the number of electrodes selected using the proposed electrode selection method. The best results were obtained by $C = 15$ $C = 9$, i.e., by top C electrodes that showed most power across all trials of subjects used in training the network. From the results obtained it is evident that our proposed method can recognize imagined word even if the system is not trained on the users EEG data. The figure 6.5 shows the electrodes that showed most power across all the trials for 12 subjects.

TABLE 6.9: Classification accuracy for three word pairs with the proposed CNN-Attention network with different number of selected electrodes C using mean power method. The results are presented in subject-by-subject manner.

Subject	("Run", "Swim")			("Run", "Write")			("Swim", "Write")		
	C=15	C=9	C=6	C=15	C=9	C=6	C=15	C=9	C=6
1	65.5	71.0	69.5	45.0	45.5	53.0	58.0	61.0	55.9
2	59.5	66.4	68.5	50.0	57.5	52.0	53.5	50.0	43.5
3	60.0	58.3	56.1	62.7	60.0	55.0	61.6	61.1	56.7
4	93.8	86.6	67.2	72.7	77.8	78.8	69.5	57.8	63.3
5	81.1	76.6	76.1	71.1	67.8	66.1	61.1	57.8	66.1
6	83.8	80.0	72.2	61.1	72.2	56.6	67.8	68.3	69.4
7	50.5	51.0	50.0	56.5	63.0	57.5	53.5	58.5	54.5
8	53.7	55.6	57.5	53.1	59.4	53.7	54.4	53.2	51.8
9	50.5	50.0	58.5	50.0	50.0	50.0	50.0	50.0	57.5
10	61.0	71.5	72.0	53.0	59.0	57.0	60.0	51.0	54.0
11	88.8	76.1	76.6	72.7	80.5	73.0	68.3	61.2	55.5
12	67.5	60.0	67.5	46.0	52.0	55.5	50.0	45.0	53.5
Average	**67.9**	**66.9**	**65.9**	**57.8**	**61.9**	**59.0**	**59.0**	**56.2**	**56.8**

TABLE 6.10: Comparison of average accuracy achieved by C selected electrodes.

Word Pair	C=15	C=9	C=6
("Run", "Swim")	67.9	66.9	65.9
("Run", "Write")	57.8	61.9	59.0
("Swim", "Write")	59.0	56.2	56.8
Average	**61.5**	**61.6**	**60.5**

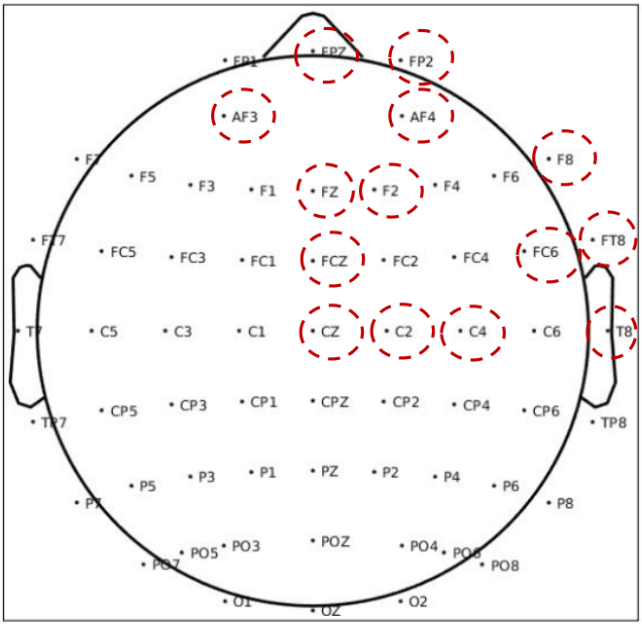


FIGURE 6.5: The electrodes that showed most power across 12 subjects. Where the electrodes showing most power across all the trials were considered most informative electrodes for recognition of imagined words.

Classification of Spectrograms of Shorter Time Frame

As it was inferred earlier in this chapter, that most task discriminative features are present few milliseconds after the stimulus onset, i.e., when the subject was asked to mentally read the presented word. Therefore, the performance of the proposed method was also evaluated on spectrograms with shorter temporal length. The dimensions of spectrogram window used in this analysis was $T \times F \times C$, where $T = 50$ is time points in the spectrogram, and $F = 86$ is the frequency points. The evaluation was only

carried out for $C = 15$ and $C = 9$. The subject-by-subject results are shown in Table 6.11 and average accuracy is shown in Table 6.12. As shown the results achieved with spectrogram representing 1000 *ms* post stimulus activity is better than spectrogram representing the whole trial. In addition, recognition rate of 63.7% achieved with $C = 15$ is better than $C = 9$, which can be attributed to each subject having different set of electrodes providing task discriminative information. In other words, a smaller set of electrodes selected from training subjects does not account for most discriminate features for word recognition in the test subject. This is caused by inter-subject variations in EEG signals. Hence, in subject-independent evaluation it is important to have large C. Therefore, the network's performance was also evaluated with C = 64,i.e., all the electrodes used during the recording. The network achieved best accuracy of 65.9% when all the electrodes were used in training and testing.

TABLE 6.11: Accuracy for three word pairs with spectrogram 1000 *ms* activity post stimulus onset. The evaluation was performed with C = 15,9. The results are presented in subject-by-subject manner.

Subject	("Run", "Swim")			("Run", "Write")			("Swim", "Write")		
	C=64	C=15	C=9	C=64	C=15	C=9	C=64	C=15	C=9
1	56.0	64.0	63.5	67.5	51.0	44.0	61.5	60.5	52.5
2	56.5	64.5	57.5	60.0	56.0	55.0	61.0	55.0	53.5
3	57.2	61.6	60.0	63.3	65.0	60.0	71.6	55.0	55.5
4	92.7	97.2	93.8	95.5	90.5	78.3	71.1	61.1	66.1
5	92.2	85.5	85.5	85.5	72.2	71.1	62.7	63.8	58.8
6	78.8	79.9	86.1	91.1	82.7	72.7	66.1	68.3	67.7
7	51.0	52.5	50.5	57.0	57.5	61.9	50.5	53.9	59.0
8	50.0	52.5	60.0	51.8	51.5	56.2	54.2	53.1	53.7
9	53.5	57.5	50.0	53.0	51.5	51.0	51.0	50.5	50.0
10	53.9	55.0	67.5	63.5	63.5	61.5	64.5	60.0	47.5
11	86.6	90.5	91.6	83.3	85.6	77.7	65.0	66.1	55.6
12	69.0	65.5	69.0	69.5	45.5	50.5	57.5	51.4	44.5
Average	**66.5**	**68.9**	**69.5**	**70.0**	**64.4**	**61.4**	**61.4**	**58.2**	**56.4**

TABLE 6.12: Average accuracy for three word pairs with spectrogram 1000 *ms* activity post stimulus onset with C = 15,9.

Word Pair	C=64	C=32	C=9
("Run", "Swim")	66.5	68.9	69.5
("Run", "Write")	70.0	64.4	61.4
("Swim", "Write")	61.4	58.2	56.4
Average	**65.9**	**63.6**	**62.4**

6.6.2 Subject-Independent Leave Trial Out (SI-LTO)

In this experiment, the network was trained on $N-1$ trails from each subject and tested on the N^{th} trial, where N is the total number of trials for a given subject. This was performed in a leave-trial-out from each subject manner, which can be referred to subject independent leave trial out cross validation (SI-LTO). This method of evaluation is easier in comparison to (LSO) performed earlier which is evident from the results, shown in Table 6.13. As shown, $C = 9$ performed best in recognition of imagined words with an accuracy of 72.9% for this subject independent evaluation. Despite the fact that the dataset contained EEG signals from multiple subjects, the proposed method achieved recognition rate of 72.9% for the imagined words.

6.6.3 Leave Trial Out (LTO) Evaluation

The third experiment focused on subject-dependent classification, only trials from one subject could be used for training and testing of the system. The proposed method was tested for the (two-class)

TABLE 6.13: Recognition of imagined word in SI-LTO manner with different number of electrodes (C).

Word Pair	C=64	C=32	C=9
("Run", "Swim")	77.0	75.8	78.0
("Run", "Write")	72.9	77.7	72.6
("Swim", "Write")	66.3	63.0	68.8
Average	**72.0**	**72.1**	**72.9**

TABLE 6.14: Classification accuracy for three word pairs with different number of electrode C selected using the proposed electrode selection method. The evaluation was performed in subject dependent manner.

Subject	("Run", "Swim")				("Run", "Write")				("Swim", "Write")			
	C=64	C=15	C=9	C=6	C=64	C=15	C=9	C=6	C=64	C=15	C=9	C=6
1	52.0	78.0	83.0	85.0	54.0	65.0	80.0	71.0	72.0	81.0	90.0	89.0
2	76.0	90.0	91.0	92.0	68.0	77.0	83.0	83.0	53.0	88.0	84.0	82.0
3	68.0	67.0	60.0	57.0	66.0	77.0	80.0	79.0	74.0	74.0	88.0	83.0
4	99.0	100	100	100	100	100	99.0	100	74	78.0	70.0	76.0
5	84.0	73.0	76.0	73.0	68.0	69.0	61.0	72.0	60.0	55.0	68.0	63.0
6	77.0	68.0	57.0	60.0	84.0	83.0	81.0	87.0	60.0	61.0	42.0	60.0
7	47.0	51.0	63.0	57.0	53.0	68.0	81.0	90.0	45.0	41.0	53.0	43.0
8	89.0	86.0	77.0	79.0	86.0	66.0	61.0	51.0	85.0	87.0	85.0	76.0
9	58.0	52.0	53.0	81.0	68.0	53.0	64.0	59.0	34.0	55.0	69.0	56.0
10	98.0	72.0	76.0	81.0	95.0	90.0	90.0	83.0	74.0	76.0	84.0	70.0
11	71.0	93.0	88.0	87.0	97.0	93.0	95.0	85.0	82.0	68.0	57.0	62.0
12	80.0	80.0	87.0	37.0	83.0	91.0	88.0	83.0	66.0	85.0	52.0	53.0
Average	**74.9**	**75.8**	**75.9**	**74.0**	**76.4**	**77.6**	**80.2**	**78.5**	**64.9**	**70.7**	**70.2**	**67.5**

TABLE 6.15: Comparison of average accuracy achieved by different number of selected electrodes C.

Word Pair	C=64	C=15	C=9	C=6
("Run", "Swim")	74.9	75.8	75.9	74.0
("Run", "Write")	79.5	77.6	80.2	78.5
("Swim", "Write")	64.9	70.7	70.2	67.5
Average	**73.1**	**74.6**	**75.4**	**73.3**

TABLE 6.16: Accuracy for three word pairs with spectrogram 1000ms activity post stimulus onset. The evaluation was performed with C = 15 and C = 9

Subject	("Run", "Swim")		("Run", "Write")		("Swim", "Write")	
	C=15	C=9	C=15	C=9	C=15	C=9
1	78.0	77.0	65.0	79.0	81.0	88.0
2	90.0	91.0	77.0	82.0	88.0	84.0
3	67.0	65.0	77.0	74.0	74.0	87.0
4	100	100	100	100	78.0	72.0
5	73.0	84.0	69.0	76.0	55.0	68.0
6	68.0	62.0	83.0	89.0	61.0	57.0
7	51.0	62.0	68.0	84.0	41.0	45.0
8	86.0	84.0	66.0	55.0	87.0	85.0
9	52.0	52.0	53.0	69.0	55.0	71.0
10	72.0	64.0	90.0	85.0	76.0	84.0
11	93.0	98.0	93.0	97.0	68.0	61.0
12	80.0	85.0	91.0	86.0	85.0	56.0
Average	**75.8**	**77.0**	**70.7**	**81.3**	**77.6**	**71.5**

classification of imagined words within pairs formed from the above three words. As the three words can be combined in three pairs, three sets of results were obtained from 12 subjects. In general, 10 trials for each class were recorded from each subject. However, after artifact rejection some subjects had only 9 trials. Results for each subject were obtained using leave-one-out cross validation, where the dataset

was divided into 90% training and 10% testing data. The model was evaluated by varying the training and test data, and the respective classification results were averaged for each subject. In order to evaluate effectiveness of the proposed electrodes selection method, we tested our system for different values of C. Further, these results were compared with the accuracy achieved by the network when trained and tested using all the electrodes $C = 64$. The results are presented in Table 6.14 and average accuracy is shown in Table 6.15. As shown, the accuracy achieved by using electrodes selected by the proposed method is higher compared to when the system is trained and tested using all electrodes. Therefore, it can be concluded that the proposed electrode selection method chose electrodes that provides discriminative information for recognition of imagined words. High recognition rate achieved in subject dependent manner can be attributed to each subject have different set of electrodes containing task discriminative information.

TABLE 6.17: Comparison of average accuracy achieved by selected electrodes C.

Word Pair	C=15	C=9
("Run", "Swim")	75.8	77.0
("Run", "Write")	70.7	81.3
("Swim", "Write")	77.6	77.6
Average	**74.7**	**76.6**

Similar to the previous evaluation in LSO, we tested our method for spectrogram representing 1000 *ms* activity after stimulus onset. This evaluation was only performed for $C = 9$ and $C = 15$. The subject-by-subject results are shown in Table 6.16 and average accuracy is shown in Table 6.17. Similar to subject-independent classification the results are better with spectrogram representing 1000*ms* post stimulus activity.

Attention Weights Visualization

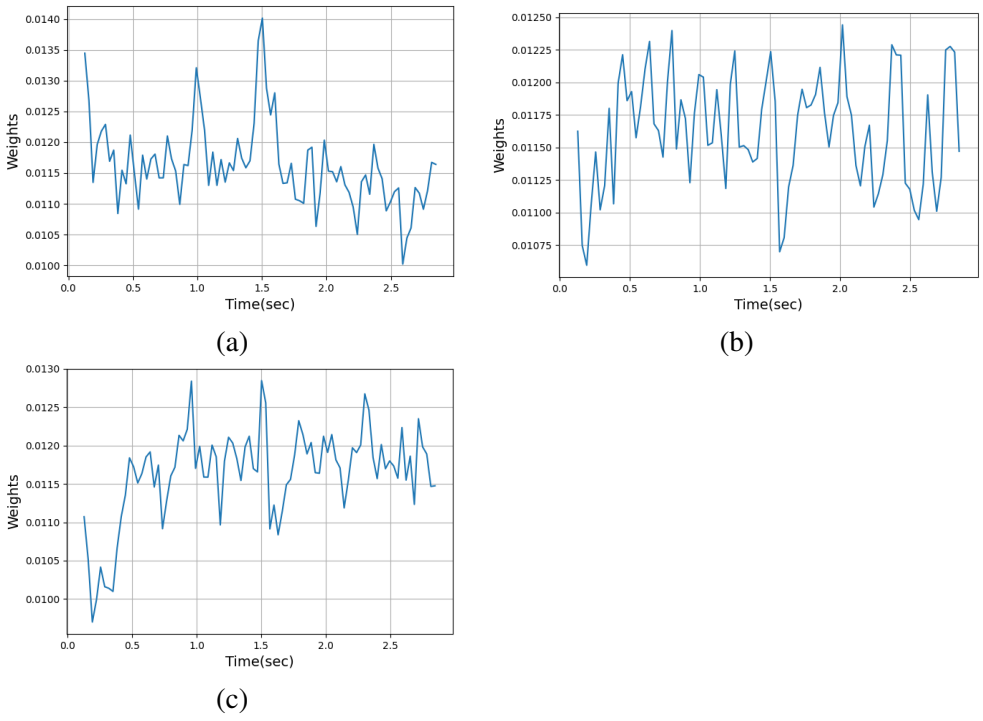


FIGURE 6.6: Global attention weights for three imagined words: (a) *Run*; (b) *Swim*; (c) *Write*.

The self-attention mechanism used in the network can provide with the time points containing important features extracted by the CNN blocks. Features that are more task discriminative are assigned higher weights by the self-attention layer. We averaged attention weights from the networks trained in subject dependent manner (leave-one-out cross validation) for C=9, which provided the best results. Further, the attention weights across all the subjects were averaged together to create the global attention weights. The global attention weights are shown in figure 6.6. The attention weights for imagined words "*Run*" and "*Write*" shows time points around $1s$ & $1.5s$ (i.e., $500ms$ & $1000ms$ after the stimulus onset) to contain the most discriminative features, figure 6.6 (a) and (C). The attention weights for the word "*Write*" contains several time points with high weights. On the other hand, imagined word "*Swim*" contains multiple time points providing important features, figure 6.6 (b). As can be seen from figure 6.6 the attention weights at the pre-stimulus time period are not the same this can be associated with trial-to-trial fluctuation in subjects concentration leading to new features (Macdonald et al., 2011). However, the proposed network is designed to extract features which associated with imagined words; therefore, highest weights are given to the time period post stimulus onset when the word was spoken silently.

6.6.4 Comparison of the Proposed Network with the Baseline

TABLE 6.18: Comparison of average accuracy achieved by the proposed network with the baseline networks in three experimental protocols; SI: subject-independent, SD: subject-dependent, and SI-LTO: subject-independent leave trial out. CLA: CNN-LSTM-Attention; CWA: CNN without Attention.

Exp	SI			SI-LTO			SD		
	C=64	C = 32	C=9	C=64	C = 32	C=9	C=64	C = 32	C=9
CLA	62.3	59.6	59.7	67.9	69.5	67.4	66.0	68.3	75.3
CWA	62.5	61.6	61.9	66.3	68.3	66.7	64.0	64.8	66.7
Proposed	65.9	63.6	62.4	72.0	72.1	72.9	73.1	73.7	75.4

The performance of the proposed architecture was evaluated in comparison with two baseline models: (1) CNN-LSTM-attention; and (2) CNN network without attention mechanism. This was done to emphasize the effectiveness of including the dense layers for dimensionality reduction and the self-attention mechanism for learning temporal dynamics in the proposed network. The details of the two networks are as follows:

- **CNN-LSTM-Attention:** We added bidirectional LSTM as it is known to be effective in learning temporal dynamics and when combined with attention mechanism has produced state-of-the-art performance in many recognition tasks (Singh et al., 2020; Zhou et al., 2016b). The CNN-LSTM-attention network had a similar architecture as the proposed CNN-attention network shown in figure 6.4, except it did not have parallel dense layer after the CNN blocks. Therefore, features extracted by the CNN were passed to the LSTM for temporal learning. The bidirectional LSTM layer had 128 units and output from the LSTM was passed to the self-attention layer to learn important temporal features.

- **CNN without Attention:** In order to show the effectiveness of the attention mechanism in conjunction with the convolutional blocks, we also evaluated tested a baseline network without attention mechanism, with the remaining network architecture similar to the proposed network in figure 6.4.

All the networks were evaluated using the three experimental protocols; SD, SI, and SI-LTO manner with different C as the number of selected electrodes. The network in a similar manner as to the proposed network. The average accuracy for 12 subjects is presented in Table 6.18. As shown the performance of the proposed network is superior to the other two networks. Further, the CNN-LSTM-attention network had 1600835 trainable parameters and required $35s$ for training. On the other hand, CNN without

attention had 1625026 trainable parameters and required 26*s* for training. However, the proposed CNN-attention network had 822467 trainable parameters and the training time of 20*s*, in subject dependent manner. Similarly, the number of parameters and the training time was less for the proposed network in SI & SI-LTO method. In addition, comparison suggests that proposed network learns temporal features without the LSTM layer.

Comparison with Previous Work

TABLE 6.19: Comparison of performance achieved by existing methods with the proposed method on our EEG dataset.

Method	*C*	SD	SI	SI-LTO
Sereshkeh et al., 2017	64	50.1%	66.9%	72.0%
Panachakel et al., 2019	64	55.5%	50.5%	50.6%
Proposed	64	73.1%	65.9%	72.0%
Proposed	9	76.6%	62.4%	72.9%

Results obtained using the proposed method were compared with existing methods. All the methods were assessed on our EEG dataset for the three word pairs and different experimental protocol. The following methodologies were evaluated in this work:

- In (Panachakel et al., 2019) representations were extracted from temporal and wavelet domain, and classification was performed individually on each channel using a multi-layer perception followed by hard voting to get the final result.
- (Sereshkeh et al., 2017) decomposed EEG signals into several levels using the discrete wavelet transform (DWT). Features such as standard deviation (SD) and root mean square (RMS) were calculated from all electrodes, which were combined into a vector. Further, these vectors were fed into a regularised neural network for classification of imagined word.

As can be seen from the results in Table 6.19, the proposed method outperformed the existing methods in recognition of imagined words from EEG signals in SD and SI-LTO manner; however, the performance (Sereshkeh et al., 2017) is better for SI evaluation.

6.6.5 Comparison of Electrode Selection Method with the State-of-the-art Optimization Methods

TABLE 6.20: Evaluation of performance achieved by the proposed electrode selection technique in comparison with the state-of-the-art optimization algorithms for selecting electrodes to recognize of imagined words. The three methods were evaluated for in the following manner; SD: Subject-Dependent; SI: Subject-Independent; and SI-LTO: Subject-Independent Leave-one-out.

Method	Accuracy			Processing Time		
	SI	SI-LTO	SD	SI	SI-LTO	SD
PSO	64.5	70.6	70.4	3.6 min	3.6 min	2.6 min
GA	63.7	69.2	73.2	2.4 min	2.4 min	1.1 min
Proposed	62.4	72.9	76.6	40 sec	41 sec	8 sec

We also compared the proposed electrode selection method to state-of-the-art optimization techniques such as genetic algorithms (GA) and particle swarm optimization (PSO). GA and PSO are both meta-heuristic algorithms that have been demonstrated to be successful in tackling complicated engineering optimization issues (Konak et al., 2006). As a result, we employed GA and PSO to choose electrodes that were then used to train a convolutional attention network for imagined word recognition.

We chose a population size of 500 for the GA which converged after the $20th$ generation, as used in (Albasri et al., 2019). The fitness metric was calculated using a logistic regression classifier. For the PSO, we employed a swarm of 20 particles with an inertial mass of 0.3. The objective function served as a proxy for accuracy when the logistic regression classifier was used, and the algorithm completed 500 iterations. The GA and PSO both were fed a two-dimensional matrix $X \in \mathbb{R}^{n \times C}$, where n represents the number of training trials and $C = 64$ represents the number of electrodes. The input was determined by averaging spectrograms for each electrode across the time and frequency axes. As shown in Table 6.20, our electrode selection approach outperformed the PSO and GA optimization techniques in SI-LTO and SD evaluation manner. Furthermore, our technique is significantly faster, requiring less processing time, which makes it suitable for BCI applications. However, PSO method outperform the proposed method in SI evaluation, proving to be more effective in group-level analysis.

6.6.6 Comparison with Chapter 4 and 5

TABLE 6.21: Comparison between the work in Chapter 6 and Chapter 4, 5

Chapter	Task	Methods	Results
4	Classification was performed between imagined speech and two other speech related activities: visual imagery and overt speech.	K-NN for electrode selection, CNN-CBAM network for feature learning and classification.	CV:82.9%, CO: 77.7%
5	Recognition of covertly spoken words.	LDA, CSP, and statistical features for feature extraction and SVM, K-NN and D-Tree for classification. Distance measure and classification using DTW and electrode Fusion.	SVM:53.7%, K-NN:53.4, DT:52.0%; DTW-Fusion: 67.4% (50-50 split), 60.3% (leave-one-out)
6	Classification was performed between covert speech (words) and non-speech activities: visual perception and resting state from EEG signals.	CNN network with dense layers.	CVP: 83.6%, CRS:87.6%
6	Recognition of covertly spoken words.	Top$K$ electrode selection, CNN-Attention.	SI: 63.6%, SD: 75.4%, SI-LTO: 72.9%

The tasks performed in Chapter 4, 5, and 6 involved recognition of covertly spoken words, and recognition of covertly spoken words from other activities using EEG signals. The work in Chapter 5 and second part of chapter 6 are directly comparable, whereas work in chapter 4 and 5 are not. Table 6.21 summarizes the comparison between work in these three chapters.

6.6.7 Limitations

The proposed electrode selection method performed well when used in subject dependent and subject independent-LTO manner. The proposed method in combination with CNN-attention network achieved, 72.7% for SI-LTO and 75.4% for subject dependent and SI-LTO task. However, for subject independent LSO evaluation, the electrode selection method failed to achieved high recognition rate. Although, the performance achieved 63.9% with C=15 is comparable, was not better than recognition rate of 65.9% with C=64. Therefore, it can be concluded that the electrode selection method is not robust in feature selection from a subject which is not part of the training data. In SI evaluation, the proposed CNN-attention network suffered from overfitting, as shown in figure therefore the network's performance was lower compared to SI-LTO and SD evaluations. Therefore, future work would include using early stopping for training the network.

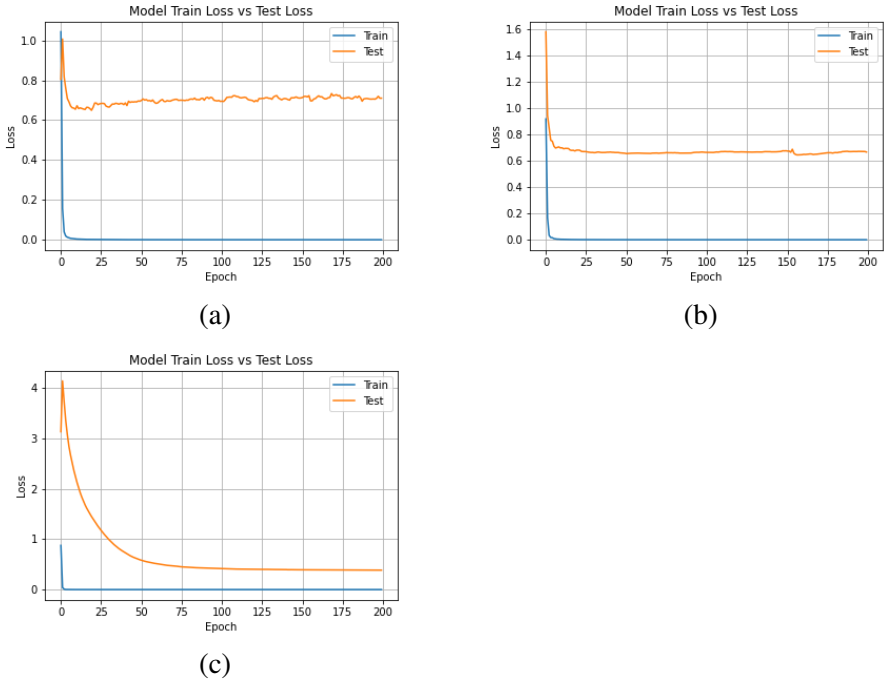


FIGURE 6.7: Training (blue) and testing (orange) loss curves indicating towards overfitting in (a) SI evaluation, where as network fits well in (b) SI-LTO ; (c) SD evaluation.

6.7 Conclusion

The study discussed in this chapter examined the feasibility of recognition between imagined words using the deep learning methods. Also, in this chapter classification was performed between imagined speech and two types of non-speech tasks, which were either visual perception or resting state (no activity). In the first part of the chapter, windows of EEG signals were used as input to the CNN to learn spatio-temporal patterns for classification between imagined speech and non-speech activity. The average accuracy achieved for 12 subjects was 87.6% and 83.6%. However, the combination of spatio-temporal input and the CNN did not perform effectively in recognition of imagined words. This concludes that spatio-temporal patterns are effective in discriminating between different activities, however the method does not perform well for recognition of imagined words. Using only the CNN had a limitation, that the temporal dynamics could not be modeled in the EEG signals. Further, EEG signals used in time domain did not contain spectral information for the particular word.

To recognize the imagined word, the second part of the chapter proposed an application of CNN-attention network from spectrograms of EEG signals. In addition, to reduce the dimensionality of the EEG data, we proposed an electrode selection method that choose electrodes by calculating mean power in the spectrogram. The combination of the proposed electrode selection technique and the CNN-attention network showed its effectiveness in recognition of imagined words. Further, the event in time-frequency domain spectrogram representing 1000*msec* post-stimulus activity produced better results. The average accuracy for three word pair across 12 subjects was 76.6% in subject dependent evaluation and 63.5% for subject dependent evaluation. Comparison with baseline network also suggested that the proposed CNN-attention architecture is successful in learning temporal features and requires less training time than the CNN-LSTM-attention.

6.8 Summary

A thought-to-text brain computer interface (BCI) technology should be able to recognize imagined words from EEG signals. Further, a practical BCI system should also be able to discriminate between imagined speech and non-speech activity, such as silence and visual perception. Therefore, in this chapter classification was performed between EEG signals recorded during imagined words and two non-speech tasks, visual perception, and non-activity (silence). In order to classify between imagined speech and non-speech activity, a CNN network was used which learned spatio-temporal features from raw EEG data. The input was divided into n windows and parallel CNNs were used for features extraction, at a later stage these features were combined and fed to dense layers (multi-layer preceptron) for classification. The proposed framework outperformed previous methods which discriminated between imagined speech and non-speech activity. Although, the method performed well in recognition between imagined speech and non-speech activity, the network performance was low in recognizing imagined words. This low performance in recognizing imagined words was associated with small number of trials available in order to train the network and inter-trial variability in spatio-temporal data. To achieve better recognition rate from imagined words, we proposed optimization to the CNN network by adding self-attention mechanism and used time-frequency features as input. In addition to reduce the spatial dimension of EEG data, we proposed an electrode selection method which choose electrodes based on mean power in the EEG spectrogram. The pipeline of spectrograms with electrode selection method and proposed CNN self-attention architecture results in better recognition rate of imagined words.

Chapter 7

Grammatical Class Recognition from Imagined Speech

In the first part of this chapter, a multi-channel CNN (MC-CNN) is proposed for recognition of grammatical class from EEG signals of imagined words. The proposed framework process the input time-frequency information for different regions of the brain separately and this information is combined at a later stage to achieve better recognition of imagined speech. The method achieves better recognition rate compared to the previously proposed method in Chapter 5. The second half of the chapter describes the optimization of the proposed MC-CNN to a multi-level framework for multi-class classification of imagined words.

7.1 Introduction

A brain-to-text BCI should be able to successfully decode imagined word from a dictionary of words i.e., predicting a class from multiple classes. From chapter 5 and 6, we can conclude that even under binary condition achieving high recognition rate can be difficult. Further, achieving high performance under multi-class classification task is very challenging especially considering the non-stationary nature of EEG signals. In order to solve this problem, in this chapter we draw inspiration from language processing in the brain. Linguistic interactions have objects and properties associated with those objects, which might be lexically reflected by the grammatical classes (Crepaldi et al., 2011). At a fundamental level, grammatically objects are regarded as nouns which promote the primary concept, whereas verbs provide context to the concept (Crepaldi et al., 2011). Subsequently, it is known that processing of verbs and nouns assume hierarchical neural circuits which depends on the semantic-features of the words, where the semantic-features could belong to sub-classes (e.g., action noun or object noun) within a given grammatical class (Popp et al., 2019; Pulvermüller, 2018). Therefore, in order to design a thought-to-speech BCI which can predict mentally spoken words, it is important to design a multi-level classification method. In such a model the first level recognizes the grammatical class of the word, and the second level recognizes the word within the recognized grammatical class.

Further, processing of language in the brain is not limited to the brain Broca's and Wernicke's area, also known as the classical language regions. In fact, information flow between areas from the left inferior frontal (Broca's area) and the premotor cortex, with auditory areas in the superior temporal lobe (Wernicke's area) have been known to be a link between language and action (Pulvermüller, 2005). On the other hand, the processing of verbs is not limited to the frontal and central (motor) region of the brain. Sound-related verbs activated auditory regions (temporal lobe) of the brain and action-related nouns activates the frontal and parietal brain areas (Popp et al., 2019). Therefore, it is important to use information from multiple brain areas in order to achieve recognition of imagined speech from EEG signals. This can be done by exploiting the spatial information from the multi-dimensional EEG signals, which are recorded using multiple electrodes covering the head.

Indeed, methods have been proposed to capture spatio-temporal information from EEG signals. An example of such algorithm is the common spatial patters (CSP) (Koles et al., 1990) which have achieved

high recognition rate for motor imagery task (Ang et al., 2008). Similarly, deep learning methods have also been proposed for capturing spatial information using the CNNs (Lawhern et al., 2018; Zhang et al., 2019a). The CNNs employed for classification of motor imagery (hand movement) from EEG signals have resulted in impressive performance. However, so far these methods have been proposed for recognition of motor actions or motor imagery from EEG signals. These methods will likely not be effective in recognition of mentally spoken words, which do not have defined events (event related synchronizes (ERS) and event related de-synchronizes (ERD) (Martin et al., 2016). Further, capturing important patterns from EEG signals in time domain containing high frequencies becomes more difficult due to presence of noise, even after pre-processing.

Therefore, this chapter propose a framework using the multi-channel CNN (MC-CNN) method for recognition of grammatical class from EEG signals of imagined (covert) speech. The MC-CNN was used in the previous studies for text recognition and has been used for multi-variate time series classification (Kim, 2014; Zheng et al., 2014). However, unlike previous models which were aimed at processing vectors, the MC-CNN proposed in this chapter is designed for extracting features from spectro-temporal information from electrodes groups. Thorough evaluation of our model on two EEG dataset containing imagined words from two grammatical classes (nouns and verbs), show that our proposed method achieves the state-of-the-art recognition under binary classification task. The idea of noun and verb classification is extended to a multi-level classification framework for recognition of imagined words in multi-class setting. At the first level (level-1), grammatical class of the given word is recognized. At the next stage (level-2), multi-class classification is performed to find the input word within the grammatical class selected at level-1. Evaluation of the proposed multi-level method shows that it outperforms the standard method, where the network was trained and tested on all the 10 classes at the same time. Further, we also compared our results with the state-of-the-art methods for multi-class classification of imagined words. The following are the contributions of this chapter:

- A multi-channel CNN (MC-CNN) framework for recognition of grammatical class of the imagined words from EEG signals. To the best of our knowledge, our work with nouns and verbs is the first to distinguish between two grammatical classes using machine learning techniques.
- Multi-level framework for classification of imagined words from EEG signals under multi-class classification task.
- Evaluation of the proposed method on two EEG datasets.

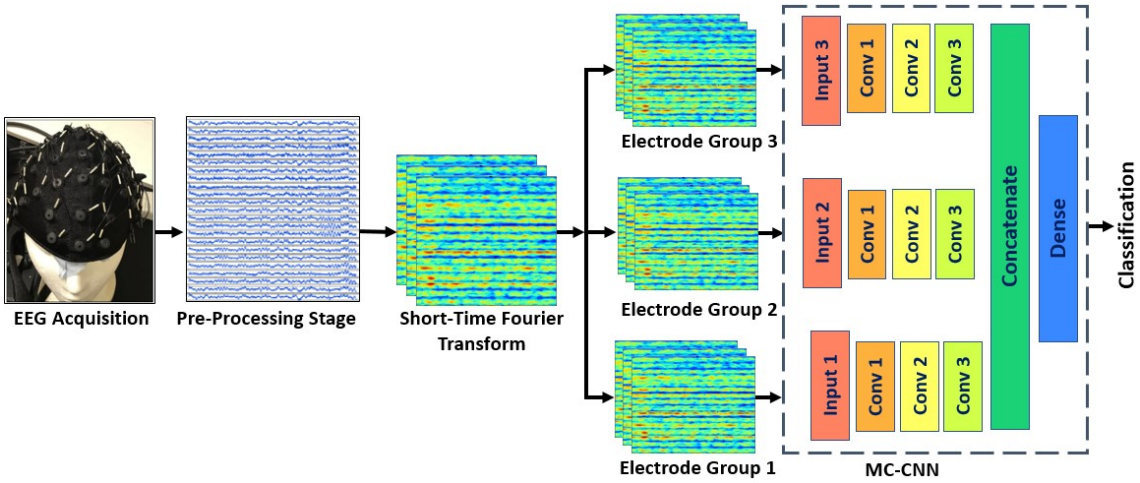


FIGURE 7.1: The proposed method for recognition of grammatical class from EEG signals acquired during imagined speech.

7.2 EEG Dataset

For the purpose of distinguish between mentally spoken nouns and verbs, two EEG database were used, our *Imagined Speech* dataset recorded for this research and the publicly available *Kara-one* EEG dataset (Zhao & Rudzicz, 2015). Details of both the datasets are provided below.

7.2.1 Imagined Speech Database

EEG signals for mentally spoken words were recorded from 12 participants, fluent in English language. The subjects were presented with 10 words, five nouns and five verbs in a random order. This was done to avoid temporal effects (Porbadnigk et al., 2009). A subject performed 10 trials for each word, in total 100 trials were recorded from each subject. Length of a trial was $3sec$ recorded at sampling rate of $1000Hz$ resulting in an EEG signal of 3000 samples per electrode. The Details of experimental protocol and trial recording are described in section 3.2. In order to have equal number of EEG signals for nouns and verbs word class, in the experiment we used ten words (stimulus), with each grammatical class having five words. The words used are presented in Table 7.1.

TABLE 7.1: Words presented as stimulus during recording of EEG signals for covert speech.

Noun:	"*Apple*"	"*Bottle*"	"*Football*"	"*Laptop*"	"*Orange*"
Verb:	"*Carry*"	"*Run*"	"*Swim*"	"*Laugh*"	"*Write*"

7.2.2 Kara-One Database

Validation of the proposed system was also conducted on a publicly available EEG dataset Kara-one (Zhao & Rudzicz, 2015). EEG signals were acquired using the 64 electrode Neuroscan Quik cap, with electrodes placed in the 10-20 system at a sampling rate of $1000Hz$. Subjects were presented with 11 prompts (stimuli) with 7 phoneme/syllables (*/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/*) and 4 words (*pat, pot, gnaw, and knew*). For each word and phoneme/syllable 12 trials were recorded. EEG dataset was recorded from 12 participants, however data from four participants was discarded due to corrupted signals. This chapter used EEG signals recorded for imagined speech task from 8 subjects. Detailed information of the database can be found in 3.3. Words presented as stimulus to record the EEG signals are presented in Table 7.2.

TABLE 7.2: Words presented as stimulus during recording of EEG signals for covert speech in (Zhao & Rudzicz, 2015).

Noun:	"*Pat*"	"*Pot*"
Verb:	"*Gnaw*"	"*Knew*"

7.3 ERP Associated with Nouns and Verbs

The Event Related Potential (ERP) for two speech parts were also investigated, i.e., nouns and verbs. Based on the literature, EEG signals from three brain region were investigated the frontal lobe, Broca's & Wernicke's area, and Occipital & Parietal lobe. However, ERP from Broca's & Wernicke's area were found to be most informative (Table 7.3). ERP was estimated by averaging trials for the same class (nouns and verbs) over all the subjects. The average ERP is shown in figure 7.2. There were four main events of interest in the ERP. The first event was a positive peak between 0.70-0.80s post stimulus. The second event was a negative peak around 0.110s which is known to be the evoked response ($N100$) to

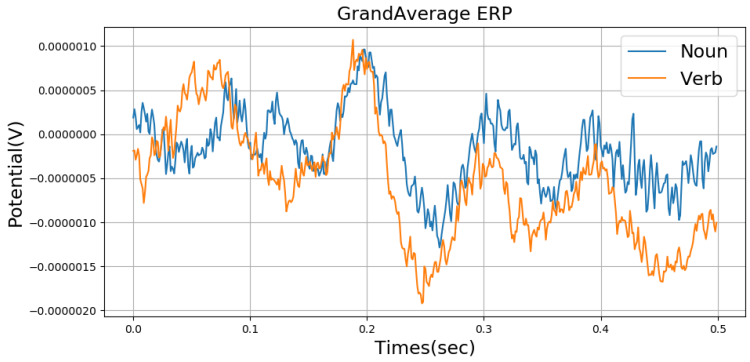


FIGURE 7.2: Event related potential (ERP) for nouns and verbs estimated from the 12 subjects. Four main components are observed, negative deflection between 0.2-0.3*sec* is the ERP component associated with processing of nouns and verbs in the brain. Image taken from (Datta & Boulgouris, 2021).

a stimulus (Sur & Sinha, 2009). The two events $P70$ and $N100$ are results of early processing of the visual stimulus presented (Preissl et al., 1995). The third event, at $0.200s$, was a very strong positive deflection known as $P200$, which is known to reflect the sensation-seeking behavior in humans (Sur & Sinha, 2009). The fourth event was a negative deflection around $0.250s$ and was estimated to be produced in response to imagined nouns and verbs. This amplitude of the event was higher for verbs in comparison to nouns which is in agreement with past studies (Preissl et al., 1995; Pulvermüller et al., 1999; Tsigka et al., 2014), where similar temporal event (ERP) was observed. This validates the presence of distinct activity produced during processing of mentally spoken nouns and verbs. These events were observed within $0.500s$ of the stimulus onset. Therefore, only ERPs from that time range are presented.

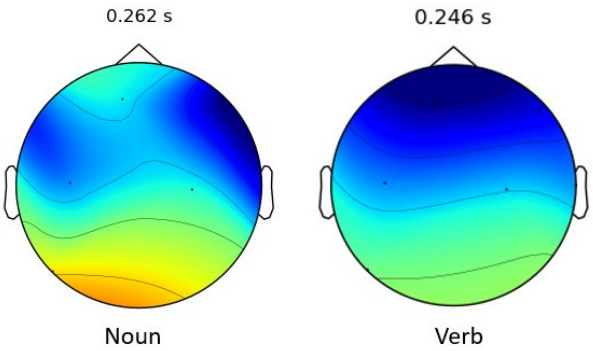


FIGURE 7.3: Topographical map for the two grammatical classes.

Along with the ERP, topographical maps are shown in figure 7.3. The topographical maps show that covertly spoken nouns resulted in reduced power at the temporal regions of the head and slightly increased power at the Occipital area. In contrast, the processing of the covertly spoken verbs resulted in reduced power in the frontal lobe. A similar observation was made by (Khader & Rösler, 2004). Association of noun and verb processing with different brain areas is in agreement with previous studies (Damasio & Tranel, 1993). As can be seen in figure 7.2, the ERPs do not provide any discriminatory information about the two grammatical classes. However, the topographical maps (figure 7.3) indicate processing of nouns and verbs at different areas of the head.

7.4 Multi-Channel CNN for Combining Information from Different Regions of the Brain

The electrode groups used in this chapter are the one found in section 5.8.2 (chapter 4). As the brain areas covered by these electrode groups are known to play important roles in the processing of nouns and verbs (Damasio & Tranel, 1993; Khader & Rösler, 2004; Popp et al., 2019). However, the number of electrodes has been reduced in two electrode groups: electrode group 1 and electrode group 2. This was done because the network performed better with fewer electrodes in these two groups.

TABLE 7.3: Groups of electrodes used as input channels to the MC-CNN network.

Electrode Group 1: F5, F3, FC3, FC5, C5, CP5, P1
Electrode Group 2: F1, FZ, F2, F4, F6, F7, FCZ, FC2, FPZ, FP2, AF3, AF4
Electrode Group 3: P1, PZ, P2, P4, POZ, PO4, PO6

7.4.1 Model Architecture

This section describes the proposed MC-CNN, designed to recognize grammatical class of mentally spoken words from EEG signals. The proposed MC-CNN is different from the ones used in past studies for natural language processing and time-series classification (Kim, 2014; Zheng et al., 2014). Instead, the proposed network was designed to learn task-specific events in spectrograms created from EEG signals and to extract information from spatial, temporal, and spectral domain. The proposed network uses the CNNs ability to learn abstract non-linear features which can adapt to the inter-trial variations in EEG signals (Bashivan et al., 2015). The filters in the CNN can help the network learn amplitude modulation patterns in the spectrograms.

Figure 7.4 shows the proposed network architecture. The network contains three channels, where input to each channel is a three-dimensional tensor. Input dimensions for each channel are $T \times F \times C$, where T refers to the number of time points, F denotes the number of frequency points, and C is the number of electrodes. The parameter C depends on the size of an electrode group i.e., the brain area, detailed in Table 7.3. In this study each channel's input was processed separately to learn important feature maps for the particular brain area and at the end of three blocks the resultant feature maps were flattened into vectors. The flattened vectors from the three channels were then combined (concatenated) and used as input to the fully connected layer.

Each channel has sequence of the convolutional blocks, and the network has three fully connected layers. The architecture of all three channels was the same. However, the input from channel 2 contained spectrograms from 12 electrodes, whereas inputs to channels 1 and 3 were comprised by spectrograms from 7 electrodes. Each block contained a single convolutional layer. The Convolutional layer in each block used small receptive field of size 3×3 that can capture local features from the spectrogram and down-sampling was performed using stride of size 2×2. The first block filtered the data with 32 kernels, whereas second and third block used 64 and 128 kernels. These layers learned hierarchical features essential for class discrimination.

The performance of the network was assessed by varying the receptive field size, among which the 3×3 filter achieved the best performance. The ability of the CNN to extract features from different time-frequency patches in the EEG spectrogram was particularly useful, as different feature maps can represent activity at different time-frequency windows. The value of the feature on the i^{th} row and j^{th} column of the k^{th} feature map at a given layer is obtained using 2.1, with $k = 1, 2, ..., K$, K = 32, 64 or 128. For the convolution operation, zero padding was used in order to preserve the spatial resolution of the input.

However, the sigmoid function was not used in the hidden layers because it suffers from vanishing gradient problem when used in the deeper layers (Goodfellow et al., 2016). To alleviate the vanishing gradient problem in the network, we used the *ELU* activation function over the more popular choice the

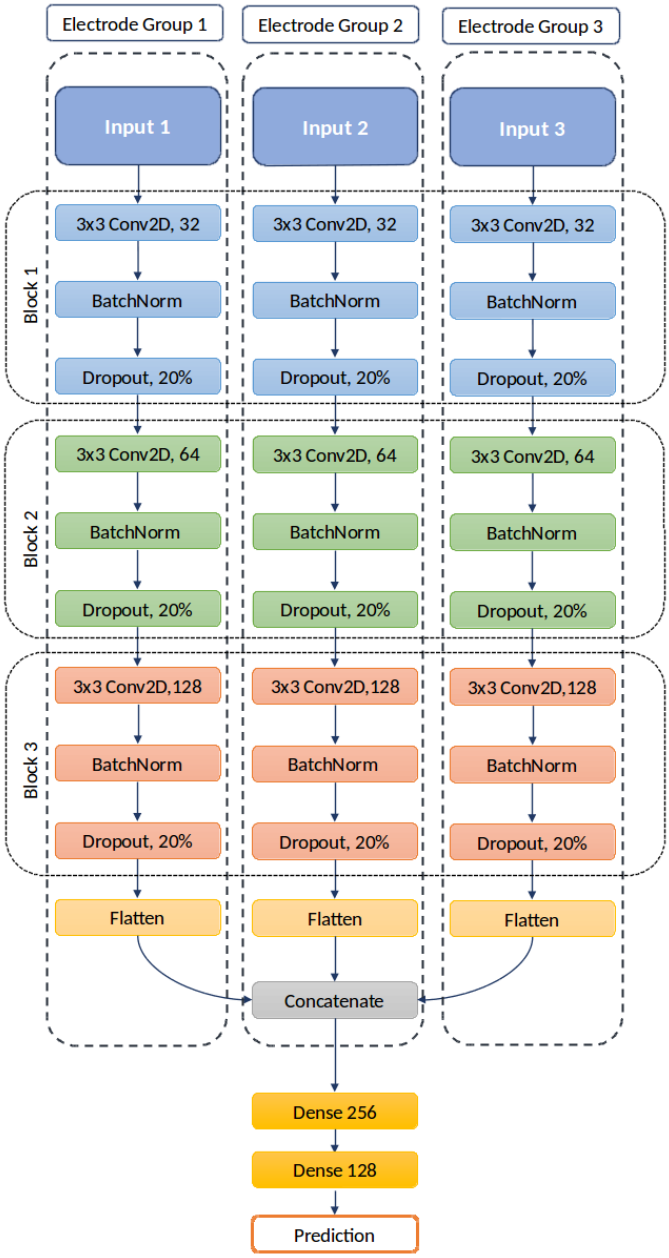


FIGURE 7.4: The architecture of the proposed MC-CNN. Each channel contains sequences of the three blocks, where each block has convolutional, batch-normalization (Ioffe & Szegedy, 2015) and dropout layers. The number of convolutional filters in each block varied.

ReLU activation function (Clevert et al., 2015; Lu et al., 2019). The *ELU* was considered a better choice over the *ReLU* function because the latter does not perform optimally when placed after the sigmoid function (Nikhil, 2018). Therefore, the rest of the network used the *ELU* activation function (Clevert et al., 2015), which endows the network with the ability to learn non-linear features.

Batch-normalization was used in the network for regularization purposes and to help speed training of the network (Clevert et al., 2015; Ioffe & Szegedy, 2015). In the fully connected layers, the number of nodes were 256, 128, and 2, with the last dense layer (classification layer) containing the sigmoid function (2.5) for binary classification. The dense layers were endowed with the *ELU* function to make the network capable of learning non-linearities at high-level features.

Hyper-parameters of the MC-CNN architecture were set after various experimentation, by varying

the activation functions, the kernel size, the filter size, and the number of input channels in the network (i.e., using more electrodes from different areas). In other words, more than three electrode groups were used to feed information to the MC-CNN through separate channels. This was done to evaluate the model performance with information from more than three brain regions. Another approach was to add more electrodes to electrode groups 2 & 3, i.e., input channel 2 & 3 of the MC-CNN network, covering more areas of the head at the Frontal and Parieto-Occipital lobe. However, both the cases resulted in reduced recognition rate. This shows the importance of electrodes from the brain areas that play an important role in the processing of nouns and verbs.

7.4.2 Network Training Parameters

The Adam optimizer (Kingma & Ba, 2014) was used for weight optimization with learning rate of 0.0001 to minimize the cross-entropy loss. In order avoid varying gradient at deeper layers in the network, a slower learning rate was selected (Bashivan et al., 2015). The network was trained for 500 iterations (epochs), with a mini-batch size of 64 in case of SI study and batch gradient descent for SD study (i.e., the batch consisted of all the samples in the training data). Over fitting was avoided, by using the batch-normalization and dropout regularization (Srivastava et al., 2014) with dropout rate of 20%.

7.5 Recognition of Grammatical Classes

In order to distinguish between mentally spoken nouns and verbs, we used EEG signals produced during imagined speech task for ten words, i.e., five nouns and five verbs. In general, 50 trials for each class (noun, verb) were recorded from each subject. However, some subjects had only 45 trials left after artifact rejection and reduction.

Three different evaluation protocols were used to validate the effectiveness of the proposed network. That is, two evaluation protocols were conducted in a subject-dependent manner, i.e., evaluation of the network's performance was done separately for each subject. On the other hand, a third experiment was performed, in which the network's performance was evaluated in a subject-independent manner, i.e., the network was trained on data from $N-1$ subjects and tested on the data from left out subject. Where, N is the total number of subjects. The three experimental protocols are summarized in Table 7.4.

TABLE 7.4: Three experimental protocols: leave one subject out (LSO) is a subject-independent experiment; leave trial out (LTO) and leave one word out (LWO) are done on subject-by-subject basis, i.e., the training and testing took place using different data from the same subject.

Exp	Training			Testing		
	Subjects	Trials	Words	Subjects	Trials	Words
LSO	All but one	All	All	one	All	All
LTO	-	80%	All	-	20%	All
LWO	-	All	All but one	-	All	All but one

7.5.1 Leave One Subject Out (LSO)

The first experimental approach assessed our network's performance in a subject-independent manner, by training the MC-CNN on EEG data from 11 participants and testing it on EEG data from a different subject. The network was trained for 100 epochs using mini-batch gradient descent with a size of 64. The recognition rate for determining whether an EEG signal corresponds to an imagined noun or verb are summarized in Table 7.5. As can be observed, the average classification rate is 80.6%, indicating that our system is capable of accurately classifying EEG data from subjects not included in training the network.

TABLE 7.5: Classification accuracy for EEG signals recorded during imagined speech of Nouns and Verbs. Results for three experimental protocols are shown.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
LSO	85.4	62.2	87.8	69.2	88.0	85.5	84.1	90.5	74.8	87.0	91.3	62.1	**80.6**
LTO	93	91.6	82.6	73.9	86.5	77.8	78.0	94.3	80.0	91.9	81.8	89.0	**85.0**
LWO	91.2	94.2	88.3	84.5	89.3	72.7	80.6	94.6	84.0	93.4	87.1	88.0	**87.3**

7.5.2 Leave Trial Out (LTO)

In the LTO evaluation protocol, 80% of the trials were used to train the network, whereas 20% were used for testing. For both training and testing, EEG data was used from all imagined words. Training and testing were conducted in a subject-dependent manner (subject-by-subject basis), and the outcomes of all the test trials were averaged for each subject. In order to eliminate variations caused in the network parameters by the stochastic nature of learning algorithm (Brownlee, 2016), the network was trained and tested ten times for each trial. The outcomes of the tests were averaged. The results for each subject are summarized in Table 7.5. As shown, the network achieved an average accuracy of 85.6%. This indicates that by observing a subject's EEG signals, it is possible to deduce whether the subject is thinking of a noun or a verb.

7.5.3 Leave One Word Out (LWO)

Another experimental approach for evaluating network performance involved training the MC-CNN network with EEG signals from four words, and testing it with EEG signals from the leftover (fifth) word from the noun or verb classes. In this scenario, 80% of the data was used for training and 20% for testing. This technique will hereafter be referred to as Leave one Word Out (LWO) cross validation. The network was evaluated using EEG data from each word separately, and the recognition rates were averaged for each subject. The classification results are shown in Table 7.5. As can be observed, the mean classification rate is 87.3%, indicating that the network is capable of classifying previously unseen nouns and verbs.

7.5.4 Evaluation on Kara-One Dataset

TABLE 7.6: Classification accuracy of our proposed model on nouns and verbs in Kara-one dataset (Zhao & Rudzicz, 2015).

Exp	S1	S2	S3	S4	S5	S6	S7	S8	Average
LSO	76.7	69.4	93.8	61	88.1	89.8	91.5	81.9	**81.5**
LTO	97.7	96.8	99.6	82.3	99.3	98.1	100	79.8	**94.1**

We also validated the networks performance using the publicly available Kara-one (Zhao & Rudzicz, 2015) EEG dataset of covertly spoken words.We analyzed the EEG data from eight people during imagined speech of four words, two nouns ("*Pat*", "*Pot*") and two verbs ("*Gnaw*", "*Knew*"). The raw signals were band pass filtered between $0.01Hz$ and $475Hz$, a notch filter was used to eliminate the $60Hz$ line, and other artefacts and noise were removed as detailed in section 3.4. Each class (noun, verb) received 24 trials, with 22 trials used to train the network and two trials utilized for testing. The training and testing trial sets were swapped in a Leave One Trial Out (LTO) cross validation procedure, with the results for each test trial calculated separately and then averaged for each subject. The results are shown in Table 7.6. As can be seen, our model attained an average accuracy of 94.1%. Additionally, we used the Kara-one database to evaluate our network's performance in a subject-independent manner. The system was trained using EEG data from seven participants and tested using data from one (different) participant, i.e., the experiment was conducted using Leave One Subject Out (LSO) cross validation. The average accuracy was 81.5%.

7.5.5 Transfer Learning

TABLE 7.7: When the network was trained on our *Imagined Speech* database and tested on *Kara-one* database. Results were evaluated in subject dependent (SD) and subject independent (SI) manner.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	Average
SD	94	50	54	93.7	83.3	62.5	92.7	50	**72.5**
SI	61.7	76.7	71	70.4	72.5	53.1	76.5	68.1	**68.7**

In order to evaluate the robustness of the proposed network, it was trained on spectrograms from our *Noun-Verb* EEG database and was tested on *Kara One* database as a transfer learning model. The pre-trained network was adjusted on *Kara-One* dataset using fine-tune last-k (Long et al., 2015; Tajbakhsh et al., 2016), where $k = 3$ in our analysis. The weights of the network were frozen, which had learned spectro-temporal patterns in the EEG spectrograms, and only the weights of the last 3 (k=3) dense layers were trained to fine tune the model on the new database. The fully connected layers had the *ELU* activation function. The network was fine-tuned with a slow learning rate of 10^{-4} to avoid over-fitting (Goodfellow et al., 2016). The network was evaluated in two different approaches, subject dependent; where the last layers were fine-tuned and tested for each subject separately. In the second approach, the network was fine tuned on the data from one subject and tested on data from all the other subjects. The results are shown in Table 7.7, and as can be seen the network achieved good recognition rate on a new dataset. This shows the robustness of the proposed method in recognizing grammatical classes from EEG signals.

7.5.6 Comparison

Although previous research (e.g., (Bierwisch, 1999; Crepaldi et al., 2011; Schilling et al., 2020)) has examined distinct brain activity related with the processing of words of various grammatical classes, to the best of our knowledge, no previous work has performed the grammatical class (noun or verb) recognition of mentally spoken words. As a result, we compared our findings to those of previously published algorithms for binary categorization of imagined speech. A direct comparison may not be completely conclusive due to the disparate datasets which were not publicly available. In our comparison, we include the method described in (Saha & Fels, 2019), which performed classification between two long words: "*cooperate*" and "*independent*", as well as the method described in (Sereshkeh et al., 2017), which performed classification between words: "*yes*" and "*no*". Additionally, we compared our findings to the state-of-the-art method described in (Nguyen et al., 2017). As shown in Table 7.8, our method surpasses the other methods in the two-class situation, despite the fact that we employed data from 16 participants, a population that is far larger than that of previous studies, which included only a few subjects. This shows the robustness of the proposed system. Additionally, despite the fact that our method utilized several nouns and verbs, which increases intra-class variation and complicates recognition, our system obtained great results across three separate experimental protocols, with a maximum classification rate of 86.4%.

TABLE 7.8: Comparison of our method with past studies in a binary classification task.

Method	Word Length	Channels	Subjects	Trials per Subject	Accuracy
Sereshkeh et al., 2017	Short	64	12	60 per class	63%
Saha and Fels, 2019	Long	64	6	100 per class	79.9%
Nguyen et al., 2017	Short & Long	64	6	100 per class	80.1%
Proposed	Short	64	12	50 per class	**85.3%**

7.6 Multi-Level Recognition System

Multi-level classification system is inspired from the hierarchical classification method such as Random Forrest. The hierarchical CNN networks have been proposed in the past (Roy et al., 2020), however most of the previous work has been done on image recognition. Therefore, in this study we introduced a hierarchical multi-class classification model for recognition of imagined words from EEG signals. As mentioned earlier, words can be divided into grammatical classes such as nouns or verbs, where each grammatical class contains multiple words which can be considered as sub-classes. Multi-class classification was done in two stages; at the first stage (level 1), system performed recognition of grammatical class of the input EEG signal. At the second stage (level 2), the recognition was performed under multi-class setting, among the words present in the given grammatical class. The multi-level structure of the proposed method is shown in figure 7.5. Due to limited data availability, words were divided into two grammatical categories. Further, different implementations of the system were investigated, for example, using the network at level 1 for feature extraction at level 2 as a transfer learning model. In other words, network used the representations learned at level 1 in word recognition at level 2. This was done by freezing the weights and adding new fully-connected layers, which enabled the network to use grammatical class specific features at level 2. The other implementation of the network was when both the level's used separate MC-CNN networks for training and testing, i.e., the network at level 2 had independent parameters.

7.6.1 Multi-level Architecture

MC-CNN architecture is similar to the MC-CNN, however the Convolutional layer in the first block of the network used the *ELU* activation. The same network architecture is used for level 1 and level 2 classification, however at level 2 the final layer of the network has the $softmax$ activation function which represents the probability that the input belongs to noun or verb class. The network at level 2 was evaluated in two different manners, independent network at level 2 and transfer learning model at level 2.

- **Independent Network at Level 2 (INL2)**: In the first method, a new MC-CNN was trained and tested at level 2, this network was independent of MC-CNN at level 1. In other words, the networks at two levels were trained and tested separately, no parameters were shared between them. In the second method, the network trained at level 1 was used as a transfer learning model at level 2. The network was trained using the Adam optimization algorithm for 300 epochs with mini-batch gradient decedent of 64 at level 1, for recognition of grammatical class. At level 2, the network was trained for 300 epochs with batch gradient-descent of 64 and learning rate of 0.0001.

- **Transfer Learning Network at Level 2 (TLL2)**: In this implementation, the MC-CNN trained at level 1 was used as a transfer learning model at level 2. All the layers of the network at level 1 were frozen and three new dense (fully connected) layers were added on top with 256, 128 and 5 neurons. The fully connected layers with 256 and 128 neurons had the *ELU* activation function. At the classification stage, the network has the $softmax$ activation function which represents the probability that the input belongs to noun or verb class. The network was trained using the Adam optimisation algorithm for 300 epochs with mini-batch gradient decedent of 64 at level 1, for recognition of grammatical class. At level 2, the network was trained for 50 epochs with batch gradient-descent of 64 and learning rate of 10^{-4}.

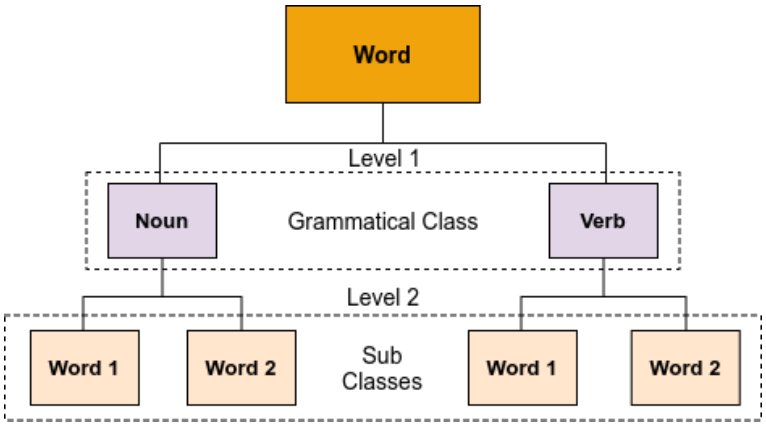


FIGURE 7.5: Proposed multi-level recognition system. At level 1, the grammatical class of the EEG signal of imagined word is recognized. At level 2, word level classification is performed from the sub-classes present in the given grammatical class.

7.7 Multi-Class Classification of Imagined Speech

In order to evaluate the proposed method for multi-class classification task, two EEG datasets were used, our own *imagined speech* data and *Kara One* data, described in section 7.2. In our *imagined speech* dataset each subject had 50 trials for each grammatical class and 10 trials for each mentally spoken word. However, after the artifact rejection some subject had only 45 trials. Similar to section 7.5, the results were obtained for both the datasets using two experimental evaluation, subject-dependent performed in leave-trial-out cross validation (LTO) manner and subject-independent performed in leave-one-subject out (LSO) manner. However, subject-dependent evaluation contained three sets of results, two set of results were obtained by using two different implementation of the multi-level system, discussed in section 7.6.1. The third set of results were obtained from a baseline method for multi-class classification.

7.7.1 Subject Dependent

TABLE 7.9: Multi-class classification accuracy of 10 imagined words from EEG signals using Multi-level recognition system. Results for two implementations of multi-level system: Independent Network at Level 2 (INL2), Transfer Learning Network at Level 2 (TLL2).

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
INL2	60.8	62.3	46.8	39.7	60.4	44.9	48.2	55.2	43.2	66.9	47.3	59.5	**52.9**
TLL2	48.4	38.5	22.2	38.8	16.4	19.6	35.9	67.2	43.3	46.7	29.4	38.9	**37.1**

In the subject dependent evaluation, two different implementations of the multi-level network were evaluated; Independent Network at Level 2 (INL2), Transfer Learning Network at Level 2 (TLL2) as described in section 7.6.1. In both the implementations, 90% of the trials were used for training the network, while 10% were used for testing, this was done in a leave-one-trial out manner. EEG data from all covertly spoken words were used for both training and testing. Training and testing took place for each subject separately and results from all the test trials were averaged together for each subject. The network was trained and tested 10 times on the same input and the results were averaged. This was done to avoid variation in network parameters, caused by the stochastic nature of the deep learning algorithms (Brownlee, 2016). Subject-by-subject results are shown in Table 7.9. As seen, the best recognition rate of 52.9% was achieved when the network is implemented in INL2 manner, with highest recognition rate of 66.9% for a single subject. The recognition rate achieved by TLL2 implementation is also above

chance level 10% for 10 classes. As best results are obtained by INL2, further experimental evaluation was conducted using INL2 implementation.

Comparison with Baseline

TABLE 7.10: Multi-class classification for 10 class, performed using single level (baseline) classification in leave-trial-out (LTO) manner.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
LTO	12.8	12.0	17.0	31.2	13.0	14.6	28.6	56.8	33.4	45.6	24.0	37.8	**27.2**

Further, the performance of the proposed multi-level classification methods was compared with the single level classification. In other words, the standard classification method where the network is trained and tested on all the classes at once. The network's performance was only evaluated in subject dependent manner (LTO). The results are shown in Table 7.10. Although, the average accuracy is above chance level, the recognition achieved in this manner is lower when compared to the proposed multi-level recognition method. Both implementations of the proposed multi-level recognition system outperformed single level classification, as can be seen from Table 7.9 and Table 7.10.

7.7.2 Subject Independent

TABLE 7.11: Multi-class classification for 10 class, performed in leave-one-subject-out (LSO) cross validation manner.

Exp	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
LSO	10.5	45.4	39.3	28.9	41.5	36.0	45.0	22.7	37.5	42.3	41.9	26.5	**34.7**

In a subject independent experimental protocol, the MC-CNN was trained on EEG spectrograms from 11 subjects and tested using EEG spectrogram from a different subject. The network was trained using INL2 implementation. The training and testing approach was repeated for all the subjects in a leave-one -subject-out (LSO) cross validation manner. At level 1, a total of 565 trials were used for grammatical class recognition and at level 2, 113 trials were available for each class for recognition of imagined word. The classification results, that is recognizing the imagined word from list of 10 words, are shown in Table 7.11. As can be seen, the average classification rate is 34.7%, with chance level of 10% for 10 classes. This shows that the proposed system can classify EEG signals under multi-class condition, from subjects that have not been used in the training of multi-level MC-CNN.

7.7.3 Kara-One Data-set

TABLE 7.12: Multi-class classification accuracy of our proposed model on 4 classes in Kara-one dataset (Zhao & Rudzicz, 2015).

Exp	S1	S2	S3	S4	S5	S6	S7	S8	Average
LSO	55.4	47.3	67.2	67	79.3	72.6	81.3	83.7	**69.2**
LTO	92.6	89.9	72.9	99	93.4	100	93.3	97	**92.2**

Further, the method was also tested on publicly available *Kara One* dataset. The dataset contains four words where two words belong to noun and two to verb grammatical class. Therefore, for multi-class classification four classes were available. Each word had 12 trials in total for one subject. The method was evaluated for 8 subjects in subject dependent and subject independent manner. The results are shown in Table 7.12.

7.7.4 Comparison

TABLE 7.13: Comparison of accuracy for multi-class word recognition with previous works. Despite more classes, our results are highest.

Method	Class Type	Classes	Channels	Subjects	No of Trials	Accuracy
Nguyen et al., 2017	Short Words	3	6	64	100 per class	50.1%
Pawar and Dhage, 2020	Short Words	4	6	64	50 per class	49.7%
Proposed	Short Words	**10**	12	64	50 per class	**52.9%**

We also made a comparison of our results with previous work that performed multi-class classification of imagined words. The comparison is performed in Table 7.13. Although a direct comparison with other studies is not always conclusive due to several factors; different data sets which were not publicly available, number of participants, number of classes used for the analysis. As shown our method is comparable to other methods with fewer classes.

7.8 Limitations of MC-CNN

The proposed MC-CNN achieved high recognition rate in predicting grammatical classes. In addition, it performed well in multi-level recognition system. However, MC-CNN suffer from overfitting in LSO and LWO evaluation method, as can be seen in figure 7.6. The learning curve for LTO evaluation method suggest slight overfitting, whereas LSO method with loss increases with number of epochs. Therefore, the future work will involve using methods such as early stopping.

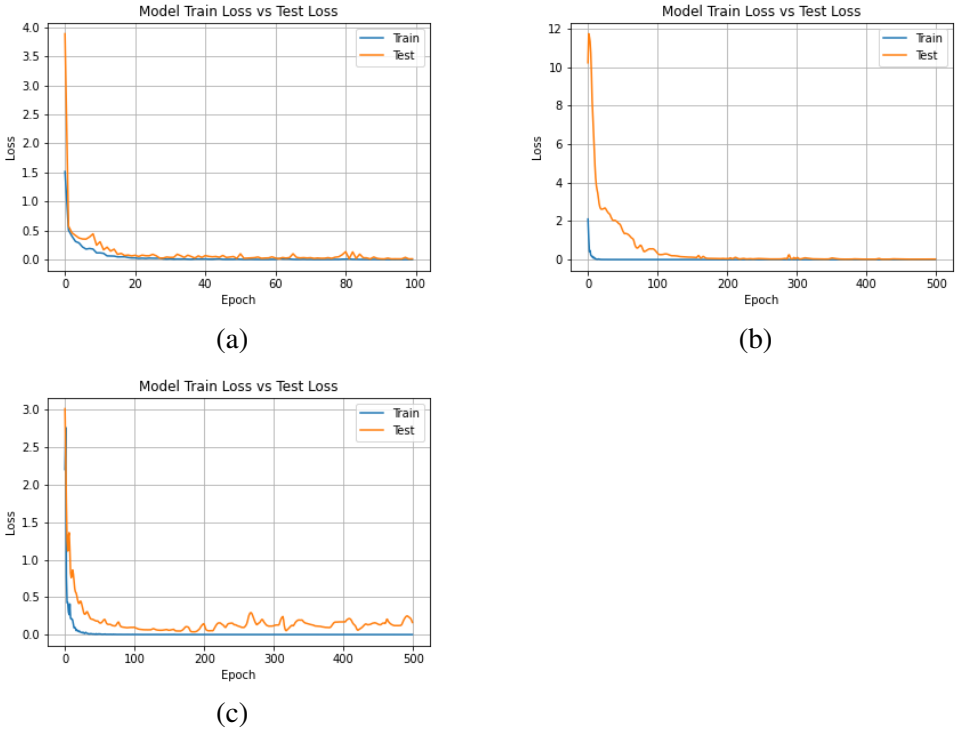


FIGURE 7.6: Training (blue) and testing (orange) loss curves for network trained in three experimental protocol (a) LSO; (b) LTO; (c) LWO. The learning curves indicate that the network trained in LWO and LSO manner suffer from over-fitting.

7.9 Conclusion

This chapter demonstrated recognition of grammatical class of imagined words from EEG signals. In addition, the chapter conducted experiments performing multi-class classification of imagined words. The evaluation of the proposed MC-CNN on two EEG dataset shows that our method is capable of recognizing grammatical class of the imagined words. Further, the analysis also demonstrates that better recognition rate can be achieved by processing information from different regions of the brain separately and combining them later. The second experiment in this chapter, showed that dividing multi-class classification of imagined words into multiple levels can achieve better accuracy of 52.9% and 37.1% compared to the standard classification method with 27.2% accuracy. However, section 7.7.1 shows that the features learned in discriminating grammatical classes does not provide much discriminative information about the underlying word. Better performance of the proposed method can be attributed to its tree like structure i.e., the multi-level processing of imagined words. The presented results shows that the proposed multi-level method is robust in classifying multiple imagined word.

7.10 Summary

A thought-to-speech brain computer interface (BCI) system should be able to predict wide range of words. However, multi-class classification of imagined words from a large dictionary of words is a challenging task.Therefore, an intuitive approach is required in classification of imagined words. In all languages, words belong to different grammatical classes. At a fundamental level, words can be divided into two basic grammatical classes of nouns and verbs. Recognizing words by first learning which grammatical class they belong to can help optimizing the accuracy of thought-to-speech BCI with large vocabulary. Therefore, this chapter first examined grammatical class recognition from EEG signals of imagined words.

In this chapter, two main experiments were performed: (1) recognition of grammatical class of imagined words from EEG signals, (2) multi-class classification of imagined words using multi-level recognition system. In the first experimental analysis, imagined words belonged to two grammatical classes of noun and verb. The network was trained and tested under several different evaluation conditions. The evaluation condition involved recognition of grammatical class of words that was not used during network training. In order to extract local features from each brain region separately, we propose an application multi-channel CNN. The proposed method was evaluated using data from 17 subjects. The proposed method is also evaluated on publicly available EEG dataset. In the second experiment, multi-class recognition of imagined words was performed. A multi-stage recognition system was proposed to recognize the imagined words from EEG signals. The proposed method recognized the grammatical class of the imagined words at level 1 and recognizes the imagined word within the grammatical class at level 2.

Chapter 8

Conclusion

This chapter presents the findings from the studies in this research. It then discusses the contributions made to the domain of recognition of imagined speech from EEG signal. Finally, it makes recommendations for future research based on the findings of this thesis.

8.1 Review of Main Findings

The purpose of this thesis was to understand the challenge of predicting covert speech from EEG signals, to propose frameworks for alleviating this difficulty, and to demonstrate that using these frameworks it is possible to build a thought-to-speech brain computer interface (BCI). Using the recorded EEG dataset and proposed methods, these studies have been able to predict imagined words from EEG signals for many challenging tasks: imagined word recognition, grammatical class recognition of imagined words, and predicting if the user (subject) was performing imagined speech or other cognitive tasks. The analysis performed in this thesis conclude that a thought-to-speech BCI can be implemented using deep learning methods which can achieve reliable response time and accuracy in offline analysis.

Chapter 3 achieved the first objective defined in chapter 1. This chapter presents a new EEG data recorded four imagined speech and three other cognitive tasks from 17 subjects. However, only 12 subjects were used in the analysis in the later chapters because of contaminated data from five subjects. Most of the previous studies had recorded EEG signals for imagined phonemes and syllables, however the EEG dataset in this study is recorded for mentally spoken words. Further, the words presented as stimulus belonged to two grammatical classes of nouns and verbs. In addition, subjects were instructed to mentally speak the presented word only once, whereas past studies have recorded EEG signals while subject mentally repeated the presented the words/phoneme/syllables multiple times. Further, advantages of using EEG signals in time-frequency domain have been summarized.

Chapter 4 demonstrated that EEG signals produced during imagined speech can be discriminated when compared with EEG signals during other cognitive tasks, such as visual imagery and overt speech. This chapter investigated the temporal window and frequency bands using multivariate pattern analysis (MVPA). This provided temporal information about the most discriminative behavior between imagined speech and other cognitive tasks. Further, the nature of neural activity between two cognitive task was also investigated, using the most discriminative frequency band. The most discriminative time window was just after the stimulus onset, and lower frequency range ($<30Hz$) performed best. Using this knowledge, we proposed a framework for recognition of cognitive tasks from EEG signals. The proposed method used K-means clustering algorithm for selecting electrodes and a CNN-attention network for spatio-temporal feature learning. The framework achieved high recognition rate and was evaluated on EEG signal of length $3000s$ and $600ms$ (obtained using MVPA), with later achieving high recognition rate. Chapter 4 along with chapter 5 shows that best performance is achieved using electrode from Parieto-Occipital lobe, this consistent performance in different experimental evaluation highlights the importance this brain region plays in recognition of imagined speech from EEG signals.

Chapter 5 addressed second objective of comparing different methods for recognition of imagined speech EEG signals. We showed that methods such as eigen features, time-frequency statistical features

and common spatial pattern features are unable to predict imagined speech EEG signals due to their trial-to-trial variability and non-stationary nature. Further, in order to overcome inter-trial variability within same class we used dynamic time warping (DTW) method to measure similarity. The DTW method with proposed electrode combining method outperformed the recognition rate achieved by features used earlier. In addition, brain areas contributing to correct recognition were also reported using DTW method. Although, DTW achieved better performance it failed to achieve high recognition rate and response time of the method was slow to be implemented in a practical BCI solution. The problem we believe made recognition difficult was inconsistency in features learned by the proposed network due to inter-subject variability in the EEG signals.

In chapter 6, a third object was achieved as deep learning methods for recognition of imagined speech from EEG signals were investigated. This introduced an electrode selection method, based on mean window power in frequency domain. Further, a CNN-attention network was proposed to learning features in frequency domain and map temporal dynamics (similar to DTW) in time domain. The network outperformed both the methods previously used in chapter 4. Further, the performance of the network was optimized by using the following methods: (1) reducing the number of electrodes used in training and testing the network, (2) reducing the size of the input spectrograms along the temporal dimension. The network performed well for subject dependent method, however the performance reduced in subject independent task.

Chapter 7 performed recognition of grammatical class of imagined words from EEG signals achieving the fourth objective. For achieving this, the chapter proposed an application of MC-CNN network for extracting local features from different brain region. The proposed method achieved high recognition rate in predicting the grammatical classes of the imagined words. Further, this method was extended to a multi-level recognition system with two levels. Level 1, recognize the grammatical class of the imagined word and level 2 performs classification among the given words within the grammatical class selected at level 1. The proposed method outperformed standard multi-class classification technique and achieved high recognition rate on two EEG dataset.

8.2 Contribution

The main contributions of this research are as follows:

- A new database containing EEG signals recorded for four tasks: imagined speech, overt speech, visual imagery, and visual perception. The dataset was recorded from 17 subjects. The imagined speech and overt speech task contain EEG signals for words belonging to noun and verb grammatical class.
- Evidence that deep learning methods outperform standard features extraction and classification techniques for recognition of imagined speech from EEG signals. The two methods were examined for EEG signals for imagined words.
- A novel multi-level recognition system which first recognize the grammatical class for recognition of imagined word from EEG signals.
- A novel electrode selection method and convolutional attention network for spectro-temporal feature learning for recognition of imagined words from EEG signals.
- Successful discrimination between imagined speech task and non-speech tasks, visual imagery and overt speech.

    1. The analysis showed that EEG signals for imagined speech were most discriminative of other cognitive tasks during the 0-500*ms* time window after the stimulus onset.

2. A novel framework for classifying imagined speech task from other language based cognitive tasks from EEG signals. The framework uses K-means clustering algorithm for electrode selection and convolutional-attention network for spatio-temporal feature extraction and classification.

8.3 Research Limitations

The current study focused developing methods for recognition of imagined words (speech) from EEG signals. Findings of this thesis suggests that deep learning models are better suited for developing a though-to-speech BCI system. Further, this work has proposed new methods for achieving high recognition rate for imagined speech from EEG signals. However, this work has some limitations:

- For imagined speech task, only 10 trials were recorded for each word from each subject. This is a small sample-size, when it comes to training and testing deep learning models for a subject dependent study. Limited dataset for each subject could negatively effective the network performance.
- The EEG signals were recorded for each subject in a single session, therefore the performance evaluation of the proposed methods is only validated on single session EEG signals. However, we acknowledge that training and testing of the proposed methods on EEG signals from different sessions could have provided better evaluation.
- EEG signals were recorded for imagined speech of 10 words. However, for a practical though-to-text BCI a larger vocabulary needs to be investigated.
- For recording EEG signals we used a 64 channel EEG cap which was connected with an external amplifier. However, such devices are not suitable for designing a though-to-text BCI for daily use. Thus, it is recommended that for future research we use a wireless mobile device with fewer electrodes.
- The repeated presentation of the blank screen for a fixed period of $1s$ during the data acquisition is a limitation because expectation of a stimulus can lead to improved performance of the task at hand (Meijs et al., 2019). Therefore, in future the experimental design should contain blank screen of random intervals between ($0.5s$ to $1.5s$)

8.4 Future Work

As discussed in the previous section, there are several problems in the current work that can be improved.

8.4.1 Experimental Design

In chapter 4, 6, and 7, it is shown that event related activity occurs between 0-500ms window. Hence, the experimental design could be improved by reducing the time for imagined speech task to $1s$. Further, it would be better to use a wireless EEG headset such as Emotive epoch capable of transferring signals to a computer at high speed. Also, the number of trials for each word could be increased and the recording can be done in multiple sessions for the same subject.

8.4.2 Methodology

The proposed CNN-attention network for learning spectro-temporal features from EEG signals of imagined words in chapter 6. On the other hand, it is known that attention mechanism in itself is capable of modeling events and features that are task discriminative (Vaswani et al., 2017). Therefore, it is important to investigate potential of purely attention-based networks, to capture spectro-temporal features from

EEG signals. Further, events in EEG signals are unaligned and using alignment techniques such as DTW improved the performance of the system (chapter 5). Therefore, a deep-learning based alignment of EEG signals could possibly improve the recognition by reducing inter-trial and inter-subject variations, especially when using spatio-temporal features.

In addition, the multi-level recognition system proposed in chapter 7 can be improved by making it an end-to-end model, similar to network proposed in (Nguyen et al., 2020; Roy et al., 2020). Also, interpretation of the features learned by MC-CNN network can provide information about the important patterns within the spectrograms. Investigation of features learned by the MC-CNN can provide important temporal or spectral information.

The networks designed and evaluated in this work suffer from overfitting, such problems can be avoided using methods such as early stopping or using a separate training, validation, and test data when evaluation the network's performance

8.5 Conclusion

The work in this thesis proposed deep learning methods for recognition of imagined speech. The work propose an electrode selection method to reduce dimensionality of EEG data and improve performance in recognition of imagined words. This results suggest that a communication based BCI system can be implemented using the deep neural networks. The ability of deep learning models to learn robust feature because of the multiple layers offer an advantage compared to standard feature extraction methods. Specifically, the networks designed using the convolutional operation can efficiently used in learning representations from the EEG signals. Although, the results indicate that the methodology is still far from real-world use, but they suggest interesting lines of future research and add relevant knowledge to the-state-of-the-art.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, *22*(10), 1533–1545.

Abdelnabi, S., Huang, M. X., & Bulling, A. Towards high-frequency SSVEP-based target discrimination with an extended alphanumeric keyboard. In: *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE. 2019, 4181–4186.

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on K-nearest neighbor classifier performance: A review. *Big data*, *7*(4), 221–248.

Al-Fahoum, A. S., & Al-Fraihat, A. A. (2014). Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains. *ISRN neuroscience*, *2014*.

Albasri, A., Abdali-Mohammadi, F., & Fathi, A. (2019). EEG electrode selection for person identification thru a genetic-algorithm method. *Journal of medical systems*, *43*(9), 1–12.

Allison, B. Z., Wolpaw, E. W., & Wolpaw, J. R. (2007). Brain–computer interface systems: Progress and prospects. *Expert review of medical devices*, *4*(4), 463–474.

Alsaleh, M. (2019). *Toward an imagined speech-based brain computer interface using EEG signals* (Doctoral dissertation). University of Sheffield.

Amit, E., Hoeflin, C., Hamzah, N., & Fedorenko, E. (2017). An asymmetrical relationship between verbal and visual thinking: Converging evidence from behavior and fMRI. *NeuroImage*, *152*, 619–627.

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE. 2008, 2390–2397.

Angrick, M., Herff, C., Mugler, E., Tate, M. C., Slutzky, M. W., Krusienski, D. J., & Schultz, T. (2019). Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of neural engineering*, *16*(3), 036019.

Antoniades, A., Spyrou, L., Took, C. C., & Sanei, S. Deep learning for epileptic intracranial EEG data. In: *2016 ieee 26th international workshop on machine learning for signal processing (MLSP)*. IEEE. 2016, 1–6.

Arjestan, M. A., Vali, M., & Faradji, F. Brain computer interface design and implementation to identify overt and covert speech. In: *Biomedical engineering and 2016 1st international iranian conference on biomedical engineering (ICBME), 2016 23rd iranian conference on*. IEEE. 2016, 59–63.

Aydemir, O., & Kayikcioglu, T. (2014). Decision tree structure based classification of EEG signals recorded during two dimensional cursor movement imagery. *Journal of neuroscience methods*, *229*, 68–75.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bakhshali, M. A., Khademi, M., Ebrahimi-Moghadam, A., & Moghimi, S. (2020). EEG signal classification of imagined speech based on riemannian distance of correntropy spectral density. *Biomedical Signal Processing and Control*, *59*, 101899.

Balaji, A., Haldar, A., Patil, K., Ruthvik, T. S., Valliappan, C., Jartarkar, M., & Baths, V. EEG-based classification of bilingual unspoken speech using ANN. In: *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2017, 1022–1025.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, *18*, 1–8.

Bashashati, A., Fatourechi, M., Ward, R. K., & Birch, G. E. (2007). A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering*, *4*(2), R32.

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.

Bierwisch, M. (1999). Words in the brain are not just labelled concepts. *Behavioral and Brain Sciences*, *22*(2), 280–282.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., & Curio, G. (2007). The non-invasive berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, *37*(2), 539–550.

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., & Yvert, B. (2016). Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS computational biology*, *12*(11), e1005119.

Borgatti, S. (2019). *Distance and correlation*. https://cmci.colorado.edu/classes/INFO-1301/files/borgatti.htm

Bostanov, V. (2004). Bci competition 2003-data sets ib and iib: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Transactions on Biomedical engineering*, *51*(6), 1057–1061.

Bowers, A., Saltuklaroglu, T., Harkrider, A., & Cuellar, M. (2013). Suppression of the μ rhythm during speech and non-speech discrimination revealed by independent component analysis: Implications for sensorimotor integration in speech processing. *PloS one*, *8*(8), e72024.

Brigham, K., & Kumar, B. V. Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy. In: *2010 4th international conference on bioinformatics and biomedical engineering*. IEEE. 2010, 1–4.

Brigham, K., & Kumar, B. V. Subject identification from electroencephalogram (EEG) signals during imagined speech. In: *Biometrics: Theory applications and systems (BTAS), 2010 fourth ieee international conference on*. IEEE. 2010, 1–8.

Brownlee, J. (2016). *Deep learning with python: Develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery.

Butts, D. A., Weng, C., Jin, J., Yeh, C.-I., Lesica, N. A., Alonso, J.-M., & Stanley, G. B. (2007). Temporal precision in the neural code and the timescales of natural vision. *Nature*, *449*(7158), 92–95.

Chen, D.-W., Miao, R., Yang, W.-Q., Liang, Y., Chen, H.-H., Huang, L., Deng, C.-J., & Han, N. (2019). A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition. *Sensors*, *19*(7), 1631.

Chen, X., Wang, Y., Nakanishi, M., Gao, X., Jung, T.-P., & Gao, S. (2015). High-speed spelling with a non-invasive brain–computer interface. *Proceedings of the national academy of sciences*, *112*(44), E6058–E6067.

Chi, X, Hagedorn, J., Schoonover, D, & D'Zmura, M. (2011). EEG-based discrimination of imagined speech phonemes. *International Journal of Bioelectromagnetism*, *13*(4), 201–206.

Chollet, F. et al. (2015). Keras.

Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*.

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. MIT press.

Coles, M. G., & Rugg, M. D. (1995). *Event-related brain potentials: An introduction.* Oxford University Press.

Crepaldi, D., Berlingeri, M., Paulesu, E., & Luzzatti, C. (2011). A place for nouns and a place for verbs? a critical review of neurocognitive data on grammatical-class effects. *Brain and language*, *116*(1), 33–49.

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, *90*(11), 4957–4960.

Das, K., Giesbrecht, B., & Eckstein, M. P. (2010). Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers. *Neuroimage*, *51*(4), 1425–1437.

DaSalla, C. S., Kambara, H., Koike, Y., & Sato, M. Spatial filtering and single-trial classification of EEG during vowel speech imagery. In: *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology*. ACM. 2009, 27.

DaSalla, C. S., Kambara, H., Sato, M., & Koike, Y. (2009b). Single-trial classification of vowel speech imagery using common spatial patterns. *Neural networks*, *22*(9), 1334–1339.

Dash, D., Ferrari, P., & Wang, J. (2020). Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Frontiers in neuroscience*, *14*.

Dash, D., Ferrari, P., & Wang, J. Role of brainwaves in neural speech decoding. In: *2020 28th european signal processing conference (EUSIPCO)*. IEEE. 2021, 1357–1361.

Datta, S., & Boulgouris, N. V. (2021). Recognition of grammatical class of imagined words from eeg signals using convolutional neural network. *Neurocomputing*, *465*, 301–309.

De Benedictis, A., Duffau, H., Paradiso, B., Grandi, E., Balbi, S., Granieri, E., Colarusso, E., Chioffi, F., Marras, C. E., & Sarubbo, S. (2014). Anatomo-functional study of the temporo-parieto-occipital region: Dissection, tractographic and brain mapping evidence from a neurosurgical perspective. *Journal of anatomy*, *225*(2), 132–151.

Dehaene, S., & King, J.-R. (2016). Decoding the dynamics of conscious perception: The temporal generalization method. *Micro-, meso-and macro-dynamics of the brain*, 85–97.

Demanuele, C., James, C. J., & Sonuga-Barke, E. J. (2007). Distinguishing low frequency oscillations within the 1/f spectral behaviour of electromagnetic brain signals. *Behavioral and Brain Functions*, *3*(1), 62.

Demb, J. B., Desmond, J. E., Wagner, A. D., Vaidya, C. J., Glover, G. H., & Gabrieli, J. (1995). Semantic encoding and retrieval in the left inferior prefrontal cortex: A functional mri study of task difficulty and process specificity. *Journal of Neuroscience*, *15*(9), 5870–5878.

Dentico, D., Cheung, B. L., Chang, J., Guokas, J., Boly, M., Tononi, G., & Van Veen, B. (2014). Reversal of cortical information flow during visual imagery as compared to visual perception. *Neuroimage*, *100*, 237–243.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, *19*(1), 158–164.

Donahue, J., Anne H., L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, 2625–2634.

D'Zmura, M., Deng, S., Lappas, T., Thorpe, S., & Srinivasan, R. Toward EEG sensing of imagined speech. In: *International conference on human-computer interaction*. Springer. 2009, 40–48.

Encyclopaedia Britannica, I. (2020). *Broca area*. https://www.britannica.com/science/Broca-area

Esfahani, E. T., & Sundararajan, V. (2012). Classification of primitive shapes using brain–computer interfaces. *Computer-Aided Design*, *44*(10), 1011–1019.

Farina, D., Do Nascimento, O. F., Lucas, M.-F., & Doncarli, C. (2007). Optimization of wavelets for classification of movement-related cortical potentials generated by variation of force-related parameters. *Journal of neuroscience methods*, *162*(1-2), 357–363.

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of broca's area in speech. *Proceedings of the National Academy of Sciences*, *112*(9), 2871–2875.

Fonken, Y. M., Kam, J. W., & Knight, R. T. (2020). A differential role for human hippocampus in novelty and contextual processing: Implications for P300. *Psychophysiology*, *57*(7), e13400.

Fuentemilla, L., Penny, W. D., Cashdollar, N., Bunzeck, N., & Düzel, E. (2010). Theta-coupled periodic replay in working memory. *Current Biology*, *20*(7), 606–612.

Fyshe, A. (2020). Studying language in context using the temporal generalization method. *Philosophical Transactions of the Royal Society B*, *375*(1791), 20180531.

Gadhoumi, K., Lina, J.-M., Mormann, F., & Gotman, J. (2016). Seizure prediction for therapeutic devices: A review. *Journal of neuroscience methods*, *260*, 270–282.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for "common sense": An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience*, *8*(2), 89–106.

García, A. A. T., García, C. A. R., & Pineda, L. V. Toward a silent speech interface based on unspoken speech. In: *Biosignals*. 2012, 370–373.

gaurdian. (2021). *Elon Musk startup shows monkey with brain chip implants playing video game*. Retrieved September 30, 2010, from https://www.theguardian.com/technology/2021/apr/09/elon-musk-neuralink-monkey-video-game

Gernsbacher, M. A., & Kaschak, M. P. (2003). Neuroimaging studies of language production and comprehension. *Annual review of psychology*, *54*(1), 91–114.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Graimann, B., Allison, B., Mandel, C., Lüth, T., Valbuena, D., & Gräser, A. Non-invasive brain-computer interfaces for semi-autonomous assistive devices. In: *Robust intelligent systems*. Springer, 2008, pp. 113–138.

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

Graves, A., Mohamed, A. R., & Hinton, G. Speech recognition with deep recurrent neural networks. In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2013, 6645–6649.

Guan, S., Zhao, K., & Yang, S. (2019). Motor imagery eeg classification based on decision tree framework and riemannian geometry. *Computational intelligence and neuroscience*, *2019*.

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., Law, R., Siebert, S. A., Bartels, J. L., Andreasen, D. S., et al. (2009). A wireless brain-machine interface for real-time speech synthesis. *PloS one*, *4*(12), e8218.

Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., Gramatica, F., & Edlinger, G. (2009). How many people are able to control a P300-based brain–computer interface (BCI)? *Neuroscience letters*, *462*(1), 94–98.

Guo, D., Zhang, S., Wright, K. L., & McTigue, E. M. (2020). Do you get the picture? a meta-analysis of the effect of graphics on reading comprehension. *AERA Open*, *6*(1), 2332858420901696.

Hartmann, K. G., Schirrmeister, R. T., & Ball, T. (2018). EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv preprint arXiv:1806.01875*.

He, K., Zhang, X., Ren, S., & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the ieee international conference on computer vision*. 2015, 1026–1034.

Heger, D., Herff, C., Pesters, A. d., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. Continuous speech recognition from ECOG. In: *Sixteenth annual conference of the international speech communication association*. 2015.

Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, *9*, 217.

Herff, C., & Schultz, T. (2016). Automatic speech recognition from neural signals: A focused review. *Frontiers in neuroscience*, *10*, 429.

Herff, C. E. (2016). *Speech processes for brain-computer interfaces* (Doctoral dissertation). University of Bremen, Germany.

Herwig, U., Satrapi, P., & Schönfeldt-Lecuona, C. (2003). Using the international 10-20 EEG system for positioning of transcranial magnetic stimulation. *Brain topography*, *16*(2), 95–99.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hong, B., Guo, F., Liu, T., Gao, X., & Gao, S. (2009). N200-speller using motion-onset visual response. *Clinical neurophysiology*, *120*(9), 1658–1666.

Hornero, R., Abásolo, D., Escudero, J., & Gómez, C. (2009). Nonlinear analysis of electroencephalogram and magnetoencephalogram recordings in patients with alzheimer's disease. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1887), 317–336.

Hu, L., Xiao, P, Zhang, Z., Mouraux, A., & Iannetti, G. D. (2014). Single-trial time–frequency analysis of electrocortical signals: Baseline correction and beyond. *Neuroimage*, *84*, 876–887.

Huang, J., Carr, T. H., & Cao, Y. (2002). Comparing cortical activations for silent and overt speech using event-related fMRI. *Human brain mapping*, *15*(1), 39–53.

Hwang, G., Jacobs, J., Geller, A., Danker, J., Sekuler, R., & Kahana, M. J. (2005). Eeg correlates of verbal and nonverbal working memory. *Behavioral and Brain Functions*, *1*(1), 20.

Hwang, H.-J., Choi, H., Kim, J.-Y., Chang, W.-D., Kim, D.-W., Kim, K., Jo, S., & Im, C.-H. (2016). Toward more intuitive brain–computer interfacing: Classification of binary covert intentions using functional near-infrared spectroscopy. *Journal of biomedical optics*, *21*(9), 091303.

Insights, C. (2019). *21 Neurotech Startups to watch: Brain-machine interfaces, implantables, and neuroprosthetics*. Retrieved September 30, 2010, from https://www.cbinsights.com/research/neurotech-startups-to-watch/

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Ismail, A., Abdlerazek, S., & El-Henawy, I. M. (2020). Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping. *Sustainability*, *12*(6), 2403.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651–666.

Jenson, D., Bowers, A. L., Harkrider, A. W., Thornton, D., Cuellar, M., & Saltuklaroglu, T. (2014). Temporal dynamics of sensorimotor integration in speech perception and production: Independent component analysis of EEG data. *Frontiers in psychology*, *5*, 656.

Jiang, W., & Yin, Z. Human activity recognition using wearable sensors by deep convolutional neural networks. In: *Proceedings of the 23rd ACM international conference on multimedia*. 2015, 1307–1310.

Khader, P., & Rösler, F. (2004). EEG power and coherence analysis of visually presented nouns and verbs reveals left frontal processing differences. *Neuroscience Letters*, *354*(2), 111–114.

khan academy. (2016). *Overview of neuron structure and function*. Retrieved 9, from https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/overview-of-neuron-structure-and-function

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in cognitive sciences*, *18*(4), 203–210.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., & Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science*, *324*(5934), 1583–1585.

Kokoska, S., & Zwillinger, D. (2000). *CRC standard probability and statistics tables and formulae*. Crc Press.

Koles, Z. J., Lazar, M. S., & Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain topography*, *2*(4), 275–284.

Kołodziej, M., & Majkowski, A. (2012). Linear discriminant analysis as EEG features reduction technique for brain-computer interfaces.

Konak, A., Coit, D. W., & Smith, A. E. (2006). Multi-objective optimization using genetic algorithms: A tutorial. *Reliability engineering & system safety*, *91*(9), 992–1007.

Kösem, A., & Van Wassenhove, V. (2017). Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*(5), 536–544.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature reviews neuroscience*, *2*(9), 635–642.

Kouijzer, M. E., de Moor, J. M., Gerrits, B. J., Buitelaar, J. K., & van Schie, H. T. (2009). Long-term effects of neurofeedback treatment in autism. *Research in Autism Spectrum Disorders*, *3*(2), 496–501.

Krishna, G., Tran, C., Carnahan, M., & Tewfik, A. Advancing speech recognition with no speech or with noisy speech. In: *2019 27th european signal processing conference (EUSIPCO)*. IEEE. 2019, 1–5.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012, 1097–1105.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNET: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, *15*(5), 056013.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, S.-H., Lee, M., Jeong, J.-H., & Lee, S.-W. Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery. In: *2019 ieee international conference on systems, man and cybernetics (SMC)*. IEEE. 2019, 4409–4414.

Leech, G., Rayson, P. et al. (2014). *Word frequencies in written and spoken english: Based on the british national corpus*. Routledge.

Levy, W. J. (1987). Effect of epoch length on power spectrum analysis of the EEG. *Anesthesiology*, *66*(4), 489–495.

Li, J., Tang, Y.-y., Zhou, L., Yu, Q.-b., Li, S., & Dan-ni, S. (2010). EEG dynamics reflects the partial and holistic effects in mental imagery generation. *Journal of Zhejiang University SCIENCE B*, *11*(12), 944–951.

Linden, D. E. (2005). The P300: Where in the brain is it produced and what does it tell us? *The Neuroscientist*, *11*(6), 563–576.

Long, M., Cao, Y., Wang, J., & Jordan, M. Learning transferable features with deep adaptation networks. In: *International conference on machine learning*. PMLR. 2015, 97–105.

Lotte, F. (2008). *Study of electroencephalographic signal processing and classification techniques towards the use of brain-computer interfaces in virtual reality applications* (Doctoral dissertation). INSA de Rennes.

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*.

Luck, S. J. (2005). An introduction to the event-related potential technique MIT press. *Cambridge, Ma*, 45–64.

Macdonald, J. S. P., Mathan, S., & Yeung, N. (2011). Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations. *Frontiers in psychology*, *2*, 82.

MacQueen, J. et al. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*. *1*. (14). Oakland, CA, USA. 1967, 281–297.

Maier-Hein, L., Metze, F., Schultz, T., & Waibel, A. Session independent non-audible speech recognition using surface electromyography. In: *IEEE workshop on automatic speech recognition and understanding, 2005.* IEEE. 2005, 331–336.

Majumder, S., Guragain, B., Wang, C., & Wilson, N. On-board drowsiness detection using eeg: Current status and future prospects. In: *2019 ieee international conference on electro information technology (eit)*. IEEE. 2019, 483–490.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R. T., & Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in neuroengineering*, *7*, 14.

Martin, S., Brunner, P., Iturrate, I., Millán, J. d. R., Schalk, G., Knight, R. T., & Pasley, B. N. (2016). Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, *6*, 25803.

Matsumoto, M., & Hori, J. (2014). Classification of silent speech using support vector machine and relevance vector machine. *Applied Soft Computing*, *20*, 95–102.

Meijs, E. L., Mostert, P., Slagter, H. A., de Lange, F. P., & van Gaal, S. (2019). Exploring the role of expectations and stimulus relevance on stimulus-specific neural representations and conscious report. *Neuroscience of consciousness*, *2019*(1), niz011.

Monesi, M. J., Accou, B., Montoya-Martinez, J., Francart, T., & Van Hamme, H. An LSTM based architecture to relate speech stimulus to EEG. In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2020, 941–945.

Muller, K.-R., Anderson, C. W., & Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, *11*(2), 165–169.

Münßinger, J. I., Halder, S., Kleih, S. C., Furdea, A., Raco, V., Hösle, A., & Kubler, A. (2010). Brain painting: First evaluation of a new brain–computer interface application with ALS-patients and healthy volunteers. *Frontiers in neuroscience*, *4*, 182.

Nair, V., & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In: *Icml*. 2010.

Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2017). Inferring imagined speech using EEG signals: A new approach using riemannian manifold features. *Journal of neural engineering*, *15*(1), 016002.

Nguyen, X.-P., Joty, S., Hoi, S. C., & Socher, R. (2020). Tree-structured attention with hierarchical accumulation. *arXiv preprint arXiv:2002.08046*.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). Determination press San Francisco, CA.

Nijboer, F., Sellers, E., Mellinger, J., Jordan, M. A., Matuz, T., Furdea, A., Halder, S., Mochty, U., Krusienski, D., Vaughan, T., et al. (2008). A P300-based brain–computer interface for people with amyotrophic lateral sclerosis. *Clinical neurophysiology*, *119*(8), 1909–1916.

Nikhil, N. (2018). *Is ReLU after sigmoid bad?* [Online Available: https://towardsdatascience.com/is-relu-after-sigmoid-bad-661fda45f7a2].

on Deafness, N. I., & communication Disorder, O. (2019). *Statistics on voice, speech, and language*. https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language

Palmer, E. D., Rosen, H. J., Ojemann, J. G., Buckner, R. L., Kelley, W. M., & Petersen, S. E. (2001). An event-related fMRI study of overt and covert word stem completion. *Neuroimage*, *14*(1), 182–193.

Panachakel, J. T., Ramakrishnan, A., & Ananthapadmanabha, T. Decoding imagined speech using wavelet features and deep neural networks. In: *2019 ieee 16th india council international conference (indicon)*. IEEE. 2019, 1–4.

Patterson, J., & Gibson, A. (2017). *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc."

Paul, Y., Jaswal, R. A., & Kajal, S. Classification of EEG based imagine speech using time domain features. In: *2018 international conference on recent innovations in electrical, electronics & communication engineering (ICRIEECE)*. IEEE. 2018, 2921–2924.

Pawar, D., & Dhage, S. (2020). Multiclass covert speech classification using extreme learning machine. *Biomedical Engineering Letters*, *10*(2), 217–226.

Pei, X., Hill, J., & Schalk, G. (2012). Silent communication: Toward using brain signals. *IEEE pulse*, *3*(1), 43–46.

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., & Schalk, G. (2011). Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage*, *54*(4), 2960–2972.

Petsche, H., Lacroix, D., Lindner, K., Rappelsberger, P., & Schmidt-Henrich, E. (1992). Thinking with images or thinking with language: A pilot EEG probability mapping study. *International Journal of Psychophysiology*, *12*(1), 31–39.

Picton, T. W. (1992). The P300 wave of the human event-related potential. *Journal of clinical neurophysiology*, *9*(4), 456–479.

Popp, M., Trumpp, N. M., Sim, E.-J., & Kiefer, M. (2019). Brain activation during conceptual processing of action and sound verbs. *Advances in Cognitive Psychology*, *15*(4), 236.

Porbadnigk, A., Wester, M., & Jan-p Calliess, T. S. (2009). EEG-based speech recognition impact of temporal effects.

Preissl, H., Pulvermüller, F., Lutzenberger, W., & Birbaumer, N. (1995). Evoked potentials distinguish between nouns and verbs. *Neuroscience Letters*, *197*(1), 81–83.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature reviews neuroscience*, *6*(7), 576–582.

Pulvermüller, F. (2018). Neurobiological mechanisms for semantic feature extraction and conceptual flexibility. *Topics in Cognitive Science*, *10*(3), 590–620.

Pulvermüller, F., Lutzenberger, W., & Preissl, H. (1999). Nouns and verbs in the intact brain: Evidence from event-related potentials and high-frequency cortical responses. *Cerebral cortex*, *9*(5), 497–506.

Qian, S., Liu, H., Liu, C., Wu, S., & San Wong, H. (2018). Adaptive activation functions in convolutional neural networks. *Neurocomputing*, *272*, 204–212.

Ramoser, H., Muller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE transactions on rehabilitation engineering*, *8*(4), 441–446.

Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., & Haider, H. (2004). Fractionating language comprehension via frequency characteristics of the human EEG. *Neuroreport*, *15*(3), 409–412.

Rong, L., Jianzhong, Z., Ming, L., & Xiangfeng, H. A wearable acceleration sensor system for gait recognition. In: *2007 2nd IEEE conference on industrial electronics and applications*. IEEE. 2007, 2654–2659.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Roy, D., Panda, P., & Roy, K. (2020). Tree-CNN: A hierarchical deep convolutional neural network for incremental learning. *Neural Networks*, *121*, 148–160.

Saha, P., Abdul-Mageed, M., & Fels, S. (2019). Speak your mind! towards imagined speech recognition with hierarchical deep learning. *arXiv preprint arXiv:1904.05746*.

Saha, P., & Fels, S. Hierarchical deep feature learning for decoding imagined speech from EEG. In: *Proceedings of the aaai conference on artificial intelligence*. *33*. 2019, 10019–10020.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, *26*(1), 43–49.

Sarmiento, L., Lorenzana, P, Cortes, C., Arcos, W., Bacca, J., & Tovar, A. Brain computer interface (BCI) with EEG signals for automatic vowel recognition based on articulation mode. In: *5th*

ISSNIP-IEEE biosignals and biorobotics conference (2014): Biosignals and robotics for better and safer living (BRC). IEEE. 2014, 1–4.

Schendan, H. E., & Ganis, G. (2012). Electrophysiological potentials reveal cortical mechanisms for mental imagery, mental simulation, and grounded (embodied) cognition. *Frontiers in psychology*, *3*, 329.

Schilling, A., Tomasello, R., Henningsen-Schomers, M. R., Zankl, A., Surendra, K., Haller, M., Karl, V., Uhrig, P., Maier, A., & Krauss, P. (2020). Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods. *Language, Cognition and Neuroscience*, 1–20.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391–5420.

Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C., & Brumberg, J. S. (2017). Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2257–2271.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Sereshkeh, A. R., Trott, R., Bricout, A., & Chau, T. (2017). EEG classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2292–2300.

Shapiro, K. A., Pascual-Leone, A., Mottaghy, F. M., Gangitano, M., & Caramazza, A. (2001). Grammatical distinctions in the left frontal cortex. *Journal of cognitive neuroscience*, *13*(6), 713–720.

Sharon, R. A., & Murthy, H. A. (2020). Correlation based multi-phasal models for improved imagined speech EEG recognition. *arXiv preprint arXiv:2011.02195*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, S. P., Sharma, M. K., Lay-Ekuakille, A., Gangwar, D., & Gupta, S. (2020). Deep ConvLSTM with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal*, *21*(6), 8575–8582.

Siuly. (2012). *Analysis and classification of eeg signals* (Doctoral dissertation). UNIVERSITY OF SOUTHERN QUEENSLAND.

Siuly, S., Li, Y., & Zhang, Y. (2016). EEG signal analysis and classification. *IEEE Trans Neural Syst Rehabilit Eng*, *11*, 141–144.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Staljanssens, W. (2013). *Brain-computer interfaces based on imaginary hand movement using EEG beamforming and a real-time cursor control application* (Master's thesis). Ghent University.

Stephan, F., Saalbach, H., & Rossi, S. (2020). Inner versus overt speech production: Does this make a difference in the developing brain? *Brain Sciences*, *10*(12), 939.

Sturm, I., Lapuschkin, S., Samek, W., & Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods*, *274*, 141–145.

Su, Y., Qi, Y., Luo, J.-x., Wu, B., Yang, F., Li, Y., Zhuang, Y.-t., Zheng, X.-x., & Chen, W.-d. (2011). A hybrid brain-computer interface control strategy in a virtual environment. *Journal of Zhejiang University SCIENCE C*, *12*(5), 351–361.

Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert systems with applications*, *37*(12), 8659–8666.

Sun, S., & Zhang, C. (2006). Adaptive feature extraction for EEG signal classification. *Medical and Biological Engineering and Computing*, *44*(10), 931–935.

Suppes, P., Han, B., Epelboim, J., & Lu, Z.-L. (1999). Invariance of brain-wave representations of simple visual images and their names. *Proceedings of the National Academy of Sciences*, *96*(25), 14658–14663.

Suppes, P., Han, B., & Lu, Z.-L. (1998). Brain-wave recognition of sentences. *Proceedings of the National Academy of Sciences*, *95*(26), 15861–15866.

Suppes, P., Lu, Z.-L., & Han, B. (1997). Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, *94*(26), 14965–14969.

Sur, S., & Sinha, V. (2009). Event-related potential: An overview. *Industrial psychiatry journal*, *18*(1), 70.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tabar, Y. R., & Halici, U. (2016). A novel deep learning approach for classification of EEG motor imagery signals. *Journal of neural engineering*, *14*(1), 016003.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, *35*(5), 1299–1312.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in psychology*, *1*, 166.

Tonin, L., Carlson, T., Leeb, R., & Millán, J. d. R. Brain-controlled telepresence robot by motor-disabled people. In: *2011 annual international conference of the ieee engineering in medicine and biology society*. IEEE. 2011, 4227–4230.

Torres, A. S. (2017). *Mudbox activity 1 (head and skull) homework*. https://www.artstation.com/artwork/R6rlA

Torres-García, A., Reyes-García, C., Villaseñor-Pineda, L, & Ramírez-Cortés, J. (2013). Análisis de senales electroencefalográficas para la clasificación de habla imaginada. *Revista mexicana de ingeniería biomédica*, *34*(1), 23–39.

Tsigka, S., Papadelis, C., Braun, C., & Miceli, G. (2014). Distinguishable neural correlates of verbs and nouns: A MEG study on homonyms. *Neuropsychologia*, *54*, 87–97.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426.

von Stein, A., Rappelsberger, P., Sarnthein, J, & Petsche, H. (1999). Synchronization between temporal and parietal cortex during multimodal object processing in man. *Cerebral Cortex*, *9*(2), 137–150.

Wang, L., Zhang, X., & Zhang, Y. Extending motor imagery by speech imagery for brain-computer interface. In: *2013 35th annual international conference of the IEEE engineering in medicine and biology society (embc)*. IEEE. 2013, 7056–7059.

Wang, L., Zhang, X., Zhong, X., & Zhang, Y. (2013b). Analysis and classification of speech imagery EEG for BCI. *Biomedical signal processing and control*, *8*(6), 901–908.

Wester, M. (2006). Unspoken speech-speech recognition based on electroencephalography. *Master's Thesis, Universitat Karlsruhe (TH)*.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, *113*(6), 767–791.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. Cbam: Convolutional block attention module. In: *Proceedings of the european conference on computer vision (ECCV)*. 2018, 3–19.

Xiang, J., & Xiao, Z. (2009). Spatiotemporal and frequency signatures of noun and verb processing: A wavelet-based beamformer study. *Journal of clinical and experimental neuropsychology*, *31*(6), 648–657.

Xie, S., Kaiser, D., & Cichy, R. M. (2020). Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In: *International conference on machine learning*. PMLR. 2015, 2048–2057.

Yamaguchi, H, Yamazaki, T, Yamamoto, K, Ueno, S, Yamaguchi, A, Ito, T, Hirose, S, Kamijo, K, Takayanagi, H, Yamanoi, T, et al. (2015). Decoding silent speech in japanese from single trial EEGS: Preliminary results. *Journal of Computer Science & Systems Biology*, *2015*.

Yang, H., Lin, Q., Han, Z., Li, H., Song, L., Chen, L., He, Y., & Bi, Y. (2017). Dissociable intrinsic functional networks support noun-object and verb-action processing. *Brain and language*, *175*, 29–41.

Yoshimura, N., Satsuma, A., DaSalla, C. S., Hanakawa, T., Sato, M.-a., & Koike, Y. Usability of EEG cortical currents in classification of vowel speech imagery. In: *2011 international conference on virtual rehabilitation*. IEEE. 2011, 1–2.

Yu, X., Bi, Y., Han, Z., Zhu, C., & Law, S.-P. (2012). Neural correlates of comprehension and production of nouns and verbs in chinese. *Brain and language*, *122*(2), 126–131.

Zabidi, A, Mansor, W, Lee, Y., & Fadzal, C. C. W. Short-time fourier transform analysis of EEG signal generated during imagined writing. In: *2012 international conference on system engineering and technology (ICSET)*. IEEE. 2012, 1–4.

Zhang, D., Yao, L., Chen, K., Wang, S., Chang, X., & Liu, Y. (2019a). Making sense of spatio-temporal preserving representations for EEG-based human intention recognition. *IEEE transactions on cybernetics*, *50*(7), 3033–3044.

Zhang, D., Gong, E., Wu, W., Lin, J., Zhou, W., & Hong, B. Spoken sentences decoding based on intracranial high gamma response using dynamic time warping. In: *2012 annual international conference of the IEEE engineering in medicine and biology society*. IEEE. 2012, 3292–3295.

Zhang, G., Davoodnia, V., Sepas-Moghaddam, A., Zhang, Y., & Etemad, A. (2019b). Classification of hand movements from EEG using a deep attention-based LSTM network. *IEEE Sensors Journal*, *20*(6), 3113–3122.

Zhang, X., Yao, L., Wang, X., Monaghan, J., Mcalpine, D., & Zhang, Y. (2019c). A survey on deep learning based brain computer interface: Recent advances and new frontiers. *arXiv preprint arXiv:1905.04149*.

Zhao, S., & Rudzicz, F. Classifying phonological categories in imagined and articulated speech. In: *2015 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2015, 992–996.

Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L. Time series classification using multi-channels deep convolutional neural networks. In: *International conference on web-age information management*. Springer. 2014, 298–310.

Zhou, H., Melloni, L., Poeppel, D., & Ding, N. (2016a). Interpretations of frequency domain analyses of neural entrainment: Periodicity, fundamental frequency, and harmonics. *Frontiers in human neuroscience*, *10*, 274.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. 2016, 207–212.

Zoumpoulaki, A., Alsufyani, A., Filetti, M., Brammer, M., & Bowman, H. (2015). Latency as a region contrast: Measuring ERP latency differences with dynamic time warping. *Psychophysiology*, *52*(12), 1559–1576.

Appendix A

Documentation for the Experiment

This section lists the research ethics approval from the Brunel Research Ethics Committee along with participants information sheet and consent form provided to the participants. The documents are in the following order:

1. Letter of approval by Research Ethics Committee to record EEG signals from human participants.
2. Information sheet provided to the participants providing instruction and explain them about the experimental details.
3. Participant Consent form provided to participants before the EEG recording which was used to collect personal information.

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

11 September 2017

LETTER OF APPROVAL (CONDITIONAL)

Applicant: Mr Sahil Datta

Project Title: Brain Signal Processing

Reference: 7361-LR-Sep/2017- 8301-1

Dear Mr Sahil Datta

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has agreed that there is no objection on ethical grounds to the proposed study. Approval is given on the understanding that the conditions of approval set out below are followed:

- The agreed protocol must be followed. Any changes to the protocol will require prior approval from the Committee by way of an application for an amendment.

- Ref Point A14. Please ensure a risk assessment is completed prior to the activity
- Ref Point A14. Please agree with your supervisor who the participants are likely to be (will they be fellow students for example) how many you will be recruiting, and how you intend to recruit them
- Ref Point A19 - The Committee recommends that you complete the Ethics Training module via Blackboard Learn prior to commencing your research project. Please click on the link below and complete the course online. https://blackboard.brunel.ac.uk/webapps/blackboard/content/listContent.jsp?course_id=_8579_1&content_id=_322757_1

Please note that:

- Research Participant Information Sheets and (where relevant) flyers, posters, and consent forms should include a clear statement that research ethics approval has been obtained from the relevant Research Ethics Committee.
- The Research Participant Information Sheets should include a clear statement that queries should be directed, in the first instance, to the Supervisor (where relevant), or the researcher. Complaints, on the other hand, should be directed, in the first instance, to the Chair of the relevant Research Ethics Committee.
- Approval to proceed with the study is granted subject to receipt by the Committee of satisfactory responses to any conditions that may appear above, in addition to any subsequent changes to the protocol.
- The Research Ethics Committee reserves the right to sample and review documentation, including raw data, relevant to the study.
- You may not undertake any research activity if you are not a registered student of Brunel University or if you cease to become registered, including abeyance or temporary withdrawal. As a deregistered student you would not be insured to undertake research activity. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Professor Hua Zhao

Chair

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London

College of Engineering, Design and Physical Sciences

Department of Electronic and Computer Engineering

PARTICIPANT INFORMATION SHEET- 2017

Study title: Brain Signal Processing

- **No persons with a history of neurological disease can participate.**
- **No persons under 18 years can participate.**
- **No persons with speech-related conditions can participate.**

What is the purpose of the study?

This research aims to record signals from a participant, captured using EEG headset, in order to recognise difference between loud speech (normal speech) and imagined speech or speech of the brain.

Do I have to take part?

No – you have the option to withdraw from the study at any point.

What will happen to me if I take part?

1. You will wear an EEG sensor. This will involve application of gel on the participant's hair.
2. You will be asked to sit on a chair 1 meter away from computer a screen and will be presented with sequence of stimuli, in form of words, objects and numbers that appear on the screen.
3. There will be 4 kinds of tasks for all the stimuli: speaking loud (overt speech) speaking in mind (covert speech), watching an image and imagine the image. You will have to speak the word/ object/number that you see on the screen in both manners above.
4. Each trial will last 2 minutes.

What sort of equipment will be used?

The Brain signal generated by speaking loud and imagined speech can be sensed and recorded (in the form of brain signals) using an EEG headset, which records the electrical activity of the brain. It is a safe, non-invasive procedure, which involves wearing a cap that contains electrodes. Each electrode is a small ceramic disc with a sintered silver coating sitting in a small rubber cup. A saline gel (similar to hair gel) will be injected into the electrodes. The gel is hypoallergenic and harmless to the normal hair and skin and is certified to use.

How much time will the whole procedure take?

There will be 10 trials and each trial will last 2 minutes, so the experiment will last approximately 20 minutes. There will be breaks in between where the subject can move around, stretch legs and talk, drink water, etc. The total length of the whole procedure could last 30-40 minutes.

Do I need to prepare for the recording?

There is no preparation required for the recording, but we cannot record from people wearing hijab or other head covering, with hair extensions, weave, thick plaited hair, or hair styled using wax, hair spray or similar product.

What do I have to do?

Sit on a chair and speak the words, objects and numbers presented on the screen loud and in mind.

What are the possible disadvantages and risks of taking part?

Not much can go wrong by sitting on a chair and speaking. There are no disadvantages or risks in taking part, other than devoting forty five minutes of your time. Headset will be placed on participants head and transfer signal through the USB port of a PC or laptop; so there is practically no risk of electrical shock.

What if something goes wrong?

The ethical guidelines and procedures put in place ensure that there is very little that can go wrong.

Will my taking part in this study be kept confidential?

All data collected will be anonymised and can be deleted upon request before **31 January 2020**. The data may be later compiled as a research database and shared with researchers outside Brunel University London. Beyond that point, you will not be able to request that your data are deleted from the database. It is currently believed that the specific information to be captured under the recording protocol cannot be used for diagnosis of medical conditions that would otherwise remain undiagnosed. It is currently believed that the captured information is not discriminative enough to be used for identification among general public.

What will happen to the results of the research study?

The result of the study may be presented in scientific journals and/or conferences.

Who is organising and funding the research?

At present, research is undertaken as part of PhD project. So currently there is no formal internal or external funding arranged. The project does not have external funding.

Who has reviewed the study?

The study has received ethics approval from Brunel University, according to universities ethic approval policy.

Brunel University is committed to the UK Concordat on Research Integrity.

The University seeks to ensure that good practice in research is an integral part of its research strategy and associated policies. This code states that the general principle of integrity should inform all research activities. Honesty should be central to the relationship between the researcher, the participant and other interested parties.

Contact for further information and Enquiry.

For further information on the research study please contact Dr Nikolaos Boulgouris (Nikolaos.Boulgouris@brunel.ac.uk) or Sahil Datta (Sahil.Datta@brunel.ac.uk).

Participants Consent Form

PARTICIPANTS CONSENT FORM: Brain Signal Processing

The participant should complete the whole of this sheet him/herself		
	YES	*NO*
Have you read the Research Participant Information Sheet?	☐	☐
Have you had an opportunity to ask questions and discuss this study?	☐	☐
Have you received satisfactory answers to all your questions?	☐	☐
Who have you spoken to?		
Do you understand that you will not be referred to by name in any report Concerning the study?	☐	☐
Do you understand that signals captured from your participation will be anonymised and may be later compiled and made available to researchers Outside Brunel University London as a research database of such signals?	☐	☐
Do you understand that you are free to withdraw from the study:		
• Any time during the experiment?	☐	☐
• Without having to give a reason for withdrawing?	☐	☐
Do you agree to take part in this study?	☐	☐

Signature of Research Participant:	
Name in capitals:	Date: