A LESS INVASIVE AND COMPUTATIONALLY EFFICIENT SILENT SPEECH INTERFACE USING FACIAL ELECTROMYOGRAPHY

Thesis

Submitted in partial fulfillment of the requirement for the award of degree of

DOCTOR OF PHILOSOPHY

by

ASIF ABDULLAH

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SRINIVASNAGAR, MANGALORE - 575025

DECEMBER 2023

DECLARATION

by the Ph.D. Research Scholar

I hereby declare that the Research Thesis entitled **"A Less Invasive and Computationally Efficient Silent Speech Interface using Facial Electromyography"** which is being submitted to the National Institute of Technology Karnataka, Surathkal in partial fulfillment of the requirement for the award of the Degree of Doctor of Philosophy in Electrical and Electronics Engineering is a *bonafide report of the research work carried out by me.* The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

.....

Asif Abdullah

Register No.177EE001, Roll No.177009

Department of Electrical and Electronics Engineering

Place: Surathkal

Date: 08/12/2023

CERTIFICATE

by the Ph.D. Research Scholar

This is to certify that the Research Thesis entitled **"A Less Invasive and Computationally Efficient Silent Speech Interface using Facial Electromyography"** submitted by Asif Abdullah (Register Number: 177EE001) as the record of the research work carried out by him, *is accepted as the Research Thesis submission* in partial fulfillment of the requirements for the award of degree of Doctor of Philosophy.

.....

Dr. Krishnan CMC
(Research Guide)

8/12/2023
.....

Dr. Dattatraya N Gaonkar
(Chairperson D.R.P.C/ H.O.D., EEE)

Acknowledgments

It is with utmost happiness and humility that I express my heartfelt gratitude and love to all those who made this achievement possible.

I would like to express my sincere gratitude to my PhD supervisor, Dr. Krishnan CMC, Assistant Professor, Department of Electrical and Electronics Engineering for his support and guidance. His constant motivation and help has come to my aid in every aspect of my journey as a research scholar. He moulded me not only into a better researcher, but also into a better person. He was the source of light and guidance during all the hours of darkness.

I am grateful to National Institute of Technology Karnataka (NITK) for giving me the valuable opportunity to carry out research in such a prestigious institution. I also thank the Ministry of Human Resource Department, Government of India for granting me the scholarship to do research.

I take this opportunity to thank my research progress assessment committee (RPAC) members Dr. Yashwant Kashyap and Dr. Partha Pratim Das, for their valuable guidance and positive feedback. I also thank the former HODs, Dr. Vinatha U, Prof. B Venkatesaperumal, Prof. K.N. Shubanga, and Prof. Gururaj S Punekar, for providing necessary facilities and resources in the department to pursue my research. I would like to thank the present HOD, Dr Dattatraya N Gaonkar, for his support in carrying out the research successfully. I thank all the faculties of the Department of Electrical and Electronics, for their valuable suggestions and support. I would like to extend my grattitude towards the non-teaching staff of the department, especially, K.M. Naik, Santhosh, Karunakar B S, Basavarajaiah, Ratnakar, Suneetha, Akshatha, Nithin, Naveen, and Ramesha for giving the required assistance in conducting the research work and associated activities.

I am truly indebted to Pratap Kumar, Nisha K.S, Manish C.P, Sibeesh P.P, Abhishek Kumar, and Vishnu P.S for all their help and support in the process of data acquisition for my work. I also thank my friends Vishnu, Nisha, Vaishak, Jyothi, Francis, Aswathy, Syam, Meghna, Sibeesh, Roopa, Aalam, Anvit, Vikas, Teena, and Rashmi who made my stay at NITK enjoyable. They motivated and supported me during the entire process.

The path of PhD was not an easy one for me as I faced multiple obstacles throughout this journey. Sometimes I became too tired and exhausted physically

and emotionally. It was my vibrant child Aidan and my loving wife Thouseem who took all my pain away. Without them this would not have been possible. I am also grateful to my parents Abdullah Koya, and Jeeja, and my uncle Sadiq, for their love and support throughout my life. I also thank my brothers Azhar, and Arfaz for their love and affection towards me.

Finally, I thank everyone who has directly or indirectly supported me during this journey to make this dissertation possible.

Asif Abdullah

Abstract

Silent Speech Interface (SSI) is one of the promising areas of Human Computer Interaction (HCI) research. The Surface Electromyography (SEMG) based SSI is a technique where the electric activity of facial muscles are used to detect speech. The existing SSI techniques use computationally expensive methods and complex machine learning algorithms for the identification of silently uttered speech. The increased computational expense prevents real time implementation of SSI models especially in cost efficient applications such as communicative assistance for laryngectomy patients. Thus the objective of this research work is to develop a less complex and computationally less expensive SEMG based SSI model with superior accuracy. To achieve this goal, investigations are done on many feature extraction methods to check if they are suitable for SEMG based SSI. Detrended Fluctuation Analysis (DFA) is found to be promising for the recognition of silent speech using SEMG. The use of computationally less expensive classification algorithms was envisioned in this research work to develop a simpler and faster SSI model. The research identified K Nearest Neighbors and Decision Trees as suitable pattern recognition algorithms for this work.

The number of channels associated with SEMG based SSI is also a matter of important concern. A state-of-the-art model uses seven channels of SEMG data for the recognition of silent speech. Considering the use of some unipolar electrodes along with the bipolar ones, the number of electrodes to be accommodated on the face usually ranges from eight to twelve. For practical applications this is a high number especially in the case of medical conditions faced by laryngectomy patients. Too many number of electrodes on the subject's face creates inconvenience to the user who have undergone laryngectomy. It can hinder facial movement and can also contribute to the occurrence of cross talk between different facial muscles. Thus the reduction of number of channels is necessary and hence it is included as an important objective of this research work. The effectiveness of using DFA for successful channel reduction is investigated thoroughly. The analysis using DFA is also compared with channel reduction performed on models that employ existing state-of-the-art methods.

The availability of reliable data is vital for every researcher to carry out fruitful research. But as far as SEMG based SSI is considered, data availability is a major concern. There are very few reliable data sets (with sufficient vocabulary) available for SEMG based research. This is primarily due to the popular research

orientation towards acoustic speech recognition. Thus the creation of an extensive database is a promising aspect to consider and the initial steps to that cause is also considered as an important goal of this research work. Hardware purchase and assembly, drafting of a detailed data acquisition methodology, and a sample data collection is done as part of the work.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

ADC	Analog to Digital Converter
ASR	Automatic Speech Recognition
ANN	Artificial Neural Networks
BCI	Brain Computer Interface
DAQ	Data Acquisition
DFA	Detrended Fluctuation Analysis
DNN	Deep Neural Networks
DT	Decison Trees
DTW	Dual Tree Wavelet
ECG	Electrocardiography
ECoG	Electrocorticography
EEG	Electroencephalography
EGG	Electroglottograph
EMA	Electromagnetic Articulography
EMG	Electromyography
EOS	Electro Optical Stomatography
EPG	Electropalatography
HCI	Human Computer Interaction
HMI	Human Machine Interface
HMM	Hidden Markov Models
HOS	Higher Order Statistics
IAV	Integral of Absolute Value
ICP	Inductive Conformal Prediction
IEMG	Integrated Electromyography
KNN	K Nearest Neighbors
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficients
LSTM	Long Short Term Memory
MFCC	Mel Frequency Cepstral Coefficients
MPF	Mean Power Frequency
MRI	Magnetic Resonance Imaging
OPG	Optopalatography
PCA	Principal Component Analysis

PMA	Permanent Magnetic Articulography
PPF	Peak Power Frequency
RAM	Random Access Memory
RF	Random Forests
RMS	Root Mean Square
RNN	Recurrent Neural Networks
RW	Random Walk
SAX	Symbolic Aggregation Approximation
SEMG	Surface Electromyography
SNR	Signal to Noise Ratio
SSI	Silent Speech Interface
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TDFV	Time Domain Feature Vector
TDPSD	Time Dependent Power Spectrum Descriptors
US	Ultra Sound
USB	Universal Serial Bus
WAcc	Word Accuracy
WER	Word Error Rate
ZCR	Zero Crossing Rate

Chapter 1

Introduction

Contents

1.1 Overview

This chapter provides an introduction to the research work elaborated in this thesis. The chapter begins with much needed background information on the research topic which is then followed by the detail of the exact research area in focus. The motivation for pursuing this research topic is discussed in the subsequent section and it is helpful for understanding the various perspectives in this research domain. It is then followed by the descriptions regarding the thesis structure which presents the entire thesis flow at a glance, thereby enhancing readability. The chapter is then concluded by a summary of the important points discussed.

1.2 Background

Communication is a vital part of human progress. The human speech can be considered as an efficient means of communication that eventually led to the development of so many languages. It started with some basic alarm sounds to alert fellow humans about an impending danger or to convey information about a prey/predator etc. Then it slowly developed into more complex forms and through time, evolved into much more complex languages with a large vocabulary and grammar associated with it. Most of the present day languages are improvised or evolved versions of certain basic, ancient languages. So they share a common structural pattern in their formation. Thus human speech can be regarded as one of the most sophisticated and reliable forms of interaction. Its enhanced efficiency, information richness, and ease of use have invited considerable attention in Human Computer Interaction (HCI) (Preece et al., 1994) research. Thus the human speech has the capability of coding highly dense information, thereby having the potential to make computer control effortless and natural. Figure 1.1 is intended to provide a general idea of an HCI system. Speech recognition (Malik, Malik, Mehmood, & Makhdoom, 2021) is an actively researched area when it comes to HCI research. Recognition of speech can serve a wide variety of applications ranging from healthcare, automobiles, education etc. to more sophisticated areas like military, spacecraft, and robotics. The emergence of big data and deep learning is revolutionising the area of speech recognition. These recent developments are facilitating a paradigm shift from conventional speaker dependent speech recognition models to speaker independent models.

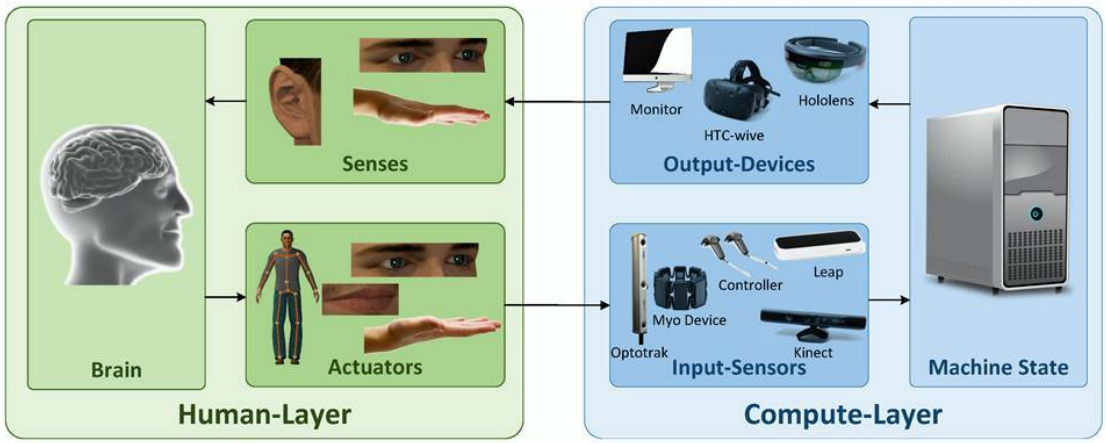


Figure 1.1: Human Computer Interaction (Source: (Bachmann et al., 2018))

In the area of speech recognition, the major research works till now are primarily focused on speech identification using acoustic data (Alharbi et al., 2021). Even though this is the most straight forward and easiest method of speech recognition, it has some limitations when it comes to certain applications where the presence of acoustic sound is absent. For example, in space we cannot always rely on detecting acoustic sound due to the absence of a sound conducting medium. Similarly detecting sound under water is also very difficult. In military applications, there are situations where the presence of audible speech is not desirable. All these situations demand a speech recognition model without the use of acoustic sound. But one might wonder that what is meant by speech when there is no sound associated with it. So that brings us to the terminology of "silent speech" where the facial movements and associated muscle movements are present, but no sound is produced by the vocal chords. We all know about "whispered speech" where facial movements are present along with a high pitched hiss. A silent speech can be characterised by only facial muscle movement and not even a whispering sound is present.

1.3 Silent Speech Interface

A Silent Speech Interface (SSI) model can be defined as a speech recognition tool where silent speech (Denby et al., 2010) is involved instead of acoustic speech that we normally see in an Automatic Speech Recognition (ASR) (Yu & Deng, 2016) model. Thus it can also be called a voiceless communication interface or electronic lip reading. In the previous section, a brief introduction about the various applications of silent speech recognition was mentioned. A detailed description regarding various existing SSI methods is presented during literature review given in the next chapter. However there is a major application to this technology from a humanitarian perspective. Silent speech identification can revolutionise the life of speech impaired people.

A brief description about speech impairment occurring due to laryngectomy is presented below which is then followed by the method envisioned in this research to address the problem. Finally the main goal of the research work is explained before moving forward to the next section where the motivation of pursuing this research is elaborated.

1.3.1 Speech Impairment due to Laryngectomy

Laryngectomy (Ceachir, Hainarosie, & Zainea, 2014) is a surgery in which the larynx of a person is removed. It is usually performed on patients with laryngeal cancer. Larynx is an organ that is essential for the production of sound. When the efforts to preserve the larynx fails or when the progression of the cancer hinders its normal functioning, a laryngectomy becomes essential. There are two types of laryngectomy, total laryngectomy and partial laryngectomy. In a total laryngectomy procedure, the whole larynx is surgically removed. This includes the hyoid bone, vocal folds, epiglottis, cricoid, and thyroid cartilage and some of the tracheal cartilage rings. In the case of partial laryngectomy only a certain diseased portion of the larynx needs to be removed. Even in case the entire larynx is removed, the facial musculature is not affected. They are able to move their facial muscles exactly like that of a normal person, and the only limitation will be the absence of sound. Thus what they are doing can be termed as 'silent speech'. The anatomical changes followed by a laryngectomy procedure is depicted in Figure 1.2.

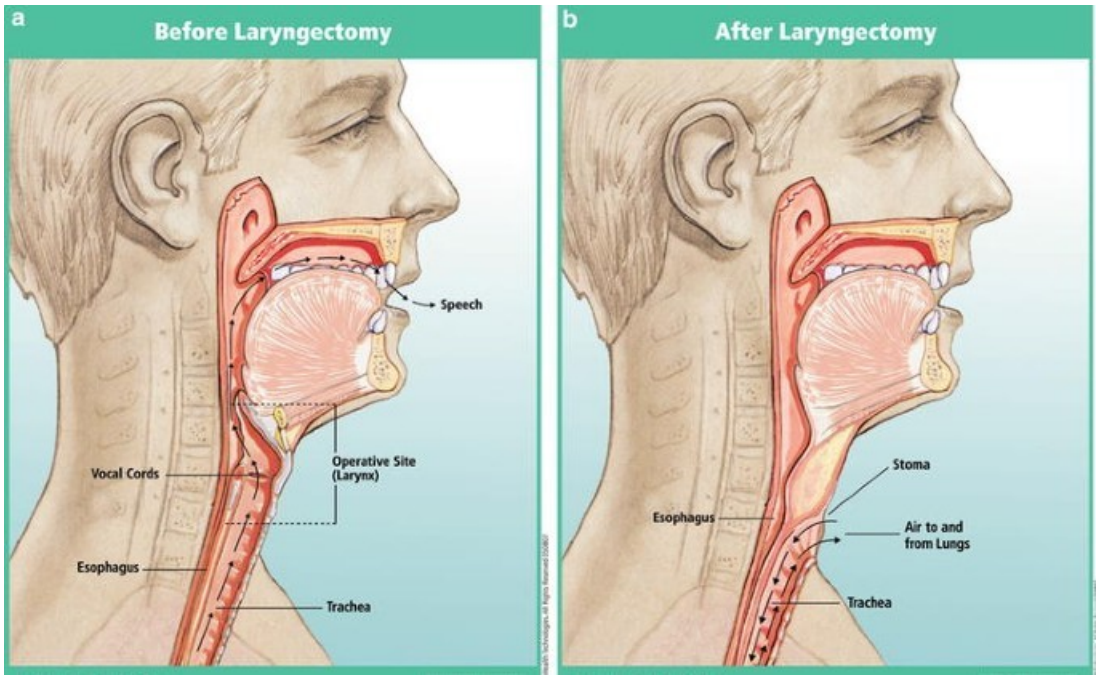


Figure 1.2: Anatomical changes following a laryngectomy (Source: (Rush, 2013))

The fundamental idea is to aid in their basic communicative requirements by identifying the words uttered. The possibility and scope of phoneme based classification is also investigated in this research work. The identification of words uttered can in turn help to identify a sentence from a fixed set of sentences. There are various assistive devices available in the market for laryngectomy patients. But

almost all of them require articulatory movement with sufficient force so that the assistive device could transform it into speech (Kapila et al., 2011). This would be a tiring exercise for people and can cause further problems. In the case of some other devices, the use of a hand is required along with the device for speech production. All these are active challenges faced by researchers in this area.

1.3.2 Surface Electromyography based Silent Speech Recognition

Surface Electromyography (SEMG) (Criswell, 2010) is a technique that involves the detection, recording, and analysis of the electric activity of muscles. When the central nervous system of the body activates any muscle fibers, the generation of tiny electric currents takes place in the form of ion flows. The movement of these currents generate potential differences due to the resistance offered by the body tissues, and can be measured on the skin. It is basically a non-invasive procedure where a single/array of electrodes are placed on the surface of the skin, on the muscles to be analyzed. The signal obtained from the SEMG setup is analyzed using a computer by extracting relevant features, which may either be time domain or frequency domain or both. At present, the SEMG technology is used on certain occasions as a tool to diagnose neuromuscular disorders (Meekins, So, & Quan, 2008). It is also used to evaluate the requirement of surgical procedures in patients with low back pain (Ambroz, Scott, Ambroz, & Talbott, 2000), and to help in assessing the prediction of disorders associated with muscle lesions. SEMG is also actively used in the area of myoelectric based prosthesis control (Powar & Chemmangat, 2019), (Powar & Chemmangat, 2020), which has huge potential due to the relative ease of the technology as compared to its counterparts. It has also been helpful in tracking the outcome of rehabilitation programs and assess the function of muscles in many day to day activities and sporting events (Vigotsky, Halperin, Lehman, Trajano, & Vieira, 2018).

Figure 1.3 shows the block diagram representation of a silent speech recognition model using SEMG. The acquired EMG signals from face are pre processed and then appropriate features are extracted from the signal. These features are then used by the classification algorithm to recognize the pattern of each of the words under consideration. The performance of the model is evaluated using the accuracy obtained during the testing of the model. The word recognition accuracy is the deciding factor for the investigation of better features so that the performance of the model can be further improved. This research work uses the data already

acquired and readily available with the research team. The methods elaborated in this work starts from the data pre processing stage to the performance evaluation stage. The hyper parameters associated with the classifier are optimized and if the performance is still poor, then further investigations are done to obtain better features.

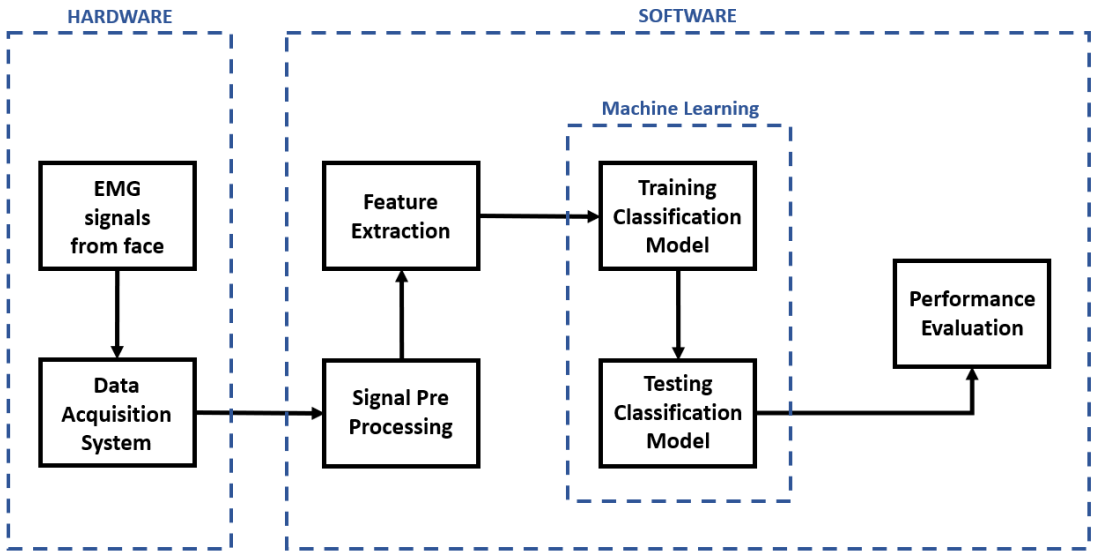


Figure 1.3: SSI block diagram

It was envisioned that the models developed have to be computationally less expensive and complex machine learning algorithms and deep learning techniques are not used. The process of improving the word recognition accuracy without using deep learning techniques requires an extensive investigation of data handling. The direct use of the raw data obtained from EMG sensors may not aid us in successful word recognition. Hence the use of an appropriate method for pre-processing the data and the extraction of relevant features from the pre-processed data plays a crucial role in improving the word recognition accuracy. The accuracy obtained using state-of-the-art features were seen to be deteriorating when channel reduction techniques were applied. So the objective of finding better features was not just limited to the improvement in accuracy, but also to the consistency in maintaining a satisfactory level of accuracy when channel reduction methods were implemented.

The entire work was done using the data developed by the researchers of the Interactive Systems Labs, University of Karlsruhe named as EMG UKA data corpus (Wand, Janke, & Schultz, 2014). The entire corpus is the result of their research work going on for more than a decade now and it is periodically updated

with new data. The creation of such a database is the vision of every research team. So as an initial step to that goal, the hardware setup with detailed data acquisition methodology and a sample data collection was included as a part of this research work. The hardware setup, methodology, and sample data collection of video and audio data acquisition was also included as part of this work so that it could be beneficial for future research prospects.

1.3.3 Word Data from Sentence Utterances

Identification of the words that are uttered in whole sentences (Wand et al., 2014) is one of the major goals of this research work. This is slightly different from the conventional word recognition models that use individual utterances of words under investigation. Pattern recognition of word utterances that are taken from sentences is more challenging due to several reasons. The accuracy of word alignments (in the case of sentence utterances) for silent speech data is relatively lower as compared to acoustic speech data. Subsequently this affects the overall word recognition accuracy of the model. In the case of independent word utterance (Zhang et al., 2020) data, no word alignment is required and hence accuracy will be better. Another reason is the pronunciation issues arising from the utterance of a whole sentence as compared to independent utterances of words. The relative speed of uttering a word is also higher when it is uttered as part of a sentence. All these factors contribute to the difficulty in identifying silent speech words when they are uttered as whole sentences. However the research was focused in this direction to address the real life scenarios.

1.4 Research Motivation

The fundamental motivation of pursuing this work is the benefit it can bring in the areas of biomedical signal based research. A particular signal and a particular methodology applied in a specific application area does not entirely confine only in that domain. This is true especially in the area of biomedical signal analysis and associated research. The suitability of a particular signal or approach can be successfully applied in another research area that may or may not have mutual similarities. An active investigation through the literature provides enough examples of situations where a specific approach in one research area demonstrated significant performance in another area. Hence the investigation of various signals, feature extraction methods, and pattern recognition techniques can be made

useful in a wide variety of research areas.

Recent trends in silent speech recognition is focused on brain signals and associated interfaces that can directly generate speech when a person articulates the speech in his/her mind. This is slightly different from other electrode based silent speech recognition methods. Given this fact, pursuing the EMG based speech recognition path was indeed a hard decision. But the insights from the literature survey provided the much required motivation to pursue this direction, so that the findings related to computationally less expensive methods can very well be adapted to other advanced methods (like brain interfaces) also in the future. Active research is also happening using visual methods and image processing techniques, but for a common application like in the case of laryngectomy patients, such advanced computing environments may not be a practical solution. This is because, in such applications the focus is not to design a precise model consisting of a large vocabulary, but to satisfy the basic communicative requirements of such voice challenged people. All of these factors provided the much required solid motivation for pursuing techniques that could perform well using limited computation resources.

1.5 Thesis Structure

The whole thesis is organized into six chapters as follows. The outline of the thesis is also depicted in the Figure 1.4.

Chapter 1: A brief introduction to human computer interaction, silent speech interface, laryngectomy, and surface electromyography is presented in the chapter. This gives a general idea regarding the research domain, application area, and the method adopted. It is followed by the motivation for choosing the research topic and the method. A brief outline of the thesis is also presented before concluding the chapter.

Chapter 2: An elaborate literature review on the state-of-the-art techniques in the area of silent speech interface and surface electromyography is presented. The detailed methodology of a silent speech recognition model as per the information obtained from the literature is then discussed. This includes information regarding SEMG signals, potential feature extraction methods, and candidates for pattern recognition. The identification of research gaps is then reported, which is followed by the formation of research objectives.

Chapter 3: Investigation of different feature extraction techniques that were attempted during this research is discussed in this chapter. The successful tech-

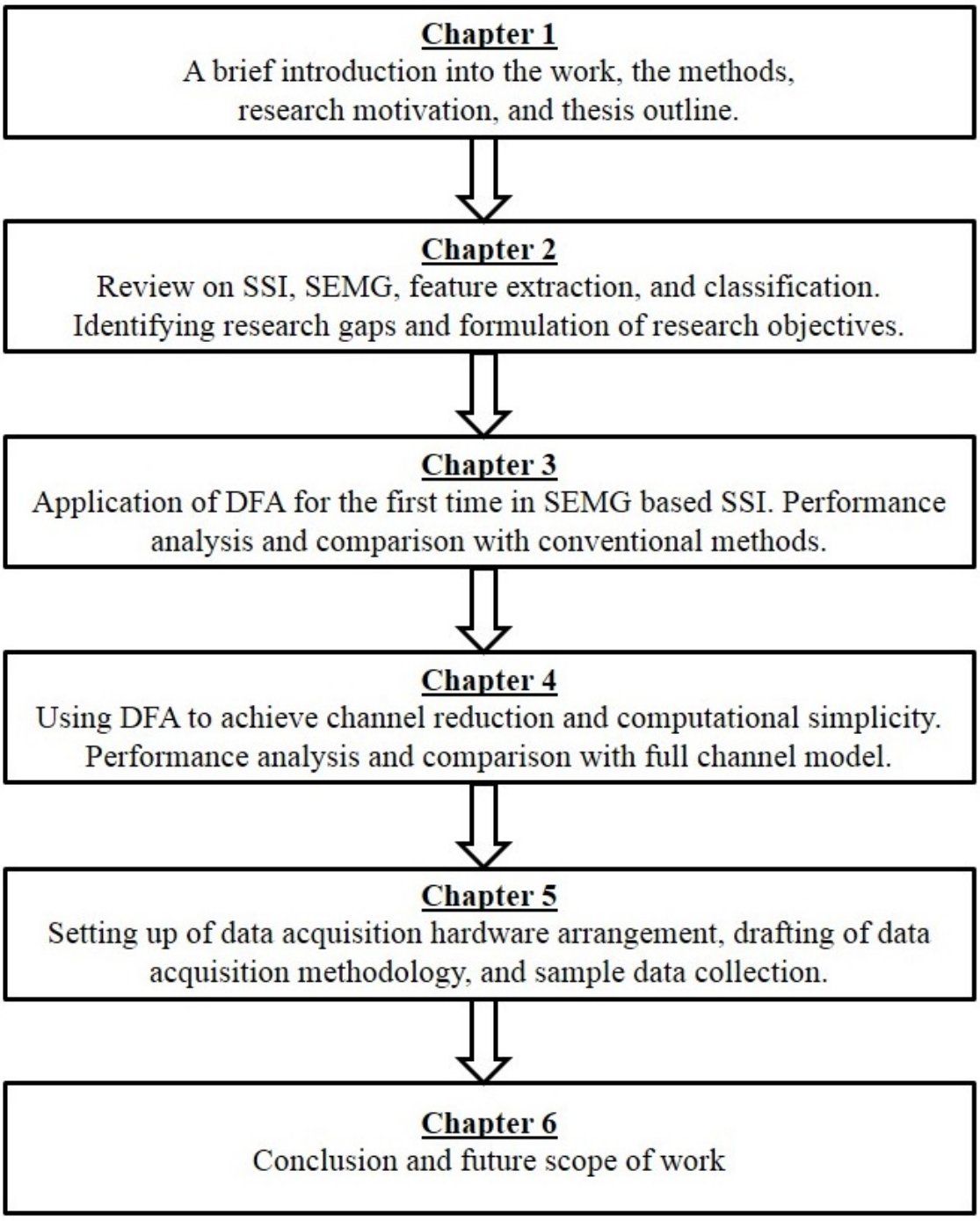


Figure 1.4: Outline of the thesis

nique is described in detail and the methods that failed are also included for the benefit of future research. The results obtained from the successful model is presented and the comparison with conventional methods is performed. The reason for the superior performance of the successful feature extraction technique is presented towards the end of the chapter.

Chapter 4: Various possibilities of achieving channel reduction is addressed in this chapter. The basic idea is to simplify the model in terms of user friendliness and hardware simplicity, while reducing the overall computational expense. The results of applying channel reduction in the model developed in this research work is compared with the results of applying channel reduction in the state-of-the-art model is presented. The contribution of novel ideas in feature extraction methods for achieving channel reduction are demonstrated in the chapter.

Chapter 5: The basic hardware requirements for SEMG based silent speech data acquisition is described in this chapter. The data acquisition methodology for successfully recording SEMG data along with the video and audio is discussed. The information regarding a sample set of data that was collected in lab is also included.

Chapter 6: This chapter presents the valuable conclusions obtained from this research work. The challenges existing at present and the scope for future research in the area is also discussed.

1.6 Summary

The objective of this chapter was to provide a basic understanding about silent speech identification and the method adopted for it. Information regarding the basic application area (i.e. laryngectomy patients) was important to get a perspective regarding the motivation behind the research work. It was also important to present a basic outline of the thesis in order to accustom the readers with a route map of the work, which is provided in the last part of this chapter.

Chapter 2

Literature Survey and Research Objectives

Contents

2.1 Introduction

This chapter presents a detailed literature survey associated with the research elaborated in this thesis. It is intended to give a clear idea about the state-of-the-art techniques in silent speech recognition and all related aspects. Existing techniques in silent speech recognition, their drawbacks, and the motivation behind the selection of SEMG based speech recognition is discussed initially. This is followed by a review on SEMG signals, its characteristics, and preprocessing methods. Detailed descriptions on the major application areas of SEMG, as identified from the literature, is then presented to get an overall idea about the possible benefits and challenges of the technology. Technical aspects related to the handling of SEMG signals, and extraction of features from them is also reviewed. An elaborate review on potential classification methods that are computationally less expensive, are also presented in this chapter. In any research endeavour, the proper identification of research gaps is of paramount importance and it is done in this work as well. Finally the chapter is concluded with the formulation of research objectives which are obtained as a result of the identification of research gaps.

2.2 Existing Techniques in Silent Speech Interface

There are multiple methods that are available for designing a model that facilitates silent speech communication. These comprise of visual methods, impedance plethysmography using electroglottograph, capturing the articulator motion using magnetic articulography, electropalatography, brain activity detection, and recording muscular activity of the face. A discussion on each of these methods regarding how they are applied, their advantages and disadvantages, possible modifications etc. needs to be done in order to understand more about the motivation of this research.

2.2.1 Visual Methods

As the name indicates, silent speech identification based on visual methods require the input of video data. Various image processing techniques are then employed to extract useful features so as to use them for pattern recognition. Limited vocabulary, higher data acquisition complexity, computationally expensive modelling, and poor accuracy are inherent drawbacks of these methods. As far as speech is considered the visual data contains far less cues for successful classification when compared with audio data (Potamianos, Neti, Gravier, Garg, & Senior, 2003). Hence the visual only methods will have a limited vocabulary. Out of several visual methods for speech recognition, two most popular visual features reported in the literature are 'appearance based features' (Potamianos et al., 2004) and 'shape based features' (Petajan, 1984). The drawback of these methods is that they follow a 'tailored for all needs' approach and such models are not suitable for visual based speech identification, because of the wide variability in the way different people speak. This is further aggravated when we cross cultural and national boundaries. Thus speech recognition using visual methods is a complex and computationally expensive process.

Ultrasound Imaging along with Lip Video

Ultrasound (US) imaging is a promising technique that can be used to gather direct information regarding the configuration of vocal tract. It is a clinically safe and non invasive procedure that enables the real time analysis of the tongue, which is one of the prominent articulators of the speech generation system. An ultrasound transducer can be placed under the subject's chin thereby enabling a

partial view of the surface of the tongue in the mid sagittal plane. (Ouisper, 2006-2009) developed an US imaging setup along with a video camera positioned facing the subject's lips. Thus the features obtained from this data (visual observations obtained from the tongue and lips) are purely non acoustic. These features were used to develop a silent speech identifier called as 'silent vocoder'. The 'US + lip video' method does not require any glottal excitation or airflow in vocal tract, hence it is suitable for laryngectomy patients. (Hueber et al., 2010) used video images of lips and US images of the tongue for developing an SSI model. The schematic of the model is given in Figure 2.1. Principal Components Analysis (PCA) was employed for image coding and Hidden Markov Model (HMM) was used to recognize phonetic targets. The challenges associated with the US + lip video model is the higher hardware requirement for data acquisition and the need for image processing methods for feature extraction.

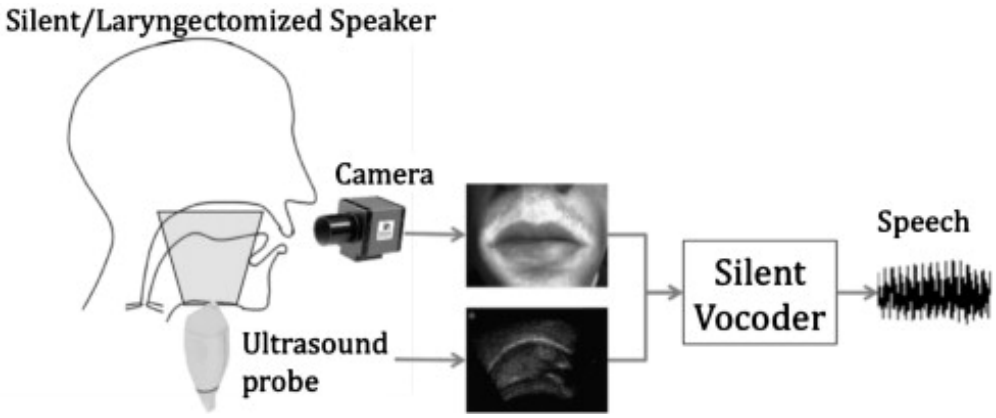


Figure 2.1: Schematic of the US + Lip video model (Source: (Hueber et al., 2010))

2.2.2 Impedance Plethysmography

This method usually uses the help of a device called electroglottograph (also known as aryngograph). During voice production, an electroglottograph (EGG) measures the degree of contact (in terms of contact area and force) occurring between the vocal cords when they vibrates. Thus an electroglottograph can be said to follow the principle of impedance plethysmography. (Dikshit & Schubert, 1995) employed an electroglottograph as a parallel source of information in addition to speech. A vocabulary of 64 words that were separately uttered was used for the research. Two independent artificial neural networks were realized where one of them used only speech data and the other used both speech and electroglottograph data. The network that used both speech and electroglottograph data gave an accuracy im-

provement of 5% than the other. The main challenge associated with this method is that the performance is reduced when it is used for silent speech recognition since the power associated with the vocal cord vibrations varies in a silent speech scenario as compared to acoustic speech. It is also important to note that the study used Artificial Neural Networks (ANN), which is a data hungry algorithm that requires a lot of data to give accurate results. The use of a large amount of data can create computational expense in addition to the challenges associated with the data acquisition.

2.2.3 Magnetic Articulography

Magnetic articulography refers to the process of measuring the movement and position of different points inside and around the mouth using electromagnetic induction. There are two major methods to apply magnetic articulography, Electromagnetic Articulography (EMA) method, and Permanent Magnetic Articulography (PMA) method. Both these methods and associated findings are presented in this section.

Electromagnetic Articulography (EMA)

EMA method (Schönle et al., 1987) uses magnetic tracers that are affixed to the articulators in order to sense their motion. The tracers generate magnetic fields and the variations in the magnetic field is measured to sense the articulator motion. Receiver coils are attached to the prominent articulators of the vocal tract. The articulatory activity is then monitored by an external equipment which is connected to these coils via wires. Transmitter coils that create alternating magnetic fields are placed near the head of the subject. Thus the spatial position of the magnetically coupled receiver coils can be easily tracked. Figure 2.2 shows the setup and the Magnetic Resonance Imaging (MRI) of EMA sensors placed on the tongue. In 2011, (Heracleous, Badin, Bailly, & Hagita, 2011) employed EMA for automatic recognition of phonemes without using additional audio data. The research was conducted on a French data corpus. (Kim, Cao, Mau, & Wang, 2017) employed bidirectional Recurrent Neural Networks (RNN) in 2017, to obtain long range temporal dependencies in the articulatory movements. Also data driven and physiological normalisation methods were used for speaker independent silent speech interface. Silent speech EMA data was obtained from 12 healthy subjects and 2 laryngectomy subjects. All of them were English speakers.

The EMA method has the advantage that the modelled articulatory dynamics

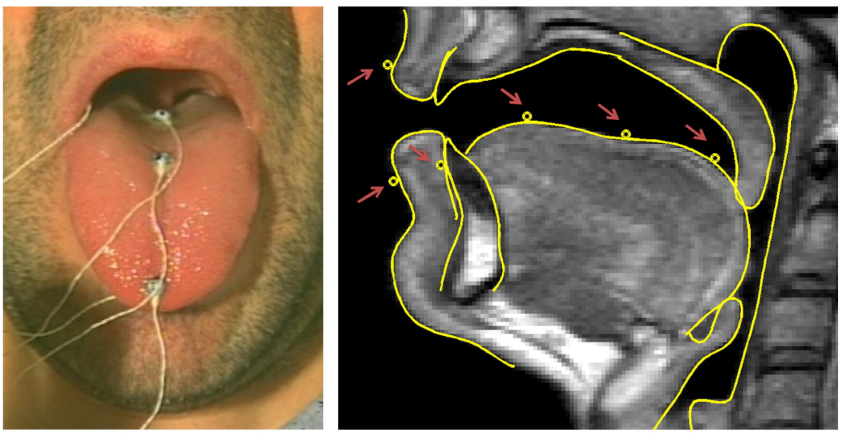


Figure 2.2: (a) EMA receptors attached to the tongue (b) MRI image of the positions of 6 active coils (Source: (Hueber et al., 2010))

will have higher temporal resolution and the requirement of feature pre processing is also minimal. The main disadvantage of this method is that it is an invasive procedure which necessitates wires running inside the mouth. It also necessitates the requirement of external transmitters that are not portable with the sophisticated connections. Thus its use is limited to laboratory experiments.

Permanent Magnetic Articulography (PMA)

The locations of the sensors are reversed in the case of PMA as compared to EMA. The transmitters are permanent magnetic and they are affixed to the articulators and the sensors to measure the magnetic field are located outside the mouth. The resultant field is the superposition of all transmitter magnetic fields from which the articulators' spatial position is decoded using sophisticated analysis. The PMA method requires the placement of permanent magnets inside the mouth which is devoid of any connecting wires and hence it is comparatively better for the subject than EMA. However the procedure is still invasive and can cause discomfort for subjects especially for laryngectomy patients. Also in PMA, the data is less explicit thereby requiring more pre processing for the correct recognition of articulatory movements and positions (Gilbert et al., 2010).

2.2.4 Palatography

The method of sensing the location and timing of tongue contacts during speech production, using electrodes positioned on inside the palate of the mouth is known as Electropalatography (EPG) (Sebkhi et al., 2019). The information regarding

the articulation of various phones is provided by the pattern of palatal contacts. A similar method is the Optopalatography (OPG), where optical distance sensors are employed to track the position of the tongue without the requirement of explicit contact with palate (Wrench, McIntosh, Watson, & Hardcastle, 1998). Most of the studies using these methods have been done in the areas of phonetic research and speech therapy. In 2008, (Toutios & Margaritis, 2008) adopted a data driven technique to map speech signal into EPG data by using PCA and Support Vector Machines (SVM). In 2011 (Birkholz & Neuschaefer-Rube, 2011) used a new technique called Electro Optical Stomatography (EOS), that combines the merits of both EPG and OPG, to obtain the information regarding the articulation of consonants and vowels. The method was later employed for the recognition of vowels using tongue contours and EPG patterns as features for a DNN based classifier (Birkholz & Neuschaefer-Rube, 2011). An extension of this research was done by (Stone & Birkholz, 2016b), (Stone & Birkholz, 2020) for recognising command words in German. In 2014 a multiple linear regression model was developed by (Mumtaz et al., 2014) for the reconstruction of tongue contour using EOS data. An error is developed if the orientation of the tongue is not perpendicular to the axes of the optical sensors, which is a major issue with OPG. In order to avoid this error, a model was suggested by (Stone & Birkholz, 2016a) that uses light propagation for arbitrary source-reflector-detector arrangements which takes into account, the complex reflective attributes of the surface of tongue due to sub surface scattering.

2.2.5 Brain Activity Detection

The technique of brain activity detection involves capturing of associated bio signals at the very origin of speech generation. It possess a big advantage that a wide variety of speech pathologies and disorders can be effectively addressed. Sensing the brain activity can help those with apraxia (Dressing et al., 2019) or dysarthria (Gupta et al., 2021) or even certain instances of aphasia (Fridriksson, Bonilha, Baker, Moser, & Rorden, 2010) in addition to assisting people with voice disorders. The brain activity can be sensed using two approaches in general - invasive techniques and non invasive techniques. Invasive techniques refers to the use of implants in the brain's speech motor cortex while non invasive techniques refers to use of non invasive electrodes placed on the scalp.

Electroencephalography (EEG)

EEG refers to the measuring of brain's electrical activity with the help of electrodes positioned non invasively on the scalp. The signals sensed at different electrode locations are the outcome of the activation of millions of neurons simultaneously. The summed voltage of these activations flows through the brain, skull, and the scalp layers (Nunez & Srinivasan, 2006). Schematic of EEG generation is represented in Figure 2.3. The electrophysiological activity of the brain is thus measured through EEG and it has effective temporal features to sufficiently characterize the neural process associated with speech generation. The major challenge of EEG is its high susceptibility to myoelectric activity, other movements, and environmental artifacts. This causes interference in EEG recordings obtained during the production of speech (Goncharova, McFarland, Vaughan, & Wolpaw, 2003). Despite these demerits in speech related research, EEG is still the most popular method used in Brain Computer Interface (BCI) (Wolpaw, Birbaumer, McFarland, Pfurtscheller, & Vaughan, 2002) technology. Despite the challenges, some efforts were made to employ EEG to synthesise speech. In one study, EEG based speech recognition was employed in an experiment where subjects either listened to pre recorded speech or read out sentences aloud, and achieved similar results in both cases (Krishna, Tran, Han, Carnahan, & Tewfik, 2020). The investigation of silent speech to text using EEG faced the same drawback of low spatial resolution. Initial attempts in this regard was made by (Suppes, Lu, & Han, 1997) in a small vocabulary. Investigations on EEG based silent speech synthesis using syllables (D'Zmura, Deng, Lappas, Thorpe, & Srinivasan, 2009) and phonemes (DaSalla, Kambara, Sato, & Koike, 2009) were also attempted, but on a very limited dataset comprising of just 6 syllables and 3 phonemes respectively.

Electrocorticography (ECoG)

When the brain's electrical activity is measured invasively using electrode implants on cortical surface, it is termed as electrocorticography. This is mostly viable on patients with severe epilepsy, where there is a need to implant electrodes temporarily for pre surgical planning or for intra operative monitoring (Crone, Miglioretti, Gordon, & Lesser, 1998). The electrode grids that are thus implanted can remain there for a time span of many days to 2 weeks. During this time the patients do consent to undergo scientific experiments. Significant progress have been made in recent years using ECoG due to its better spatiotemporal resolution. In 2020, (Makin, Moses, & Chang, 2020) achieved word error rates as low as 3%

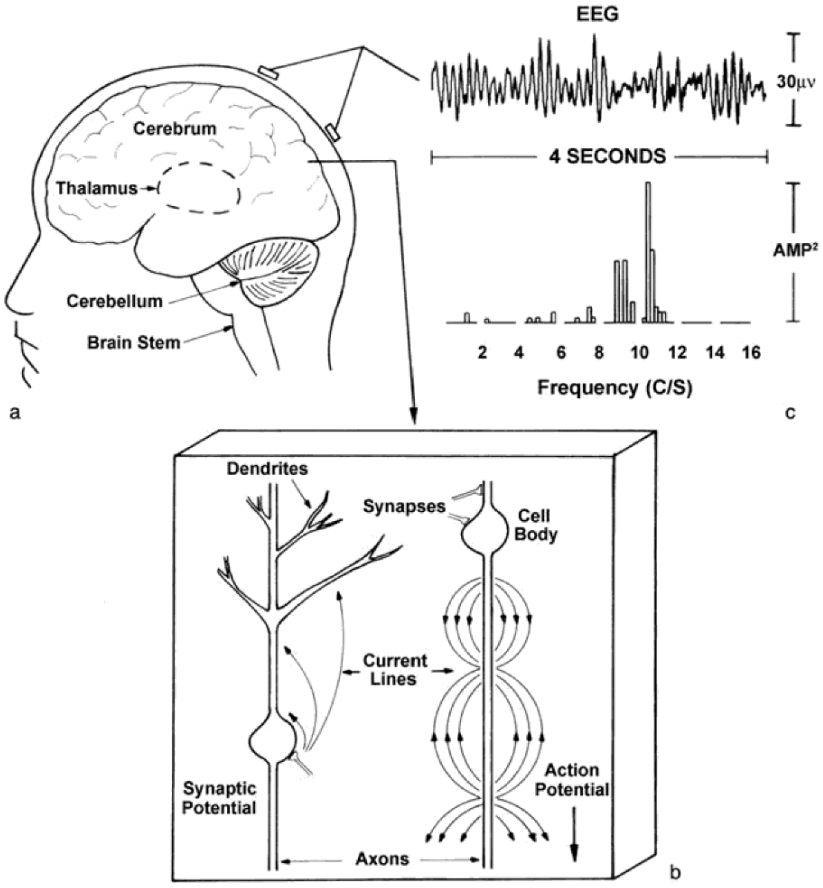


Figure 2.3: (a) Human brain (b) Generation of micro currents due to synaptic and action potentials in the cerebral cortex (c) A sample EEG signal and its power spectrum (Source: (Nunez & Srinivasan, 2006))

for an acoustic speech recognition model using ECoG. The major drawback of this method is the invasiveness of the procedure. It is also important to note that most of the studies employed either complex machine learning algorithms or deep learning techniques for pattern recognition.

2.2.6 SEMG based Silent Speech Recognition

In 1985,(Sugie & Tsunoda, 1985) employed EMG for silent speech identification for the first time in literature. They used 3 SEMG electrodes to distinguish between Japanese vowels, and presented a pilot model that demonstrated this work in real time. During almost the same time, (M. S. Morse & O'Brien, 1986) published their initial work in the same research area. They analysed the EMG signals generated from the muscle activity occurring in the head and neck to classify 2 words. The study was extended in the following years that lead to the identification of 10 words uttered separately (M. S. Morse, Day, Trull, & Morse, 1989), (M. Morse, Gopalan, & Wright, 1991). They obtained an accuracy of 70% on a 10 word vocabulary but the model performance showed a dramatic reduction for a slightly larger word count. The accuracy reported was only 35% for a vocabulary of 17 words and hence the result was not comparable with traditional silent speech identification methods. A better performance was offered by (A. D. Chan, Englehart, Hudgins, & Lovely, 2001) by achieving an accuracy of 93% on a vocabulary comprising English digits. Chan also combined an EMG based model with a traditional system to achieve superior performance in removing the ambient noise (A. D. C. Chan, 2003).

The application of EMG for the recognition of sub auditory speech was performed by (Jorgensen, Lee, & Agabont, 2003) in 2003. A set of 6 English words were uttered by 3 subjects of age 24, 35, and 55. The data acquisition was done with only two pairs of EMG electrodes which were positioned below the jaw and on the larynx. Each subject was made to record 100 repetitions of each word over a span of six days. For feature extraction, short time fourier transform and dual tree wavelet coefficients were used. A total of 5 neural network algorithms were tested for performance and a scaled conjugate gradient based network was finally chosen for pattern recognition. A word recognition accuracy of 92% was obtained using the dual tree wavelet feature based model.

While initial studies in the area used less number of words, the research began to gain pace in early 2000s by improving the size of the vocabulary. In 2006, (S.-C. Jou, Schultz, Walliczek, Kraft, & Waibel, 2006) achieved 70% accuracy in a 100

word vocabulary developed using a single speaker. Successful use of bigger vocabularies requires the breakdown of words into components of sub-word units such as phonemes or syllables. (Jorgensen & Binsted, 2005) performed the classification of consonants and vowels using phonemes as units, while (Walliczek, Kraft, Jou, Schultz, & Waibel, 2006) analysed different units derived from continuous speech with a hundred word vocabulary. In 2007 (S. Jou, Schultz, & Waibel, 2007) used articulatory features for augmenting phoneme units and in 2010 (Schultz & Wand, 2010) elaborated the incorporation of context dependant phonetic features which helped in enhancing the the accuracy to 90% for the hundred word vocabulary. This was a case of speaker dependant silent speech recognition. In 2009 (Wand & Schultz, 2009) reported the initial studies on speaker adaptive and speaker independent silent speech recognition based on SEMG which was done on a large corpus of EMG data acquired from a total of 78 speakers who read sentences in both silent as well as audible mode.

By the second decade of the twenty first century, researchers began actively looking for better pattern recognition techniques with the rise of AI and the recent works in this research domain is motivated by the advances in deep learning methods and associated technologies. Hybrid systems that used EMG along with Deep Neural Networks (DNN) for silent speech recognition have been developed by (Wand & Schmidhuber, 2016). In 2020 (Y. Wang et al., 2020) reported that the idea of transfer learning was observed to be fruitful for recognising EMG based silent speech by using neural networks whose training was done on an image classification task. More recently, (X. Wang et al., 2020) conducted an empirical study to find out the impact of the number of channels in SEMG based silent speech recognition. In recent years, direct synthesis of speech from EMG signals (Janke & Diener, 2017), (Diener, Janke, & Schultz, 2015), (Diener, Herff, Janke, & Schultz, 2016) made remarkable progress due to the developments in deep learning and array based EMG sensors. A specific merit of EMG when compared to other methods for sensing articulator motion is that the sensing of EMG signals can be done approximately 60 ms before the actual motion of the articulator. This advantage helps in the realization of real time models with low latency (Diener et al., 2016), (Diener & Schultz, 2018) thereby minimising the delay between articulator movements and synthesised acoustic feedback. The influence of different system parameters like the size of training data, size of DNN, frame shift etc. on the quality of speech was investigated by (Diener & Schultz, 2018). The research was conducted using a real time direct speech synthesis model and the analyses was done with the use of objective quality metrics.

Even though considerable progress has been made, SEMG based SSI still face many challenges that are under active investigation. A major issue is the heavy dependence of the models on the training session. The use of array based electrodes can reduce this effect since the relative location of the electrodes remains constant. But still there are considerable differences between data recorded in different sessions (Janke & Diener, 2017), (Diener, Felsch, Angrick, & Schultz, 2018). (Wand & Schultz, 2014) suggested an unsupervised adaptation method to address this problem that allowed the incorporation of new data with each recording session. They employed domain adversarial training method to adapt front end of the SEMG based SSI model to the target session data in an unsupervised way. The issues arising from data discrepancies are usually dealt by deep learning algorithms through extensive computation using a large corpus of data. The discrepancies that are unable to accommodate are treated as outliers and are excluded from pattern recognition. The loss of such data does not create much of a problem for the algorithm when there is a large amount of available data. However a more scientific approach would be to study about such discrepancies, its causes, and methods to overcome them. The investigation on better data processing methods, and feature extraction techniques can open up solutions to such challenges.

The comparison of all these methods based on their advantages and disadvantages are tabulated in Table 2.1 for an easy analysis.

2.3 Surface Electromyography (SEMG) Signals

The details regarding the choice of an SEMG based silent speech recognition model have been explained so far in this chapter. Now it is necessary to provide an elaborate discussion on the various aspects of SEMG signals, including their generation, characteristics, and pre processing.

2.3.1 Myoelectric Signal Generation

A muscle in the human body is made up of several of its smallest subdivisions called motor units. The nerve cell body, axon, and a number of fibres constitute a motor unit. The number of motor units in a particular muscle depends on the muscle type and it generally varies between 2 to 2000. The brain has a part called the motor cortex which generates a series of nerve impulses whenever a muscle

Table 2.1: Comparison between different SSI methods

SSI Method	Advantages	Disadvantages
US + Lip Video	non invasive, glottal excitation is not required, suitable for laryngectomy patients	video/image processing required, complex hardware setup
EGG	non invasive, more accurate for acoustic speech recognition	less accurate for silent speech recognition, requires classifiers like ANN
EMA	higher temporal resolution, minimal feature preprocessing	invasive with wires inside mouth, requires external transmitters
PMA	comparatively less invasive than EMA,	invasive, data is less explicit, more preprocessing required
EOS	suitable for computationally less expensive classifiers	invasive, complex hardware setup
EEG	non invasive, effective temporal features	high susceptibility to signal interference
ECoG	better spatiotemporal resolution than EEG	highly invasive, needs surgical implants, requires deep learning methods
SEMG	non invasive, suitable for computationally less expensive classifiers	too many electrodes on face, requires better feature extraction methods

contraction is required. Depending on the required degree of contraction, the simulation of appropriate number of motor units takes place. Upon the reception of the impulse train, the motor units produce motor unit trains, the summation of which is known as myoelectric signal. Figure 2.4 gives a concise idea about the process.

2.3.2 Electrodes

EMG signals are acquired using electrodes which are positioned over the muscles to be analysed. There are two types of EMG electrodes, surface electrodes and needle electrodes. The former is a non invasive method while the latter is invasive which requires medical skill for electrode placement and can cause discomfort and pain to the subject. The data used in this work uses surface electrodes to acquire the EMG signals required and hence the method is termed as Surface Electromyography (SEMG). Ag/AgCl standard gelled electrodes with a diameter of 4 mm is employed and the electrodes have a circular recording area. The diagram of electrodes is given in Figure 2.5.

EMG signal acquisition can be done using electrodes affixed on the face either by positioning them over particular muscles (Maier-Hein et al., 2005) or by arranging them in a grid as an array (Wand et al., 2013). The signal can be expressed as the difference in potential between 2 electrodes. The measurement can be done either in the monopolar configuration (active versus reference) or in the bipolar mode (active versus active). The EMG recording setups currently in use is not much user friendly and hence they are less practical in real life scenarios. For example, the number of electrodes to be accommodated on the face is more, and it needs to be firmly affixed to the skin for a long duration. Manabe et.al. addressed this problem by designing ring shaped electrodes which can be wrapped around any two fingers and the thumb (Manabe, Hiraiwa, & Sugimura, 2003) (Manabe & Zhang, 2004). Whenever there is a requirement to capture the signal, the subject has to press the fingers against the face in a specific manner. This is a promising method in developing a model for mobile interface, which can be utilised in both noisy as well as silent environments.

2.3.3 SEMG Properties

The difference in potential occurring between 2 electrodes is measured by using differential amplifiers and it also helps in eliminating artifacts. The unamplified SEMG signal falls in the 2 - 5 mV range and hence it is usually amplified at a gain

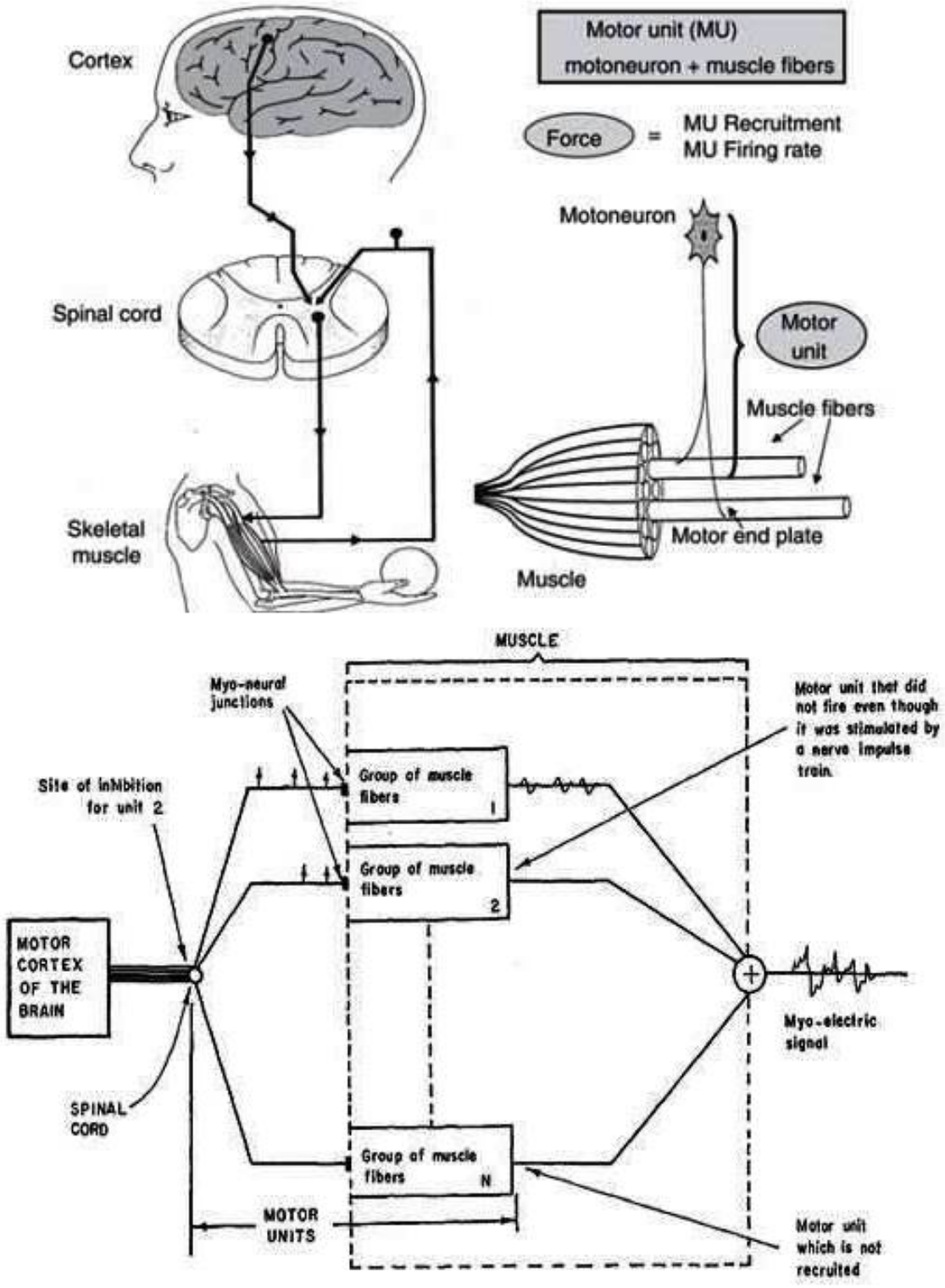


Figure 2.4: Myoelectric signal generation (Source: (Brody et al., 1974))

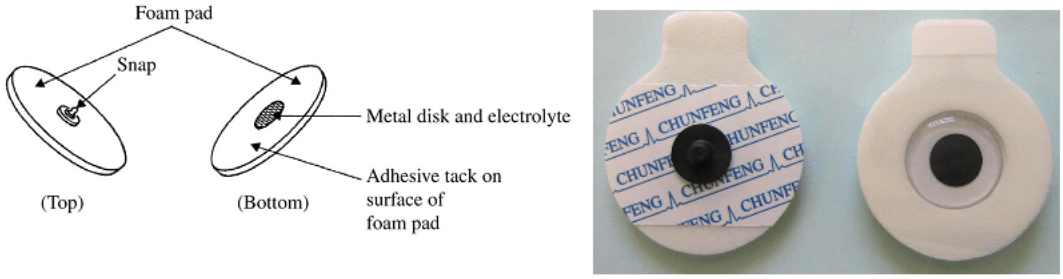


Figure 2.5: Ag AgCl Electrodes (adapted from (Webster, 2009))

ranging from 500 to 1200. Digital conversion of the signal is the next step. Analog to Digital Conversion (ADC) is generally done at a sampling frequency of 1000 Hz since most of the signal components happen to be in the range of 20 - 500 Hz. This is followed in accordance with the Nyquist criterion which states that the minimum sampling frequency needs to be twice as that of the highest frequency component in the signal in order to accomplish the successful reproduction of the signal.

2.3.4 SEMG Preprocessing

The acquired EMG signals contain different types of noise. The noise content in the signal can be attributed to various factors such as noise associated with the equipment, cross talk due to muscle overlap, environmental factors, and other interference. The first part of EMG pre processing consists of signal conditioning which is aimed at removing the noise content in the signal. The signal conditioning of EMG almost always uses digital filters in addition to the regular anti aliasing filters, The second part of signal pre processing consists of normalizing the EMG signal using an appropriate method. Overall the pre processing step is very helpful in enhancing the reliability and accuracy of the speech recognition model. The filtered and amplified SEMG signal and the corresponding normalized signal is given in Figure 2.6.

2.4 Applications of SEMG

SEMG has been useful in several medical applications, and Human-Machine Interfaces (HMI). These are the two areas where SEMG is popularly used. In medical field itself it has found application in a wide variety of sub domains like the diagnosis of neuromuscular diseases, prosthesis control, assistive devices for speech

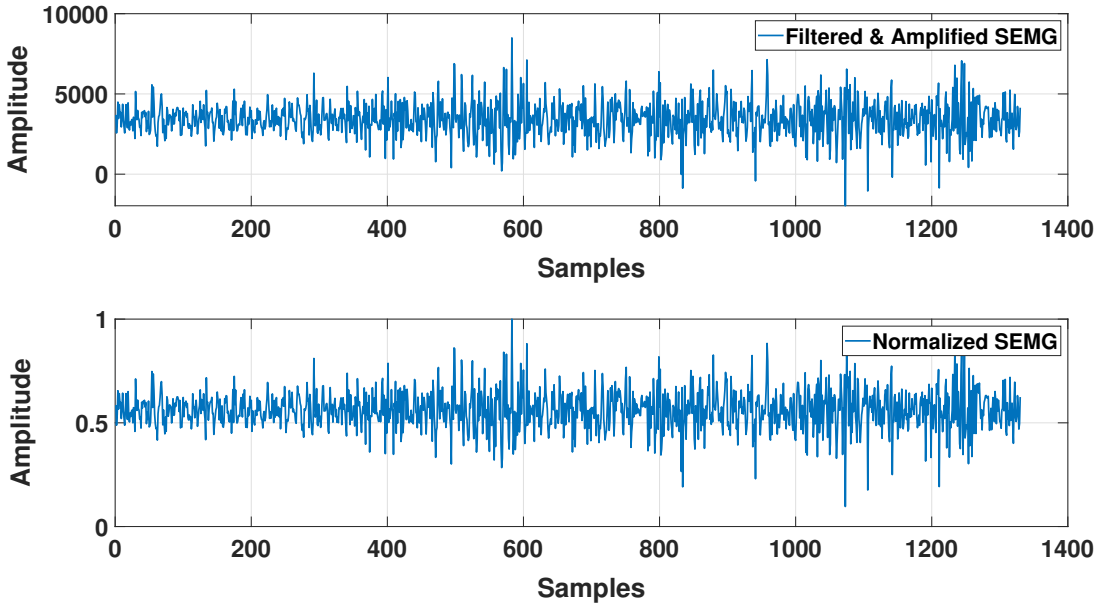


Figure 2.6: SEMG Normalization

restoration etc. This section describes the major areas where SEMG have been applied so far.

2.4.1 Medical Field

A major application of surface electromyography since its discovery is in the area of medicine. It has marked its presence in many specific medical applications by forming sub branches like electronystagmography (eyes), electrocardiography (heart), electroencephalography (brain), etc. A few instances of the application of SEMG in medical fields will be discussed in the paragraphs below.

SEMG is a prominent tool in the diagnosis and analysis of neuromuscular disorders. In 1982, (Muro, Nagata, Murakami, & Moritani, 1982) came up with the first work in this domain. They employed Integrated Electromyography (IEMG), Peak Power Frequency (PPF), and Mean Power Frequency (MPF) as the EMG parameters for the analysis of neuromuscular pathology. Since then considerable research has been done in this area. Lot of research groups have employed Muscle Fibre Conduction Velocity (MFCV) as an effective parameter for SEMG based neuromuscular research (Hilfiker & Meyer, 1984), (Ramaekers et al., 1993). Recently, (Chandrasekhar, Vazhayil, & Rao, 2020) developed an eight channel SEMG model to acquire real time data for different upper limb movements so as to assess the motor impairment. In addition to its application for neuromuscular disorders, this low cost portable system can also be used for patients undergoing post stroke

rehabilitation as well.

(Stepp, Heaton, Rolland, & Hillman, 2009) developed an electrolarynx that is controlled using EMG. Electrolarynx is a popular rehabilitative device for patients whose larynx are removed through a total laryngectomy procedure. A common drawback of an electrolarynx is that it lacks pitch control thereby requiring the constant use of a hand. EMG signals were used to regulate the pitch and the onset, offset controls of the device. The experiment used 7 Delsys 2.1 differential electrodes for SEMG signal acquisition from face and neck. Eight persons who underwent total laryngectomy articulated speech with the help of a normal electrolarynx as well as an EMG based electrolarynx. All the eight subjects displayed better speech performance when EMG from the electrode positions 4, 5 and 6 were used. These electrode locations corresponds to the ventral neck midline, submental midline, and corner of the mouth respectively. All subjects could articulate serial and running speech without using hands when the EMG based electrolarynx was used.

In 2010 (Stepp, Hillman, & Heaton, 2010) also demonstrated that SEMG can be used to diagnose hyperfunctional dysphonia which is also known as vocal hyperfunction. This is a usual condition that is associated with several voice disorders. Eighteen individuals with normal voice and 18 persons having vocal nodules were selected for their study. Myoelectric signals generated from the neck of all the above subjects were measured with 2 double differential electrodes (Delsys 3.1), positioned over the sternohyoid, omohyoid, and thyrohyoid muscles. The electrodes employed were beneficial in increasing spatial selectivity while minimizing cross talk.

The application of EMG in the research of finger braille was performed by (Miyagi, Nishida, Horiuchi, & Ichikawa, 2006) in 2006. Finger braille is a means of communication devised for the benefit of deaf blind and during that time it was quickest and accurate. Prosody in linguistics is the analysis of speech elements which are not actual phonetic segments but are characteristics of syllables or larger units of speech. It includes functions like tone, rhythm, and stress that can transmit the structure of a sentence, its prominence and the emotions associated with it. The objective of this work was to analyze prosody by measuring the electric activity of the muscles associated with the index, middle, and ring fingers while using finger braille. A finger braille interpreter was asked to type a given sentence during which the myoelectric activity was measured. The results indicated that the typing strength improves at the starting of a phrase as well as with a prominent phrase. Thus for the deaf blind, the hopes of using prosody in

the interpreter system was opened up.

An investigation into pattern recognition of SEMG applications was performed by (Khadivi, Nazarpour, & Zadeh, 2005). The objective of the work was to utilize Higher Order Statistics (HOS) in the feature extraction of SEMG for a successful classification problem that finds application in the control of upper limb prosthesis. The classification of 4 primitive motions of the upper limb was the intention. The subject in this study was a man aged 24 years and 2 pairs of electrodes (Ag/AgCl) were affixed on his triceps brachii and biceps brachii. The purpose of using HOS in feature extraction was to obtain higher classification rate as well as to increase robustness to interference and noise as compared to methods such as Integral of Absolute Value (IAV). The accuracy of classification obtained by using HOS+IAV was 90.70% as compared to 81.41% obtained using only IAV. Similar classification rates was obtained for other feature extraction methods, but only at the expense of higher computational complexity.

2.4.2 Human-Machine Interactions

EMG technology has proven to be a valuable approach in the development of HMI. It offers a novel method of control, replacing traditional means of communication such as mice and touch, and introducing a new way of interaction based on 'basic thinking'. Electromyographic signals serve as a means to capture a person's intentions, and these signals are typically obtained through EEG by placing electrodes on the scalp. However, this method is plagued by low Signal to Noise Ratios (SNR) and often necessitates multiple electrodes. As a solution to these issues, SEMG is being employed as an alternative to address and rectify these challenges.

(Larson, Terry, & Stepp, 2012) conducted the design and testing of a human-machine interface that utilized EMG to capture signals from a discreet location. They collected bilateral signals from the auricularis posterior muscle in five participants. These individuals underwent training to modulate their muscle activity, enabling cursor movement in two dimensions and producing different vowel sounds as target points. The success rate, measured by the participants' ability to accurately position the cursor on the target location within 15 seconds, was established as a percentage. The results indicate that participants achieved an average accuracy of 67% in learning SEMG control of vowel synthesis solely through auditory feedback. Furthermore, this skill demonstrated the ability to generalize to new vowel targets.

(Hashimoto, Takahashi, & Shimada, 2009) developed a nonverbal interface that

enables hands-free control of an electric wheelchair. The objective of this study was to identify specific gestures made by an operator to control the linear and turning motions, velocity, and steering angle of the wheelchair. The recognized gestures included jaw closure, forehead wrinkling, and left and right gaze. EEG, EMG, and EOG signals were obtained as input data through three dry electrodes embedded in a headband. To validate the system's functionality, an operator provided simple commands to the wheelchair using the gesture interface, resulting in corresponding movements. As a demonstration of the feasibility of wheelchair operation through the gesture interface, an indoor navigation experiment was conducted, covering a distance of 150 meters.

2.5 Windowing

The EMG signal is highly non stationary in nature which means that whenever a particular facial motion is performed, the signal shape varies. This happens even when identical words are uttered. The reason for this nature is due to the fact that even when identical words are spoken, the set of motor units recruited are different. Thus time domain windowing is a necessary step to overcome this problem. During feature extraction step, the number of samples or the time duration (W) considered for acquiring EMG data is termed as a window. The time required by the computer for the extraction of features and do pattern recognition of the signal is called computation time. The classification outputs cannot be made available instantaneously due to the computation time.

There are two types of windowing methods, overlapping window and non overlapping window. In an overlapping windowing method, two adjacent windows are overlapped and few samples appear in both windows. While in non overlapping windowing method, a set of samples will be unique to that particular window. There are also several windowing schemes depending on the type of window used, like rectangular window, hamming window, hanning window etc. An overlapping rectangular window has been employed in this research work.

2.6 Feature Extraction

Feature extraction refers to the process of identifying underlying structures that are hidden in any type of data. It also filters the interference and irrelevant data present in the signal.Feature extraction can also be termed as the conversion of

a signal pattern into a set of signal features. In a classical pattern recognition problem, feature extraction is the most crucial step that determines the overall performance and accuracy of the model. An appropriate windowing technique is employed during the computation of features to obtain the temporal characteristics from the EMG signal, while minimizing spectral leakage. The extracted features are supposed to possess less computational expense, robustness, and superior class separability.The extracted features can either be in time domain or in frequency domain. It can even be in time-frequency domain. A comprehensive literature review of various EMG features is presented below.

2.6.1 Time Domain Features

Time domain features are widely used due to the fact that there is no need for any kind of signal transformation and hence it is computationally faster to calculate(Westerink, Van Den Broek, Schut, Van Herk, & Tuinenbreijer, 2008), (Phinyomark et al., 2013). A time domain analysis often gives the transitory response of the system under investigation. It is also able to provide a better understanding regarding the flow of electrical and mechanical energies of the system. Usually, this includes structural variations in a system, wave propagation, and electric potential created by external sources.

- **Jou et al.** In 2006, (S.-C. Jou et al., 2006) drafted a set of six time domain features for the silent speech recognition model that employed SEMG signals. These features are still used as the state-of-the-art features in this research area. Nine-point double averaged SEMG, rectified nine-point double averaged SEMG, power of nine-point double averaged SEMG, power of rectified Nine-point double averaged SEMG, and zero crossing rate, are the features presented in this paper. The features were used in accordance with three contextual filters to generate new features. The contextual filters used in the work are delta filter, trend filter, and stacking filter.

- **You et al.** The Root Mean Square (RMS) value of SEMG was used as features by (Yau et al., 2008) in 2008. The RMS of SEMG is considered to be associated with the count of active muscle fibers as well as the activation rate of the muscle fibers. Hence it can be considered as an ideal measure to evaluate muscle activation strength. The research work also recommended the integration of the RMS in order to address the problem of changes in speed and pronunciation of vowels.

- **Wang et al.** Symbolic Aggregation Approximation (SAX) can be defined as an effective method for data transformation that is popularly used in the analysis of time series data. In 2013, (J. Wang et al., 2013) used symbolic representation of the time series silent speech data as features for pattern recognition. The data used was acquired using an electromagnetic articulograph which is another method widely used in silent speech recognition. SAX also helped them in efficient dimensionality reduction as well.

2.6.2 Frequency Domain Features

Analysis in frequency domain is the process of representing a signal or mathematical function with respect to the frequency content of the signal, instead of time. A frequency domain plot shows the amount of signal existing in a particular frequency range while a time domain plot shows only the signal variation over a specified time span.

- **Meltzner et al.** (Meltzner et al., 2017) used Mel Frequency Cepstral Coefficients (MFCC) as the baseline technique in feature extraction for the identification of silent speech uttered by post laryngectomy subjects. A hamming window having suitable window length and window shift adapted for each subject, was employed for feature extraction. The next step was cepstral analysis that led to the generation of a cepstral feature set with 7 dimensions. This was followed by the computation of delta cepstral features and both these were concatenated to obtain the final feature set.

2.6.3 Time-Frequency Domain Features

- **Jorgensen et al.** (Jorgensen et al., 2003) used time-frequency domain features in 2006, for recognising sub auditory speech. Short Time Fourier Transform (STFT), and coefficients of Dual Tree Wavelet (DTW) were employed to generate necessary features for speech recognition.

- **Maier-Hein et al.** In 2005 (Maier-Hein et al., 2005) used windowed STFT to compute required features for the recognition of non audible speech using SEMG. A total of 18 coefficients were evaluated in which, the delta coefficients obtained from STFT formed the initial 17 and the 18th one consisted of the average value of the time domain data points in the particular observation window. Normal STFT coefficients, RMS values, Linear Prediction

Coefficients (LPC), and cepstral coefficients were also tried as features but they could not contribute much to the model performance.

- **Phinyomark et al.** (Phinyomark, Phukpattaranont, & Limsakul, 2012) employed the fractal analysis technique for identifying low level muscle activations when EMG signals are used. The method is also known as Detrended Fluctuation Analysis (DFA). The work was aimed at identifying upper limb movements, but it is important to note that silent speech articulation is also associated with low level muscle activation.

2.7 Pattern Recognition

The feature extraction step is followed by pattern recognition of the signal. In this case an appropriate classification method is used for identifying the word uttered. The investigation of various potential classifiers is presented in this section. The objective in mind was to choose the ones with reduced computational expense and less model complexity.

2.7.1 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a popular machine learning algorithm that requires considerably less computation. It is a non parametric classification technique where a sample is assigned a particular class by the plurality votes of its neighbours. In KNN classification method, local approximation of the function is done during training and all computation is performed during testing. The use of normalized data can greatly improve the performance of the algorithm and the consistency increases with the quantity of data. These features of KNN prompted its use in this research work. (Powar, Chemmangat, & Figarado, 2018), and (Çerçi & Temeltaş, 2018) employed KNN for classifying EMG signals generated from the movement of various muscles in the human hand. It was also used by (Ma et al., 2019) to classify a set of ten silently uttered Chinese words using surface electromyography. (Chatterjee, Pratiher, & Bose, 2017) applied KNN to segregate EEG signals using multi-fractal DFA as the feature set.

2.7.2 Decision Trees (DT)

Decision trees (DT) (Rokach & Maimon, 2008) perform classification using a very simple mechanism. A tree of different values as nodes are developed by the al-

gorithm concerning the grouping of data. More nodes and branches are added according to the number of classes required. This is done using a mathematical index to identify the most suitable split criterion. This process is relatively simple as compared to other machine learning techniques and hence the computational expense associated with DT based classification is less. There are some other advantages of DT when it comes to speech recognition performed in this work. The processing of SEMG data poses some serious issues like the presence of outliers, missing samples, non-linear factors etc. DT based model is immune to these issues related to data. Splitting of data to develop a tree is not affected by missing data samples or outliers. This is because the splitting is not dependant on absolute values and instead it depends on the amount of samples present inside the split ranges. Also, it does not require the data to be always linear. DT was successfully implemented for SEMG based silent speech recognition previously in this research area (Abdullah & Chemmangat, 2020). (Povey et al., 2011) employed DT, along with other deep learning techniques, for speech recognition.

2.7.3 Random Forests (RF)

Random forests is a learning method that employs multiple algorithms to achieve better prediction results as compared to the results achieved by the individual use of any of the participating algorithms. This kind of learning is also known as ensemble learning method. Random forests construct several decision trees during training in order to perform machine learning tasks such as classification and regression. In a particular classification problem, the class chosen by the majority of decision trees is selected as the output class of the algorithm. In a regression model, the mean predicted value of individual decision trees is given as the output of the model. The problem of over fitting in the case of decision trees is generally addressed by random forests (Hastie, Tibshirani, Friedman, & Friedman, 2009). In 1995, Tin Kam Ho developed the first random forest algorithm (Ho, 1995) using a method called random subspace, (Ho, 1998) which is a technique to employ stochastic discrimination method for classification. Random forests are able to provide reasonable predictions for a wide variety of data with minimal configuration requirements and hence they are widely employed in business as black box models.

2.7.4 Long Short Term Memory (LSTM)

LSTM (Hochreiter & Schmidhuber, 1997) is a popular artificial neural network utilised in the area of deep learning and artificial intelligence. While standard neural networks use feed forward connections, an LSTM employs the feed back method. This kind of a neural network is known as a recurrent neural network and they have the ability to not only process individual data points (e.g. images), but also a whole series of data (e.g. speech). This property of LSTM makes it a suitable method for analyzing and predicting data. They are able to perform tasks like robot control, speech activity detection, machine translation, speech recognition, connected handwriting recognition etc. The name LSTM is obtained from the idea that a standard RNN possess both "short term memory" as well as "long term memory". The biases and weights associated with the network connections vary once per each training episode, which is similar to the mechanism of long term memory storage by the physiological variations in synaptic strengths. However the network's activation patterns do vary once per each time step, which is similar to the momentary variations of the brain's electric firing patterns associated with short term memory storage (Jeffrey et al., 1990). The architecture of an LSTM is aimed at providing a short term memory for the network that has the ability to last for thousands of time steps and hence the name "long short term memory". LSTM has been successfully used for the identification of silent speech that uses facial electromyographic signals (Janke & Diener, 2017).

2.8 Research Gaps

Based on an extensive literature survey, the following research gaps were identified.

1. At present there are several advanced technologies for the identification of silently uttered speech. Most of them use video/image processing techniques and/or employ deep learning algorithms for pattern recognition. Even the methods that use electromyographic or ultrasound or brain signal data also employ complex machine learning or deep learning techniques for classification. These are computationally expensive and necessitates the requirement of advanced high computing processors. So the lack of computationally less expensive methods for silent speech recognition is evident from the literature survey.

2. According to the literature, SEMG based speech recognition involves the

usage of multiple channels and subsequently an inconvenient number of electrodes needs to be placed on the subject's face. This contributes to both hardware as well as software complexity in addition to the discomfort caused to the user. Successful channel reduction strategies are very rarely reported in literature and mostly along with deep learning methods. So the challenge of channel reduction without using computationally expensive models need to be addressed.

3. One of the major challenges associated with SEMG based speech recognition is the unavailability of reliable data. There are only very few data sets available right now. This poses a serious challenge in pursuing research in this area. Developing an extensive silent speech data set with a decent vocabulary, is a task that requires several years and expertise of several researchers. At least a start towards that goal can be beneficial for researchers in the long run.

2.9 Research Objectives

A detailed literature review facilitated an in depth understanding about the computational complexity of the existing SEMG based silent speech recognition models that employ deep learning techniques. This helped in drafting the research objectives for a computationally efficient SEMG based silent speech recognition model which is presented below.

1. There are a fixed set of sentences that needs to be recognised using SSI model using SEMG. The sentence identification can be done if the words are recognised in the proper sequence by the classification algorithm. The identification of words can either be done using the method of direct recognition of the word or by recognising the phonemes that constitute the word. The objective is the development of a silent speech recognition model using either phoneme based approach or word based approach so that it can be used to identify the fixed set of sentences.

2. The conventional SSI setup that uses seven channels causes difficulty for the user (especially for laryngectomy patients) and at the same time it increases the computational expense. An investigation into methods that can reduce both the computational expense and the number of channels is important. The target is to find suitable methods to achieve channel reduction for the

silent speech recognition model. It is also envisioned to use computationally less expensive classification algorithms while maintaining sufficient word recognition accuracy.

3. The future research work in the area of SSI is dependant on the availability of reliable data. The setting up of hardware and data acquisition methodology is envisioned as an initial step of developing an extensive database. Hence this is included as an important objective of this research work. Hardware setup constitute of equipment purchase and assembly and it is not just limited to EMG. Video and audio recording setup is also employed for possible use in the future. The data acquisition methodology describes in detail about the steps to be followed for successful data acquisition. Collection of a sample set of data is also included in the objective.

2.10 Summary

The existing techniques for silent speech recognition, their merits, and demerits that has been reported in literature was discussed in detail in this chapter. An extensive survey of the methodology and applications of SEMG technology was also reported. This helped in identifying the research gaps in this area and hence effective formulation of the research objectives was possible, which is presented in the chapter.

Chapter 3

Fractal Analysis as Feature Extractor for Facial Electromyography

Contents

3.1 Introduction

The development of an accurate silent speech interface model depends on several factors like accuracy of data acquisition, choice of suitable data pre processing methods, selection of appropriate feature extraction techniques, and the ability of the pattern recognition algorithm employed. Out of all these factors, selection of appropriate feature extraction techniques plays a vital role in the overall success of the model since the performance of the classifier highly depends on the features input to it. If the extracted features are not appropriate the model performance deteriorates even if the classifier employed possess superior abilities. Thus it can be said that the feature extraction step acts as the backbone of the silent speech recognition model. This chapter describes in detail about the various research directions undertaken for finding a suitable feature extraction technique. The discussion is not just limited to the successful technique, but it also includes the failed methods that eventually became the stepping stones to success.

3.2 Feature Extraction

The SEMG data acquired cannot be directly fed to a classification algorithm since the raw data is unable to provide useful patterns for the classification algorithm. Appropriate features have to be derived from the raw SEMG data and these features are then fed to an intelligent machine learning algorithm to recognize words. Hence feature extraction refers to the process of investigating more about the salient characteristics of the signal under consideration, in order to apply appropriate tools, such that a suitable mathematical manifestation of the signal is possible.

The major work done as part of this research work is in the investigation of suitable features. It started with an extensive literature survey through which possible candidates (feature extraction methods) were identified. The whole area of human computer interaction, speech recognition (audible as well as silent), and

areas that use similar signals (like EEG, ECG etc.) was covered in this process. Then these techniques were implemented to see if they yielded any positive outcome. In a research it is always important to mark the wrong approaches and report the failed attempts so that future researchers benefit from it. The failed techniques of this research work is presented before moving on to the successful one.

3.2.1 Time Dependent Power Spectrum Descriptors

The quantitative measures used to represent the spectral properties of a signal with respect to time is denoted as Time Dependent Power Spectrum Descriptors (TDPSD). It is used to obtain the information regarding the variation of energy or power in various frequency bands of a signal with respect to time. The methods like wavelet transform or short time fourier transform are usually employed to compute TDPSD. These descriptors are very much helpful in the analysis of signals associated with a wide range of domains like vibration analysis, biomedical signal processing, speech recognition, and audio processing, among others (Amini, Pedram, Moradi, Ouchani, et al., 2021), (Khushaba, Takruri, Miro, & Kodagoda, 2014). They provide valuable insights into the temporal and frequency content of the signal and also facilitate the extraction of relevant features. The wide use of TDPSD in literature prompted the investigation of its potential use in the area of silent speech recognition but it was unable to perform well in that area.

3.2.2 Mel Frequency Cepstral Coefficients

In automatic speech and speaker recognition, Mel Frequency Cepstral Coefficients (MFCC) are commonly used (Ittichaichareon, Suksri, & Yingthawornsuk, 2012). Davis and Mermelstein introduced them in the 1980s, and they have remained the state-of-the-art ever since. The main feature type for automatic speech recognition prior to the introduction of MFCC was LPC and Linear Prediction Cepstral Coefficients (LPCC). A major concept to grasp about speech is that the sounds generated by a person are filtered by factors such as the shape of their vocal tract, teeth, tongue, etc. It is the shape that decides what sound is coming out. If the shape can be determined correctly, then it can lead to the accurate identification of the word or phoneme that is being generated. The vocal tract shape is manifested in the short time power spectrum envelope, and MFCC performs the task of correctly representing this envelope. Even though MFCC has contributed substantially in the area of acoustic speech recognition, the usage of it in the area

of silent speech recognition was a failure. It couldn't help in the successful identification of either phonemes or words in silently uttered speech.

The successful feature extraction techniques as far as this research work is considered is presented in the following sub sections.

3.2.3 Time Domain Features

Research in the area of acoustic speech recognition is closely linked with frequency domain features. There are many features which have demonstrated their superiority when it comes to speech recognition based on audible sound. Mel Frequency Cepstral Coefficients (MFCC), mentioned in the previous section is an example of such a feature used for acoustic speech recognition. However they are seen to fail in the case of identifying EMG based speech recognition. The computational expense of frequency domain features is also more than that of time domain features. This was the motivating factor to investigate more about the time domain features that are useful in this research area. The various time domain features reported in the literature as state-of-the-art (S.-C. Jou et al., 2006) for silent speech recognition which are also used in this work are given below.

Normalized SEMG ($\bar{x}$)

If the basic SEMG signal is represented by $E(n)$, then Normalized SEMG is given by $\bar{x}$, where

$$x(n) = \frac{E(n) - min(E)}{max(E) - min(E)} \tag{3.1}$$

$\bar{x}$ = frame-based time domain mean of $x(n)$

Nine Point Double Averaged SEMG ($\bar{w}$)

Nine point double averaged signal $\bar{w}$ can be obtained from

$$w(n) = \frac{1}{9} \sum_{k=-4}^{4} v(n+k) \tag{3.2}$$

where

$$v(n) = \frac{1}{9} \sum_{k=-4}^{4} x(n+k) \tag{3.3}$$

$\bar{w}$ = frame-based time domain mean of $w(n)$

Rectified Nine Point Double Averaged SEMG ($\bar{r}$))

Rectified high frequency part of the signal is denoted by $\bar{r}$, where

$$r(n) = |x(n) - w(n)| \tag{3.4}$$

$\bar{r}$ = frame-based time domain mean of $r(n)$

Power of Nine Point Double Averaged SEMG (P_w)

Power of $w(n)$ is given by

$$P_w = \frac{1}{9} \sum_{k=-4}^{4} |w(n)|^2 \tag{3.5}$$

Power of Rectified Nine point Double Averaged SEMG (P_r)

Power of $r(n)$ is given by

$$P_r = \frac{1}{9} \sum_{k=-4}^{4} |r(n)|^2 \tag{3.6}$$

Zero Crossing Rate (z_x)

z_x = frame-based zero crossing rate of $x(n)$

3.2.4 Detrended Fluctuation Analysis

Detrended Fluctuation Analysis (DFA) (Phinyomark et al., 2012) is a feature that effectively use the characteristics of both time domain and time-frequency domain. In DFA method, the SEMG signal is first integrated to transform it into a Random Walk (RW). It is then split into rectangular windows of same size, without any overlap. For each window, a least square fit is calculated to illustrate the semi local trend of that window. The last step is to calculate the Root Mean Square (RMS) fluctuation of each of the windows to obtain the detrended time series.

The DFA feature can thus be given as:

$$F(i) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_i(k)]^2} \tag{3.7}$$

where the window number is denoted by i , N is the total signal length, and $y(k)$ gives the random walk transformation of the SEMG signal $x(n)$ and is expressed as:

$$y(k) = \sum_{n=1}^{k}[x(n) - \overline{x(n)}], \qquad k = 1, ..., N \tag{3.8}$$

where $\overline{x(n)}$ gives the average value of $x(n)$.

The DFA process is depicted in Figure3.1. The first part shows the SEMG signal and the second part shows the same signal integrated into a random walk conversion. The second part of the figure also shows equal sized windows and a least square fit applied to each of the chosen windows.

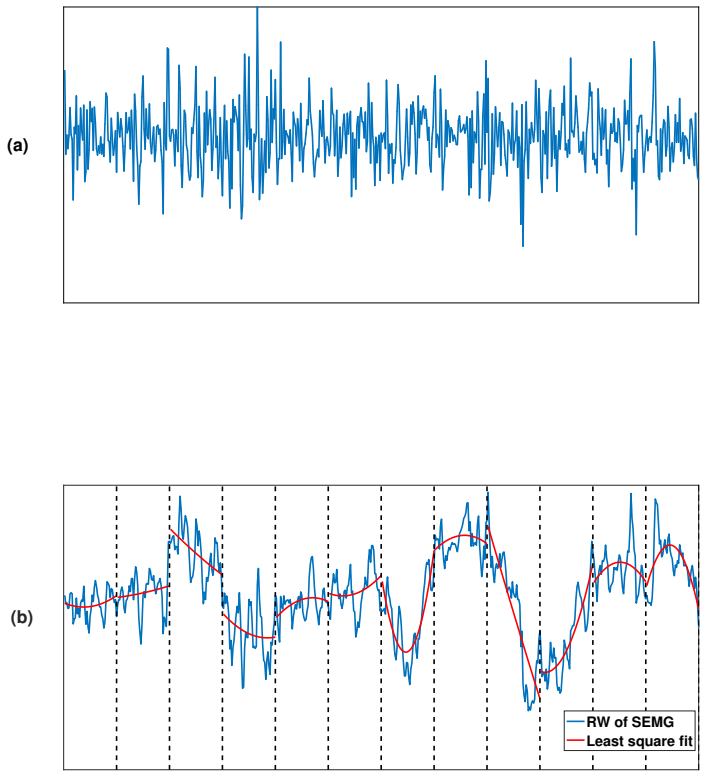


Figure 3.1: (a) SEMG signal (b) the integrated random walk conversion of the signal with different windows of size n and the least square fit applied for each window

3.3 Materials and Methods

The research materials and methods associated with the first objective is described in detail in this section. It consists of the data used, system architecture employed, and the methodology of approaching the objective.

3.3.1 Dataset Employed

The research work presented in this thesis uses the SEMG data developed by the researchers of the Interactive Systems Labs, University of Karlsruhe (Wand et al., 2014) named as EMG UKA data corpus. The data corpus is developed for assisting research associated with speech recognition and analysis. It contains SEMG data as well as acoustic data of speech performed by different people. The data acquisition was done in 3 modes - audible, whispered and silent - so that, the classification algorithm can get an insight into different force levels of human speech. EMG UKA corpus consists of a total of 7 hours and 40 minutes of data, which is further sub-classified into 63 sessions. The corpus consists of a total of 1 hour and 46 minutes of data in silent mode and a total of 1 hour and 47 minutes of data in whispered mode. The rest of the data is in audible mode. A total of 8 speakers have contributed to the corpus.

The acquisition of SEMG data was done with an electrode arrangement (Maier-Hein et al., 2005) comprising of 6 channels (excluding the reference electrode). The location of electrodes were on 6 major articular muscles namely: zygomaticus major, levator anguli oris, depressor anguli oris, platysma, anterior belly of the digastric and the tongue. Two channels (2 and 6) are bipolar and the remaining four channels (1,3,4, and 5) are unipolar. The positioning of electrodes is represented in Figure 3.2.

Varioport biosignal recorder developed by Becker Meditec, Germany, was used to perform the recordings. It is powered by a battery and also consists of an insulating device to isolate electric currents of the computer controlling the recorder and the amplifier. It has a 16 bit ADC, amplification factor of 1170, 0.9 - 295 frequency range, and a resolution of 0.033 mV/bit. Sampling of EMG signals were performed at a sampling rate of 600 Hz. The recording of acoustic data was performed using a close talking microphone that had connections to a stereo USB sound card and the sampling rate was 16 KHz.

The recordings were facilitated in generally quiet rooms without any provisions for electrical shielding in order to closely match real life scenarios. Thus the usage of professional recording studios were avoided. The supervision of all recordings

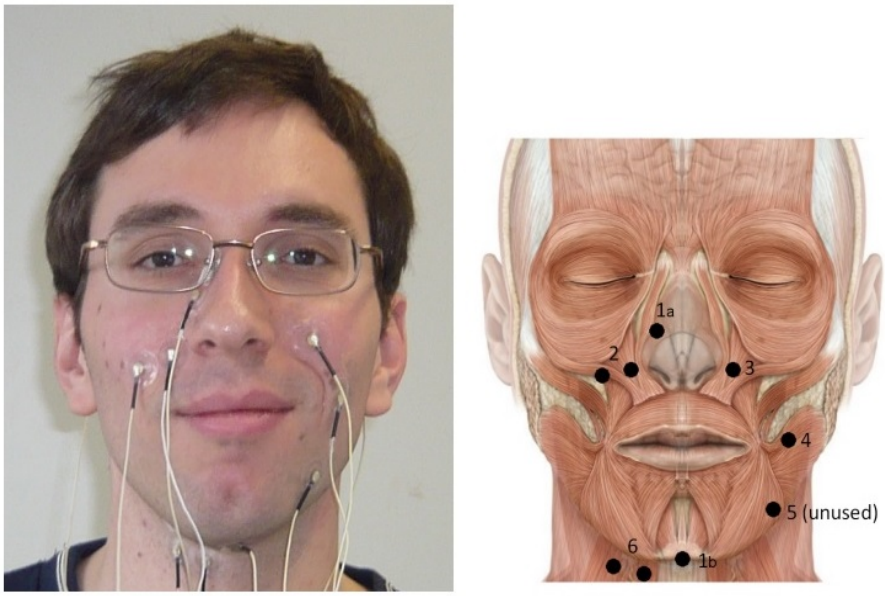


Figure 3.2: Positions of electrodes for EMG-UKA corpus (Maier-Hein et al., 2005) (facial musculature adapted from (Schünke & Schulte, 2006)). Channels are numbered, reference electrodes of unipolar channels 3, 4, and 5 (behind the ears) are not shown.

were done by professional recording assistants and the research team. All subjects who contributed to the data collection belonged to the university student population. Hence they were not necessarily native English speakers. However it was ensured that the pronunciation of the sentences were correct. The nature of the research was explained to the subjects before obtaining their written consent to use the data for future research and distribution. The data was also anonymized using neutral IDs to safeguard the privacy of the subjects.

The corpus consists of well-marked and documented data. There are separate markings for both phonemes and words, which facilitates the research for speech recognition in both phoneme based method as well as word based technique. This work follows the word recognition based technique for speech identification. A set of 1100 words is taken from the corpus for this work. The words were selected based on their number of repetitions in the database. Only those words having atleast 5 utterances in the whole database was selected. This was done so that the classification model could get enough data samples to train on that particular word. It was also made sure that the selection of articles, pre-positions, and post-positions were minimal. It is also important to note that the words used in this work are taken from the complete utterances of the sentences, using signal markers. Word based SSI implemented throughout the literature uses separate utterances

of each word and hence the data extraction errors will be minimum. The classifier performance will be better in that case. Since the work presented here aims at improving the word recognition accuracy of words taken from sentence utterances, such a dataset is selected.

3.3.2 System Architecture

The entire work including data processing, feature extraction, and classification is done using MATLAB software. It is also important to include the system architecture that is used for classification since a quantitative comparison in terms of computation time is presented. The overall performance and computation time of the model depends on the system architecture factors such as the processor, its clock frequency, and RAM capacity. The details of the processor used in this work for feature extraction and classification are provided below.
Processor : Intel(R) Core(TM) i7-4770 @3.40 GHz
RAM : 24.00 GB
System Type : 64-bit Operating System, x64-based processor

3.3.3 Methodology

The total features that are used in this work are described in detail in section 3.4. As mentioned in the section, six time domain features and one time-frequency domain feature is used in this work. There are a total of seven channels and hence the total number of features accounts to 49. Dimensionality reduction is reported widely in literature where a lot of features are extracted for use with deep learning techniques. In this research work, the number of features are optimal for the classifiers selected and hence dimensionality reduction is excluded.

Thus the Feature Matrix for evaluation can be denoted as TD7 and it is given as:

$$\mathbf{TD7} = [\bar{x}, \bar{w}, \bar{r}, P_w, P_r, z_x, F] \tag{3.9}$$

The first six features of TD7 represents state-of-the-art time domain feature vector that is popularly used in literature for EMG based silent speech recognition, while the last one represents the novel DFA feature that is used for the first time for SEMG based silent speech recognition. The features for all channels are stacked in this manner. Seven unique combinations of these features are used to obtain

the effect of each of the feature in classification accuracy. The first combination consists of only the six time domain features and the DFA feature is not considered in this combination. In all other combinations the DFA feature is present and one among the six time domain features is excluded. The combinations used are given as:

(1) $[\bar{x}, \bar{w}, \bar{r}, P_w, P_r, z_x]$
(2) $[\bar{x}, \bar{w}, \bar{r}, P_w, P_r, F]$
(3) $[\bar{x}, \bar{w}, \bar{r}, P_w, z_x, F]$
(4) $[\bar{x}, \bar{w}, \bar{r}, P_r, z_x, F]$
(5) $[\bar{x}, \bar{w}, P_w, P_r, z_x, F]$
(6) $[\bar{x}, \bar{r}, P_w, P_r, z_x, F]$
(7) $[\bar{w}, \bar{r}, P_w, P_r, z_x, F]$

3.3.4 Classifiers

The objective of this work is to demonstrate the compatibility and performance of the new feature vector that incorporates DFA along with the state-of-the-art features used in EMG based silent speech identification. This is achieved by comparing the Word Accuracy (WAcc) obtained when seven unique combinations of the features are used. The result is validated by comparing between two different classifiers used for pattern recognition. The two classifiers that are included for this investigation are K-Nearest Neighbors (KNN) and Decision Trees (DT). The reason for choosing these two classifiers is their similarity when it comes to model complexity and computational expense. Both these classifiers offer less model complexity and computational expense as compared to other machine learning and deep learning techniques. The RF and LSTM classifiers were also tried in this research work, but they couldn't offer much in terms of model performance.

Before going into the results of the study, it is important to note the parameter specifications of the classifiers employed. The number of nearest neighbours in a KNN algorithm is represented by the 'k' value of the algorithm which was taken as '3' for this work. The distance measure employed for the algorithm was 'euclidean'. Trial and error method was performed for the selection of the appropriate distance measure and optimal k value. The split criterion used by the DT classifier was 'gini's diversity index (gdi)'. It also used a 'maximum number of splits' of '80000'. Both these were found out using trial and error. Exhaustive grid search technique was employed to determine the optimal parameters of both classifiers.

3.3.5 Statistical Significance Test

The validity of the methods devised in the research work can be established by evaluating the statistical significance of the results. There are several statistical tools to perform such an evaluation. In this research work, a paired sample t test is used as an effective evaluation tool for establishing the statistical significance of the results. The details of the test is explained in this section.

The paired sample t-test, which is also known as the dependent sample t-test, is a statistical method used to ascertain whether the mean difference between two observation sets is zero. In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations. Repeated measures designs, case control studies etc. are some popular applications of the test. Similar to many other statistical methods, the paired sample t-test also has two competing hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis assumes that the true mean difference between the paired samples is zero. Under this model, all observable differences are explained by random variation. Conversely, the alternative hypothesis assumes that the true mean difference between the paired samples is not equal to zero.

Statistical significance is evaluated by considering the 'p' value. The 'p' value denotes the probability of observing the test results under the null hypothesis. The lower the 'p' value, the lower the probability of obtaining a result like the one that was observed if the null hypothesis was true. Thus, a low 'p' value indicates decreased support for the null hypothesis. However, the possibility that the null hypothesis is true and that we simply obtained a very rare result can never be ruled out completely. The cutoff value for determining statistical significance is ultimately decided on by the researcher, but usually a value of 0.05 or less is chosen. This corresponds to a 5% (or less) chance of obtaining a result like the one that was observed if the null hypothesis was true. In this research work also a 'p' value of 0.05 is chosen for the paired sample t test.

3.4 Results and Discussion

The results presented here is based on the work done on a decisive vocabulary comprising of 1100 words and it successfully provides a conclusive idea about the influence of different features in improving the word recognition accuracy of an SEMG based model. The actual data consists of 7 channels of EMG from which 6 time-domain features and 1 time-frequency domain feature per channel are

calculated. Thus the total number of features extracted for classification accounts to 49. The speech recognition model devised in this work is a speaker independent model and hence no categorical predictors to identify the speaker is provided to the model. The word recognition accuracy reported here is the mean value of a total of 50 trials. All 1100 words used in this study are included in each of these 50 trials. 80% of the total available data was utilised for training and the testing was done on the remaining 20% data. All available data is used in each trial but the data for training and testing is randomized for each trial. Both training and testing of the model was implemented in frame level. To recognize the uttered word, voting was performed on the classified frames once testing process was finished. After the voting process, the word recognition accuracy was calculated for each word in the testing samples.

The impact of introducing the DFA feature along with state-of-the-art time domain features is presented in the following sections. A comparison in terms of both word recognition accuracy as well as computation time is provided. The trials for all these combinations are performed for both of the classifiers in order to confirm the importance of the DFA feature in improving the word recognition accuracy. Ten fold cross validation was implemented on a separate trial for both classifiers to ensure the reliability of the accuracy obtained. The evaluation time is also obtained for all cases in order to denote the computational expense associated with each classifier and combination. It can thus give an insight into the practicality of introducing the DFA feature in the area of silent speech recognition.

3.4.1 Choice of Optimal Window Length and Window Shift

The feature extraction process usually consists of the use a suitable windowing method. The choice of the windowing method, window length, and window shift is an important step in the success of the process. In this research work, rectangular windowing scheme is used in the extraction of all the features. A window length of 54 ms and an overlapping window shift of 1.6 ms is found to be optimal for the work. The selection of optimal window length and window shift was done using a grid search method. The range of values for the grid search was obtained from literature where EMG signals were used for silent speech recognition (Wand et al., 2014) (Janke & Diener, 2017) (Wand et al., 2013). The classification accuracy improves with an increase in window length and it saturates after a particular point. The choice of the optimal window length and window shift can be comprehended from Figure 3.3 and 54 ms is chosen as the window length for faster

computation.

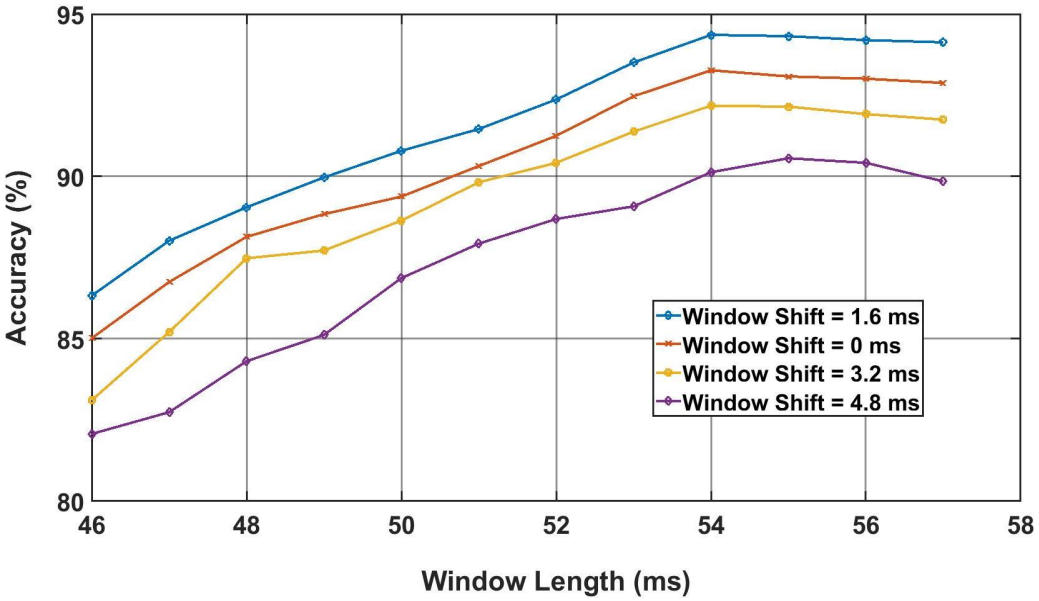


Figure 3.3: Choice of optimal window length and window shift

3.4.2 Phoneme based SSI

Phonemes can be termed as the fundamental building blocks of any word in a vocabulary. The English language has a total of 44 phonemes using which the pronunciation of any word in the English vocabulary can be obtained. The phonemes in English are constituted by either a vowel, or a consonant, or by the combination of both. Phoneme based silent speech recognition was attempted as part of this research work using computationally less expensive classifiers. But the results obtained were not satisfactory. Throughout the literature, phoneme based models are developed either using computationally expensive machine learning methods or by using deep learning techniques. Due to this challenge, the focus of realizing a computationally less expensive SSI model was hence shifted to word based silent speech recognition.

3.4.3 Classification using KNN

The word recognition accuracy for all the combinations across 50 trials is plotted in Figure 3.4. The total results obtained from the KNN based model is presented in Table 3.1. The combination 1 that did not use the DFA feature for classification reported the lowest accuracy while combination 7 that used the DFA feature gave

the highest accuracy. All the combinations that used DFA can be seen to perform significantly better than the combination 1 that did not use DFA. The time domain feature that is excluded in combination 7 is normalized SEMG. There is an increase of more than 2% in accuracy when the DFA feature was introduced. A paired sample t-test was done on both of these combinations and it yielded a 'p' value much less than 0.05 and hence it can be ascertained that the improvement in accuracy for combination 7 as compared to combination 1 is statistically significant. The training and testing times are computed per sample of the feature vector and is inclusive of both feature extraction and classification. It is also tabulated along with the classification accuracy and standard deviation. The computation time taken by the model is very less thereby making it feasible for hardware implementation with reduced cost.

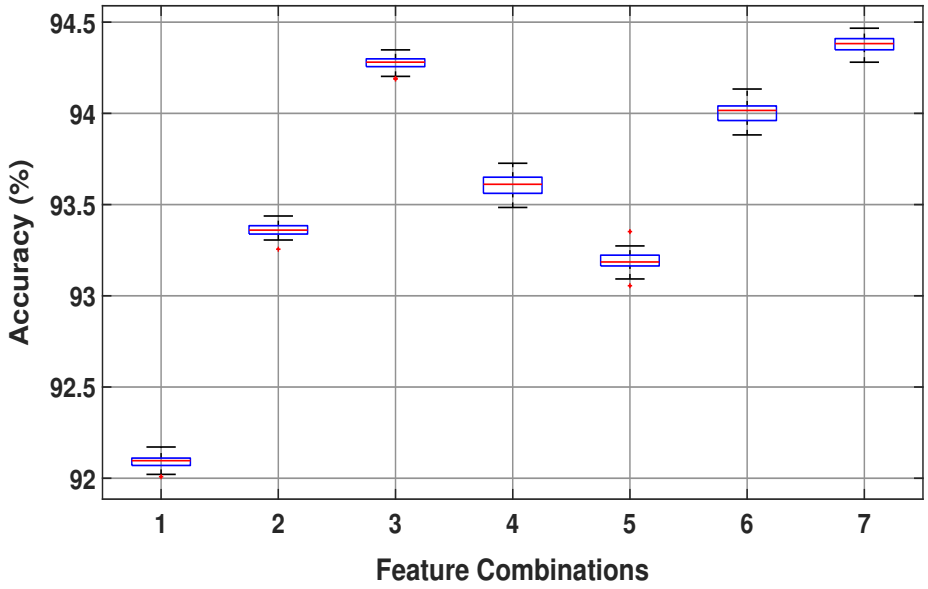


Figure 3.4: Word recognition accuracy using KNN

The training time required for KNN is much less than the testing time since most of the computation is done during testing. The testing phase is when the algorithm computes distances to identify the nearest neighbours of the test samples. There are slight variations in classification accuracy for different words and this can be attributed to errors associated with data acquisition and extraction. In a word based SSI system, the data is usually obtained by separate utterances of each word, as in the case of the 10 Chinese words used by (Zhang et al., 2020). The words used in this work are taken from the utterances of sentences as a whole, using time markers of the signal. This can pose a higher possibility of errors associated with data extraction.

Table 3.1: Word recognition accuracy and computation time for KNN based model

Combination Number	Accuracy (%)	Standard Deviation (%)	Training Time (μs)	Testing Time (ms)
1	92.0953	0.0373	8.36	11.5
2	93.3605	0.0365	8.89	11.8
3	94.2752	0.0368	9.20	11.6
4	93.6089	0.0564	10.13	11.7
5	93.1898	0.0513	11.08	11.3
6	94.0099	0.0569	10.42	11.6
7	94.3770	0.0432	8.71	11.4

3.4.4 Classification using DT

The classification using DT also yielded similar results as in the previous case. The word accuracy plot and the total results are given in Figure 3.5 and Table 3.2 respectively. There is an increase of 3% in accuracy for the combination that uses the DFA feature as compared to the combination where DFA is not used. The paired sample t-test for this model also yielded a similar result as in the case of KNN based model. The 'p' value obtained was much less that 0.05 thereby demonstrating that the improvement in accuracy is statistically significant.

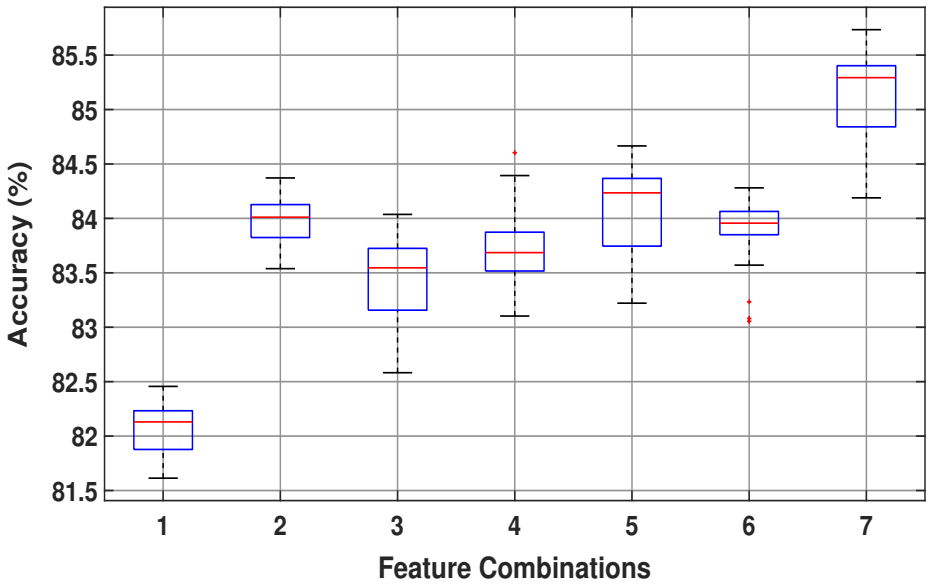


Figure 3.5: Word recognition accuracy using DT

Table 3.2: Word recognition accuracy and computation time for DT based model

Combination Number	**Accuracy (%)**	**Standard Deviation (%)**	**Training Time (ms)**	**Testing Time (μs)**
1	82.0778	0.2250	1.09	28.71
2	83.9940	0.1985	1.14	29.13
3	83.4722	0.3302	1.13	28.56
4	83.7016	0.3163	1.08	28.47
5	84.0895	0.3814	1.04	29.88
6	83.9049	0.2517	1.23	30.61
7	85.1506	0.3801	1.14	29.39

3.4.5 Cross Validation of the Results

The data used for training and testing in each of the 50 independent trials is ensured to be randomized using random indices. However as a double check to ensure the reliability of the classification accuracy, a cross validation trial is performed separately in addition to the 50 trials. Only the two combinations under direct comparison is included in this trial. The results of cross validation performed for combination 1 and 7 for both classifiers are presented in Table 3.3.

Table 3.3: Word recognition accuracy using cross validation

Classifier	**Combination Number**	**Accuracy (%)**	**Testing Time**
KNN	1	92.07	11.6 ms
	7	94.36	11.4 ms
DT	1	82.05	28.02 μs
	7	85.20	28.74 μs

3.4.6 Comparison with the Accuracy Benchmark

In 2017, (Meltzner et al., 2017) performed silent speech recognition on the data of eight people in which eight sensors were employed. The eight people who were part of this work had earlier underwent total laryngectomy. The data was recorded after a minimum period of six months from the surgical procedure. The model used in the work was phoneme based and 39 English phonemes (out of a total of 44 phonemes) that are most commonly used, were considered for the research. They got a Word Accuracy (WAcc) of 89.7% (Word Error Rate, WER of 10.3%)

for the SSI model. They also obtained a WAcc of 86.4% while using a model that employed only four sensors. The work was carried out on a vocabulary of 2500 words. In 2018, the accuracy was further improved to 91.1% (Meltzner et al., 2018) in a 2200 word vocabulary.

In 2020, (Zhang et al., 2020) employed Inductive Conformal Prediction (ICP) technique for SEMG based SSI, to obtain a maximum classification accuracy of 88.5% on a dataset of 10 Chinese words. The classifier employed was random forests and the major objective was to utilize ICP for ensuring confidence and reliability in prediction. They used a technique called test time data augmentation, where ICP was employed for utilizing unlabelled data so as to improve model performance. The work used word based approach for silent speech recognition. It is also important here to note that the approach adopted here uses the words that are independently uttered and not as part of the utterance of a whole sentence. It is relatively easier to identify words when they are uttered separately as compared with the identification of words which are part of the sentence utterance.

The effect of DFA based feature in improving the word classification accuracy is evident from the results presented in this chapter. The results from both the classification methods establish the importance of this time-frequency domain feature in the area of silent speech recognition using surface electromyography which is dominated by time domain features. It is also important to note that the KNN based model is able to achieve an accuracy that is already benchmarked in the literature. The KNN model devised in this research work achieved a word accuracy of 94.37% that is comparable with the accuracy benchmark of 89.7% (which was further improved to 91.1% on a slightly reduced dataset) obtained by (Meltzner et al., 2018) where word-phoneme hybrid method is used and deep learning techniques are employed. Same is the case when compared with the work mentioned in (Zhang et al., 2020). They obtained an accuracy of 93.6% on a very limited vocabulary comprising of just 10 words that were uttered independently. The work reported in this thesis is performed on word utterances taken from the complete utterances of sentences and is able to achieve better accuracy by using only SEMG data and computationally less expensive pattern recognition techniques. A comparison between the model discussed in this research work and the model developed in the benchmark papers is presented in Table 3.4. The accuracy benchmark studies are included just to demonstrate that the overall accuracy achieved by this model does not fall short of the existing methods.

Table 3.4: Comparison with accuracy benchmark

	This Work	**Meltzner et.al**	**Zhang et.al**
Number of Words	1100	2500	10
Type of the Data	SEMG	SEMG	SEMG
SSI Method Adopted	Word based	Word-Phoneme	Word based
Classification Methods	KNN , DT	HMM-GMM	Random Forests
Word Recognition Accuracy (Standard Deviation)	94.4 % (0.04 %), 85.1 % (0.38 %)	89.7 % (5.3 %)	93.6 % (-)
Model Complexity	Low	High	Moderate

3.4.7 Discussion on the Superiority of DFA

The reason for the success of DFA based feature in the classification of SEMG signals could be the ability of DFA to exploit the non stationary nature of EMG signals thereby helping the classifier for an effective pattern recognition. DFA has the ability to exploit the characteristics of both time domain and time-frequency domain and hence it possess a unique advantage as compared to other features in the area of SSI. DFA offers better class separability than other conventional features during low level muscle activation (Phinyomark et al., 2012). In silent speech recognition, low level muscle activation is an inherent phenomenon and its magnitude varies from person to person and between different accents as well. The ability of DFA to categorize EEG signals into focal and non-focal groups is described by (Chatterjee et al., 2017). This result provided an insight into the possible effectiveness of using DFA where more cross talk between channels are anticipated. Facial muscle activity is always prone to the occurrence of cross talk from adjacent muscles. The possibility of cross talk is particularly challenging in the case of technologies like the use of array of electrodes that is recently reported in several research papers. So DFA has the capacity to offer better performance in such cases where channel optimisation or reduction is required. It is also important to note that DFA could perform well alongside the state-of-the-art features (that are time domain) which is a good indication of its possible use in similar research domains.

3.5 Summary

This chapter presents the main novelty of this research work, where a highly promising feature extraction technique (DFA) is introduced for the pattern recognition of SEMG signals to identify silently uttered speech. DFA has been used previousle for many biomedical signals including SEMG, but this is the first time it is being used for silent speech recognition and it has proved its superiority in that domain. The chapter provides a strong foundation regarding the potential benefits of DFA that acts as a guiding step towards the following chapters of this thesis, where other modalities in the area of silent speech recognition is investigated.

Chapter 4

Realization of Channel Reduction and Model Simplicity for Facial Electromyography based Silent Speech Interface Model

Contents

4.1 Introduction

The EMG data used in this work consists of 7 channels and the locations of the electrodes on the human face is given in Figure 4.1. It can be seen that the electrodes occupy a considerable portion of the face which causes difficulty to the user. Thus minimising the number of channels leads to reduced hardware complexity. A reduction in the number of channels can also provide the advantage of decreasing the computational expense of the model. The implementation of channel reduction on larger vocabularies of data requires the use of better performing features as compared to the conventional ones used by researchers.

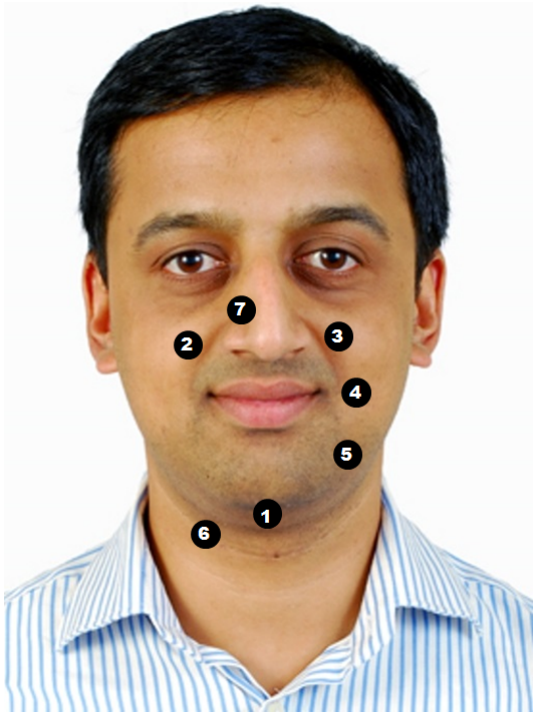


Figure 4.1: Actual electrode locations

The aim is to reduce the number of channels by at least two. Hence there would be a total of 5 electrodes (including the reference electrode) remaining. To achieve this, it is important to find out the impact of each of these channels on word recognition accuracy. So different channel combinations have to be tested for the optimal selection of the channels. It is also important to analyze the impact of the DFA based feature in the whole channel reduction process. Hence the comparisons between the combinations that use the DFA feature and the ones that doesn't use it needs to be done.

4.2 Channel Reduction vs Channel Optimisation

In literature some techniques have been discussed for the optimisation of channels. The methods such as the use of only bipolar electrodes and the use of an array of electrodes have been widely used. There are also studies that applied channel reduction by replacing certain channels with some other technology such as electroglottograph or facial plethysmogram. But these methods do possess many drawbacks as compared to a proper channel reduction methodology that just reduces the number of electrodes used. Figure 4.2 shows an electrode arrangement method were an array positioning strategy is adopted. Some advantages of this method includes shorter setup time than the separate electrode method, better ease of use for the subject, and reduced dimensionality of features. The major drawback of this method is that since there is no much reduction in channels, the hardware complexity and computational expense are not reduced. Also the electrode positioning is somewhat compromised in this method thereby causing challenges of cross talk between adjacent muscles and subsequent loss in accuracy.

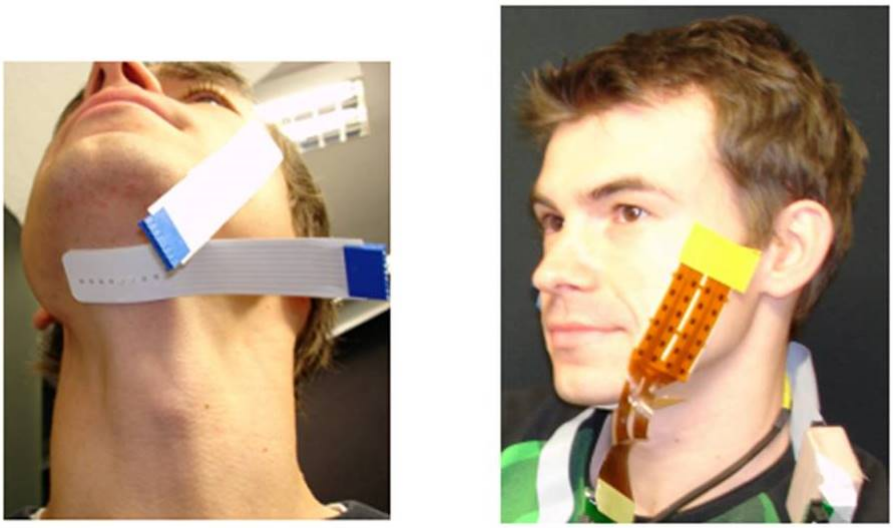


Figure 4.2: Positioning of EMG electrodes as an array (taken from (Wand et al., 2013))

The drawback of using only bipolar electrodes is that some muscle positions that are physically apart on the face actually compliment one another when it comes to pattern recognition. Only unipolar electrodes can be applicable in such situations. So the use of only bipolar electrodes may not be advantageous in all cases. But efforts can be made to employ more bipolar electrodes instead of unipolar ones such that the objective of reducing the number of electrodes on the

subject's face can be achieved. The use of other technologies such as electroglottograph or facial plethysmogram instead of some channels is also not much helpful when hardware complexity and computational expense are considered. Further, the accuracy gains from such hybrid models are not much lucrative as compared to the additional efforts in designing speech recognition models that use hybrid data.

4.3 Materials and Methods

The materials and methods associated with the objective is presented in this section. First, the information regarding facial musculature is provided such that an effective comprehension of the target muscles and their locations is possible. This in turn can help to draft a suitable electrode positioning strategy without considerable effect of cross talk between muscles. The data used, and the various methods such as the estimation of predictor importance and channel importance, that were attempted to achieve effective channel reduction are also explained.

4.3.1 Facial Musculature

A brief idea about the human facial muscles is presented in this section that can help in identifying the target muscles associated with silent speech recognition using SEMG. This information is particularly important when the investigations regarding channel reduction is done. Another important aspect to note is that the facial muscles that are under consideration here are generally not affected by the laryngectomy procedure, thus enabling researchers in successful acquisition of facial EMG data for silent speech recognition research.

The human facial musculature is capable of expressing diverse types of information such as communicative gestures, emotions, and reflex actions. So the facial muscles have a great potential for various modes of information exchange. For employing SEMG, it is very important to identify and choose those facial muscles that are related to human speech. The effectiveness of the data acquired depends on the proper choice of these muscles and the corresponding electrode locations. There is also an effect of cross-talk happening due to the overlap of various muscles, which has to be taken care of while choosing the location of electrodes. The topography of the facial muscles associated with speech is shown in Figure 4.3. The location of the electrodes has to be chosen, considering the possibility of minimizing the number of electrodes. Less number of electrodes can reduce

the hardware complexity of data acquisition as well as the computational expense when it comes to feature extraction and classification. It will be comfortable and not much invasive to the user as well.

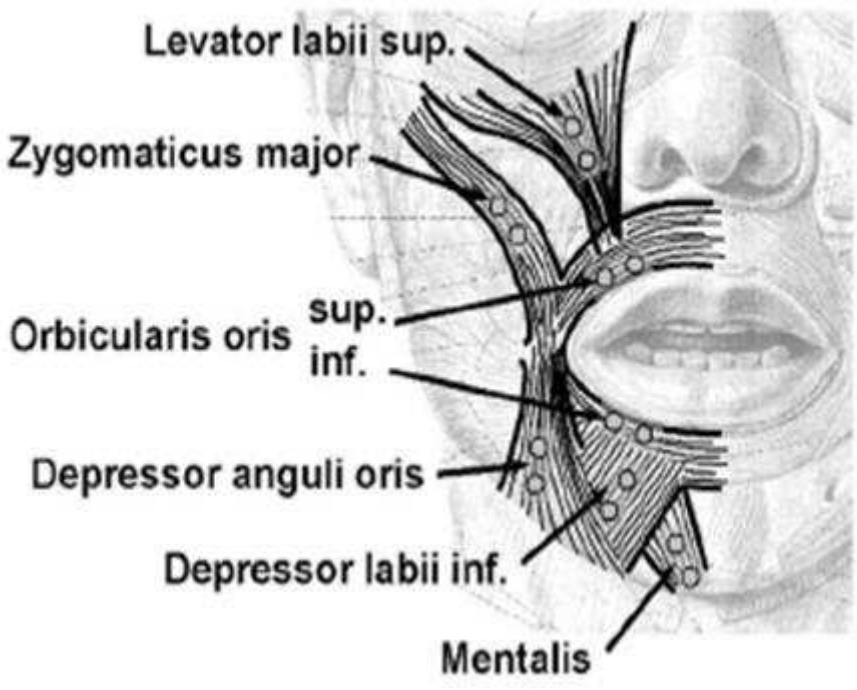


Figure 4.3: Facial musculature (Source: (Yau et al., 2008))

4.3.2 Data Used

The objective of channel reduction is also performed on the same dataset (that was used for DFA based models in the previous chapter) developed by the researchers of the Interactive Systems Labs, University of Karlsruhe (Wand et al., 2014). The vocabulary used for channel reduction analysis consisted of the same 1100 words that was used in the DFA based models earlier. The initial research was focused on using only the existing state-of-the-art features for achieving channel reduction. During that time, the suitability of DFA based feature in SEMG based SSI was not identified. Later when the effectiveness of DFA in this research area was ascertained, it was further used to achieve effective channel reduction as well. The availability of error free and reliable data helped in carrying out investigations of channel reduction in SEMG based SSI.

4.3.3 Predictor Importance and Channel Importance

There is a tool in MATLAB which estimates the importance of each of the predictor columns. The predictor columns correspond to each of the feature under each of the channel. A total of seven features and seven channels are employed in this work, which constitutes a total of 49 predictors. The predictor importance tool in MATLAB was used to estimate the impact of each of the predictors in the successful classification of the words. The predictor importance measures were then used to compute the channel importance of the model. However the result of this channel importance calculation did not favour the research. The algebraic calculation of channel importance from the predictor importance could not yield favourable outcomes and hence it became necessary to explore other methods. This prompted the investigation into different channel combinations.

4.3.4 Channel Combinations

A basic trial and error approach was performed on the dataset comprising of 1100 words. This was done by performing classification while removing one channel at a time such that the impact of each of the channels in classification accuracy was visualized. Various combinations of channels were then tried out based on this result to finally arrive at the best possible channel combination. The same process was performed over multiple trials and the average values of the results obtained were not different.

4.4 Results and Discussion

The results obtained from the investigations related to channel reduction and the discussion associated with it is presented in this section. The classifiers, KNN and DT are used once again for the investigations and the comparisons of the channel reduced models with the full channel models are presented. The vocabulary, feature extraction methodology, and classifier parameters used here is same as that of the previous chapter. A total of 50 trials is performed for each of the investigations in channel reduction as well. The results presented here is the mean value of all the trials.

4.4.1 Investigation of the Channel Combinations

The search for the best performing channel combination started with the investigation of the impact of each of the channel in the word recognition accuracy. To achieve this, trials were performed for different combinations where one channel was excluded at a time. The resultant accuracy was an indication to the impact of the excluded channel on the word recognition accuracy. Figure 4.4 shows the accuracy of the combinations under consideration.

It can be seen that the channels 1,2,5 and 6 had a greater impact on word recognition accuracy as compared to other channels. So the channel combination 1,2,5,6 was also tested and the resultant accuracy can be seen to perform well enough as compared to the all channel combination. The performance of a silent speech recognition model is based on factors such as choice of muscles, optimal electrode number and positioning, muscular crosstalk etc. among several other factors. The performance of a particular feature for a particular channel combination is dependent on these factors as well. The reason behind channel combination 1,2,5,6 performing better than other combinations is the ability of DFA to capture distinguishable patterns for those particular muscle locations. The information regarding facial musculature for electrode positioning is given in the section 4.3.1.

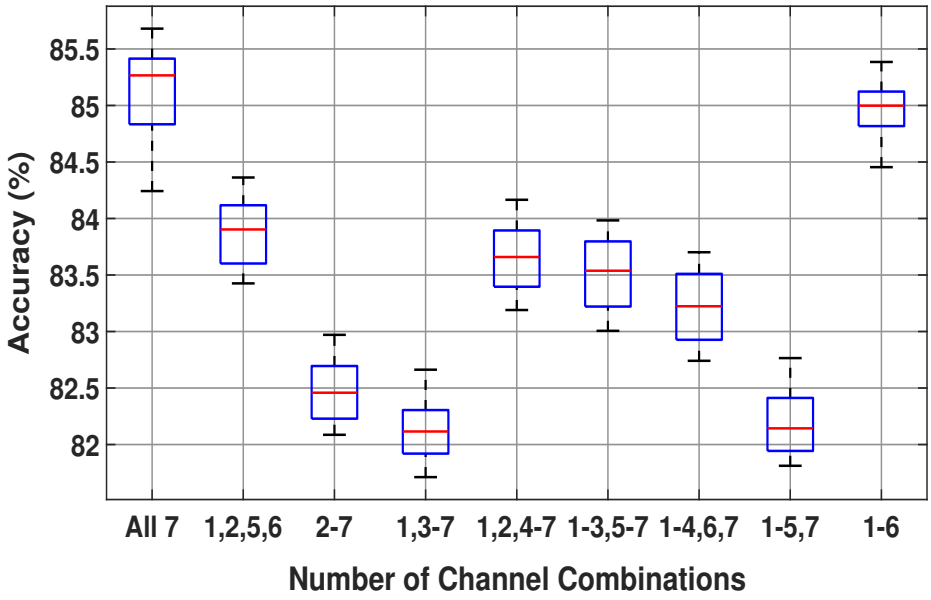


Figure 4.4: Accuracy boxplot of various channel combinations

The investigation of different channel combinations was first performed on DT based model and the results were validated using KNN model. Based on the results obtained from different channel combinations, the reduced channel locations can

be visualized in Figure 4.5.

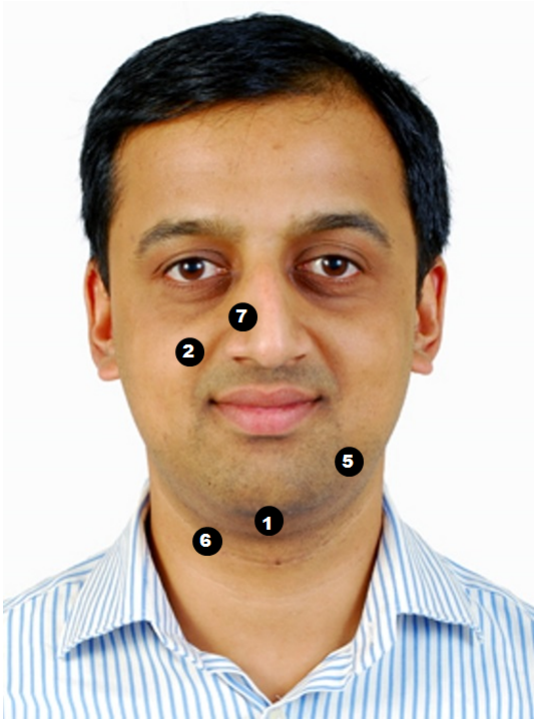


Figure 4.5: Proposed electrode locations after channel reduction

4.4.2 Impact of DFA in Channel Reduction of KNN based Model

The application of DFA as an additional feature along with the existing time domain features had a crucial impact in improving the model performance which is visible in the accuracy plot given in Figure 4.6. The mean values for all the 50 trials, the standard deviation, and computation times are presented in Table 4.1. The results presented here pertains to the KNN based silent speech recognition model. Channel reduction when implemented on the state-of-the-art model suffers poor performance by a considerable margin, as compared to combinations that employ DFA.

Confusion matrices plotted for both of the KNN based channel reduced models - one that uses TDFV-DFA and the one that uses only TDFV - is given in Figure 4.7. The reduction in overall accuracy for the TDFV only model is clearly visible in the vertical axis. However the confusion matrix plot appears crowded due to the presence of high number of output classes, i.e 1100 classes for the vocabulary under consideration.

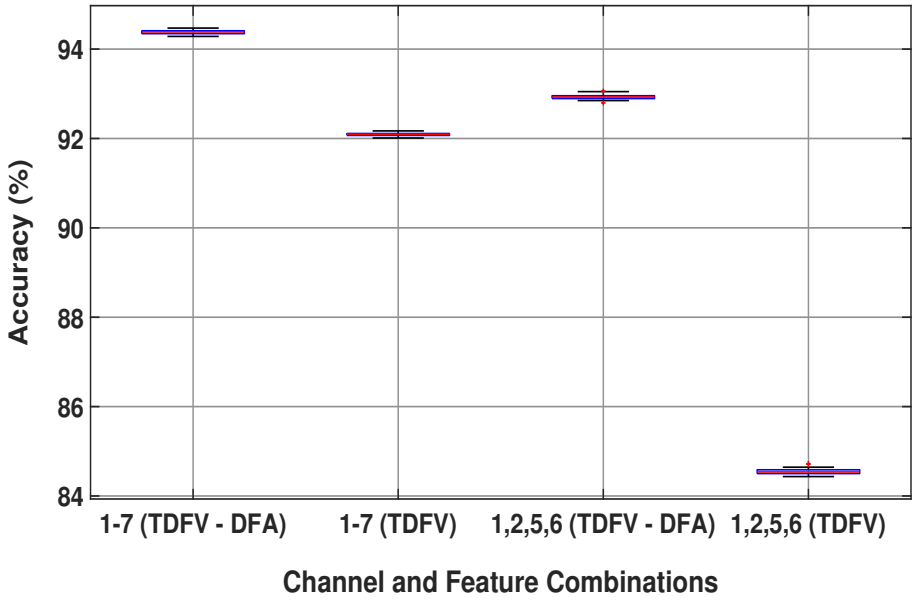
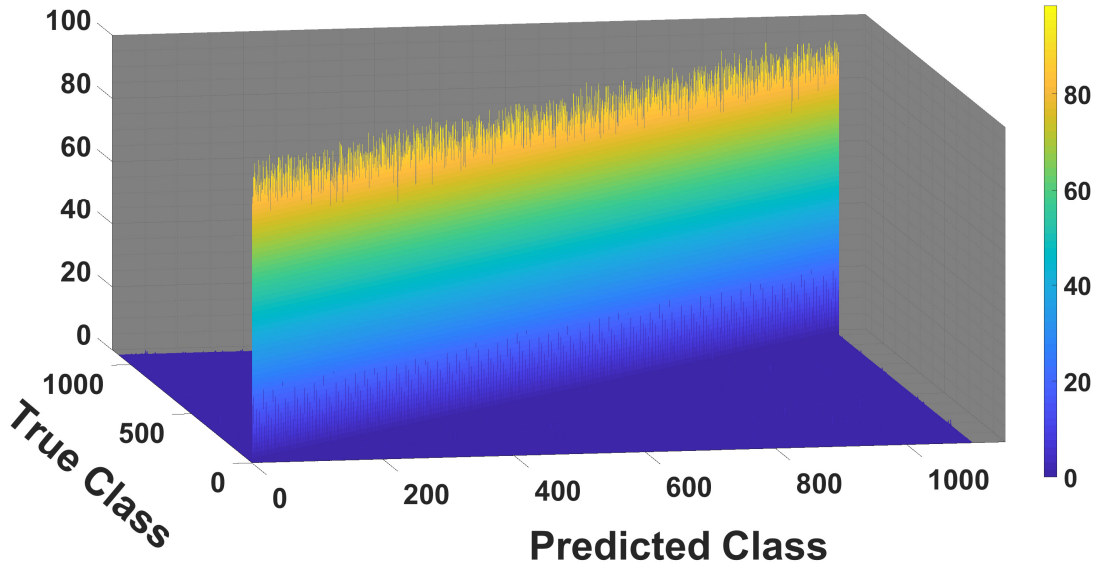


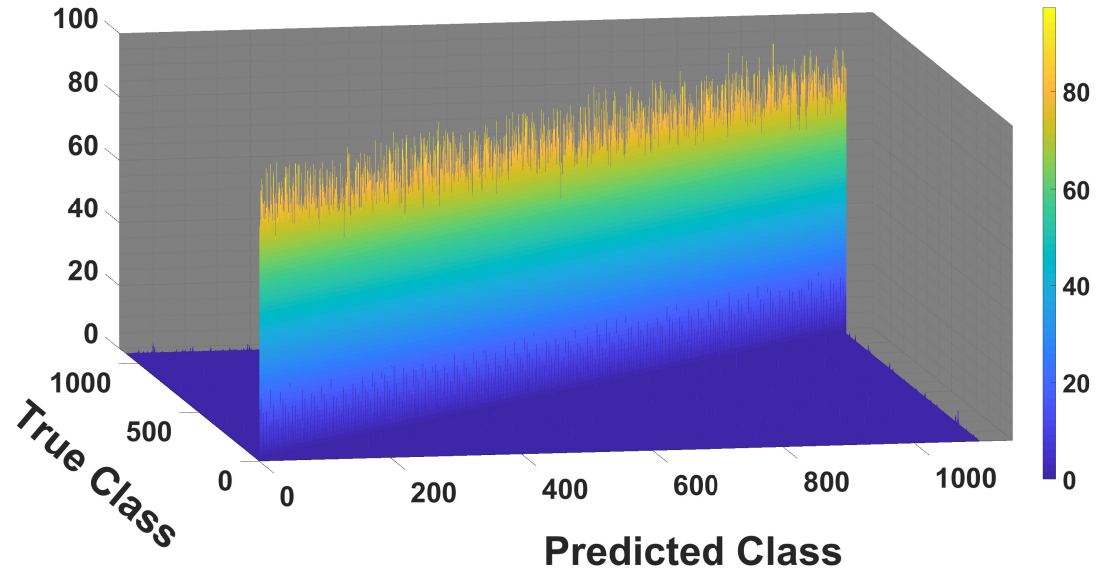
Figure 4.6: Accuracy comparison using KNN (TDFV vs TDFV-DFA)

Table 4.1: Comparison of accuracy for KNN based model

Channels Taken	**Features Used**	**Accuracy (%)**	**Standard Deviation (%)**	**Training Time (μs)**	**Testing Time (ms)**
All Channels	TDFV, DFA	94.3770	0.0432	8.71	11.4
	TDFV Only	92.0953	0.0373	8.36	11.5
Channels 1,2,5,6 Only	TDFV, DFA	92.9263	0.0511	8.42	10.7
	TDFV Only	84.5468	0.0630	8.27	10.3



(a) Channels 1,2,5,6 (TDFV-DFA)



(b) Channels 1,2,5,6 (TDFV Only)

Figure 4.7: Confusion matrices for KNN based model

4.4.3 Impact of DFA in Channel Reduction of DT based Model

It is important to try out the same method in a model that uses a different classifier, so that the results obtained is more reliable. As mentioned in the previous sections of this thesis, DT is the second classifier that proved useful in this research work. The accuracy plot obtained by using a DT based model is presented in Figure 4.8. The mean values of accuracy, standard deviation, and computation times are tabulated in Table 4.2. The effect of DFA in improving accuracy is better quantified in the results of the DT based model because as far as the KNN model is considered, the accuracy values are nearly saturated while there is further scope for improvement in the case of DT based model.

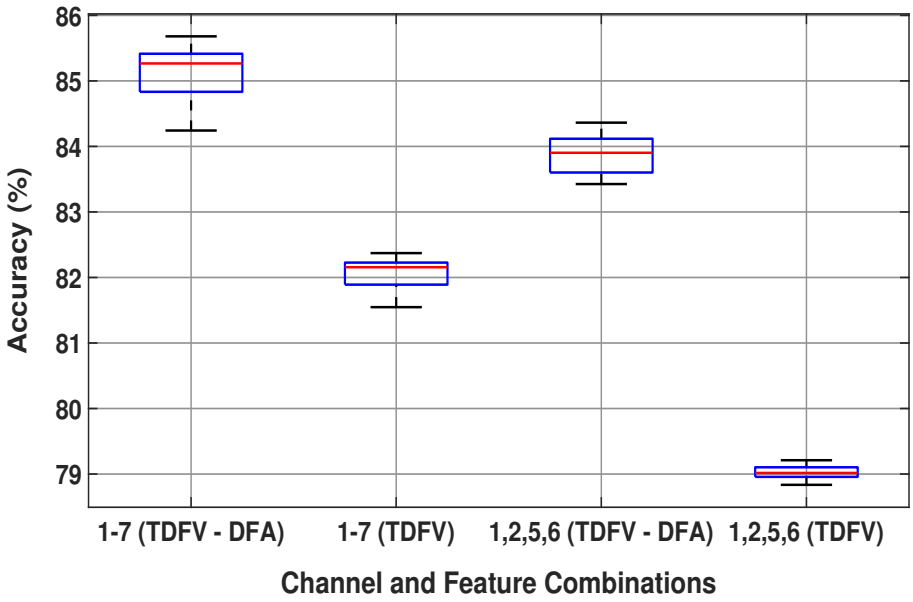
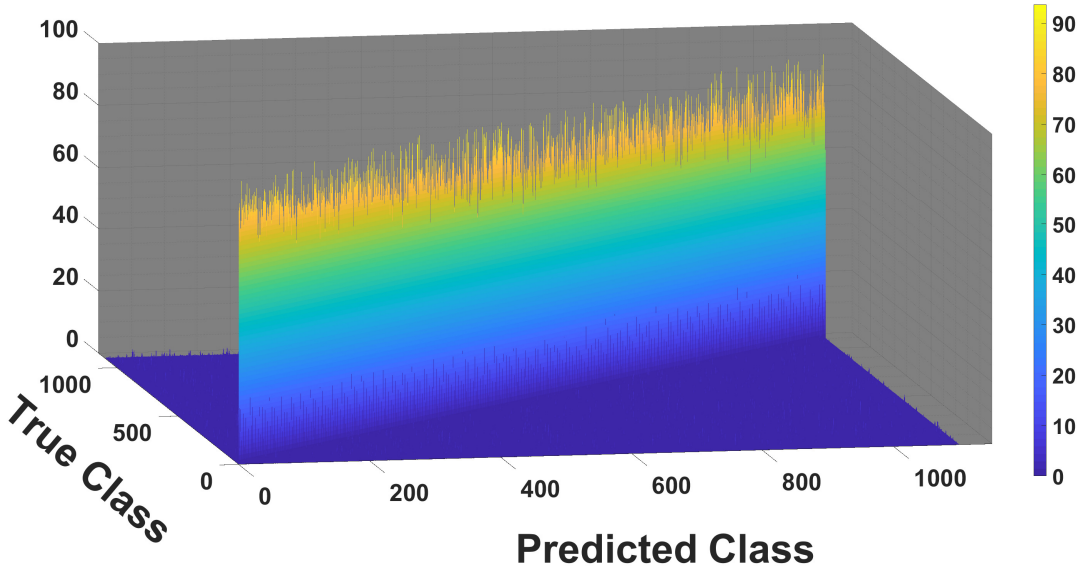


Figure 4.8: Accuracy comparison using DT (TDFV vs TDFV-DFA)

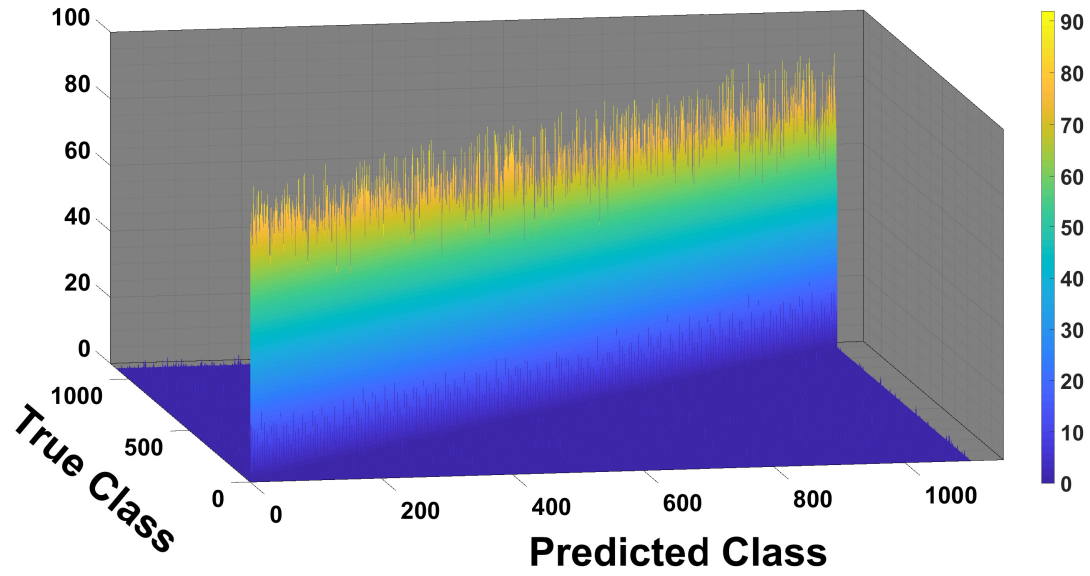
The confusion matrices for the DT based models are plotted in Figure 4.9. Here also the high number of output classes is a hindrance in better representation and visualization of the plot. However the difference between both confusion matrix plots can be seen when observed carefully.

4.4.4 Discussion on Channel Reduction using DFA

Two important findings can be comprehended from the results. One is the impact of the DFA based feature. For both the full channel and the channel reduced combination, the presence of DFA has contributed significantly in maintaining



(a) Channels 1,2,5,6 (TDFV-DFA)



(b) Channels 1,2,5,6 (TDFV Only)

Figure 4.9: Confusion matrices for DT based model

Table 4.2: Comparison of accuracy for DT based model

Channels Taken	**Features Used**	**Accuracy (%)**	**Standard Deviation (%)**	**Training Time (ms)**	**Testing Time (μs)**
All Channels	TDFV, DFA	85.1506	0.3801	1.14	29.39
	TDFV Only	82.0778	0.2250	1.09	28.71
Channels 1,2,5,6 Only	TDFV, DFA	83.8841	0.2865	1.08	27.23
	TDFV Only	79.0240	0.0847	1.05	26.59

the accuracy. The second important observation is the closely matching profile of the TDFV-DFA channel reduced combination with that of the TDFV-DFA all channel plot. This demonstrates the ability of DFA in maintaining accuracy even when a significant reduction in data occurs and hence the suitability of DFA feature in channel reduction is established. In some situations there will be certain words whose recognition becomes difficult due to several potential reasons and are often treated as outliers. In this work, the vocabulary used in establishing the superiority of DFA in the previous chapter, is used as such for channel reduction studies as well. Neither new words were added, nor the existing ones were omitted. The confusion matrices provides the consistency of the results across the entire vocabulary.

4.5 Summary

This chapter presented various analyses regarding the implementation of channel reduction in silent speech recognition using SEMG. The difference between channel reduction and channel optimisation is described and the information regarding various approaches to achieve successful channel reduction is discussed in the beginning of the chapter. The results of the investigation of different channel combinations and the identification of the most fruitful combination is presented. The results of four silent speech recognition models (TDFV Only full channel model, TDFV-DFA full channel model, TDFV Only reduced channel model, and TDFV-DFA reduced channel model) were presented for each of the two classifiers. An elaborate discussion on the impact of DFA in achieving successful channel reduction is also presented in this chapter.

Chapter 5

Data Acquisition Setup and Sample Data Collection

Contents

5.1 Introduction

The future research work in the area of SSI is dependent on the availability of reliable data. At present there is a scarcity of reliable data sets in the research domain of SEMG based silent speech recognition. By having diverse data sets from different parts of the world, the challenges associated with various accents can be effectively addressed. These factors led to the thought of laying the foundation stone for creating a SEMG based SSI dataset. The setting up of hardware and data acquisition methodology is envisioned as an initial step for developing such an extensive database. Hardware setup constituted of equipment purchase, drafting of data acquisition methodology, and finally the overall setup. Data acquisition is not just limited to EMG. Video and audio recording setup was also included for possible use in the future. The data acquisition methodology describes in detail about the steps to be followed for successful data acquisition. A sample data is also collected using the data acquisition setup. However, silent speech recognition is not performed on the sample dataset since the acquired data is limited.

5.2 Hardware Components for SEMG

The hardware components required for the setup of a SEMG data acquisition system is described in detail in this section. Before that it is important to have a basic idea about the processes involved in SEMG data acquisition. Figure 5.1 depicts the basic block diagram representation of an SEMG data acquisition system. The SEMG electrodes are affixed on the subject's face using sticky gel mechanism. The electrodes are then connected to the sensors with the help of snap connections. The sensor outputs are connected to the sensor isolator, the output of which is fed to the Data Acquisition (DAQ) module via a terminal block (used for easier interconnection). The DAQ module basically consists of an acquisition device housed on a chassis. The output of DAQ module is connected to the data acquiring computer via a USB connection.

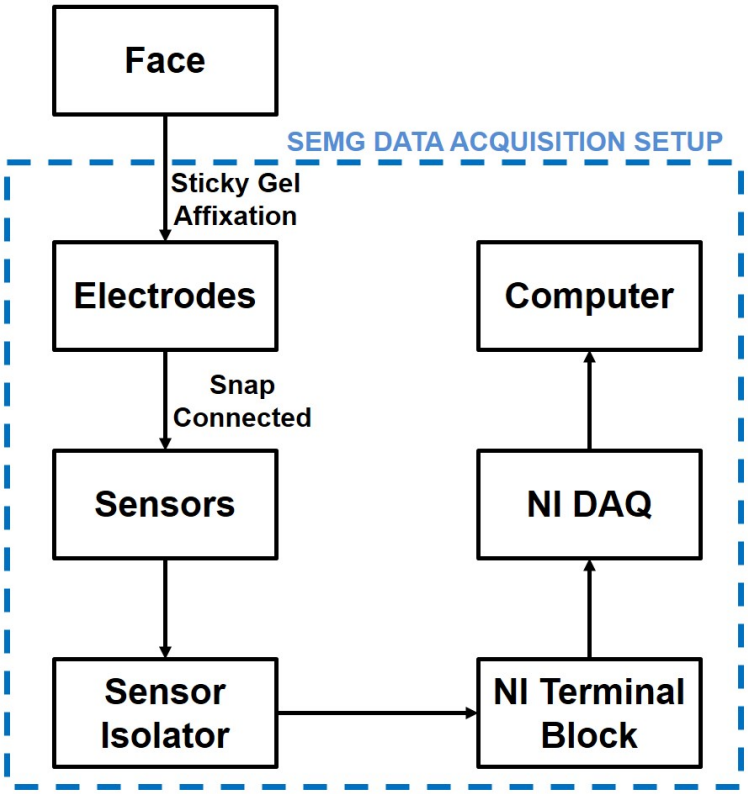


Figure 5.1: Block diagram of SEMG acquisition setup

5.2.1 SEMG Electrodes

The electrodes used for data acquisition is "$3M^{TM}RedDot^{TM}$ Multi Purpose Monitoring Electrodes with Sticky Gel 2560 Electrode" which is shown in Figure 5.2. It has a circular sensor area of 3.48 sq cm which is sufficient for facial EMG extraction. The sensor material is Ag/AgCl which is the standard material used for manufacturing EMG electrodes. It has a stainless steel snap connector that can be attached to the sensors easily. The detailed specifications of the electrode is given in Appendix A.

5.2.2 Sensors

"MyoScan EMG Sensor (SA9503Z)" developed by Thought Technology Ltd. is used for connecting the electrodes attached to face to the sensor isolator. The sensors and electrodes are connected via snap connections. It has a signal input range of 0 – 2000 µV RMS. A bipolar electrode method is used the data acquisition setup which is demonstrated in Figure 5.3. The detailed specifications of the sensors is given in Appendix A.

Figure 5.2: SEMG electrode used (Ag/AgCl)

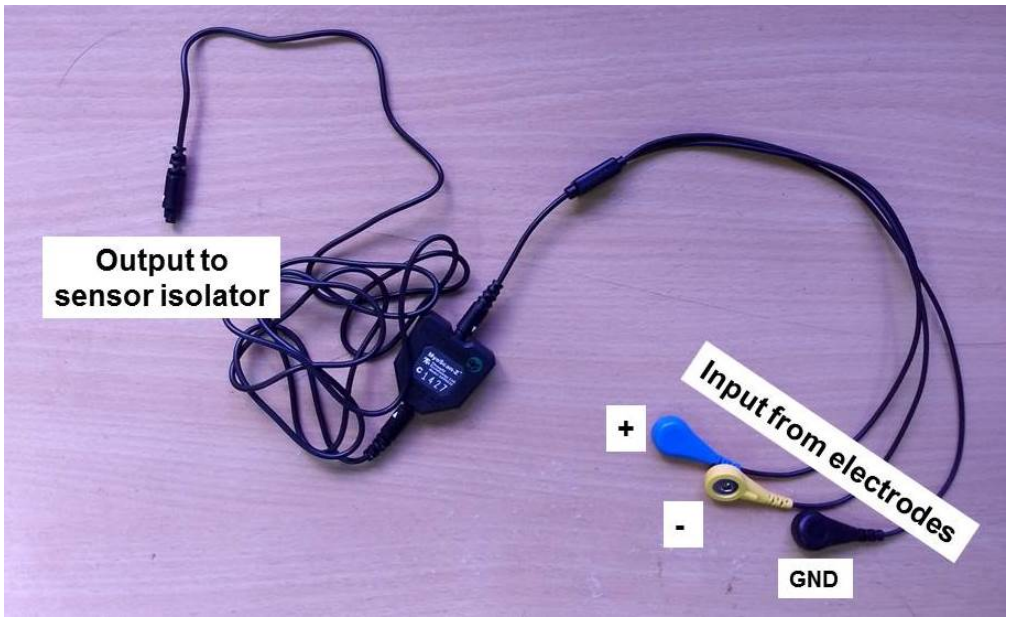


Figure 5.3: Sensors

5.2.3 Sensor Isolator

The interface device used to accommodate sufficient degree of electrical isolation is called a sensor isolator. It facilitates the safe interfacing of EMG sensors with the analog inputs of line powered systems like computers with DAC cards. The sensor isolator used here is manufactured by Thought Technology Ltd. and is shown in Figure 5.4. The sensor isolator used here has the capacity to accommodate four channels of SEMG data. A total of two sensor isolators are used in the sample data collection to accomodate for five channels. The detailed specifications of the isolator is given in Appendix A.

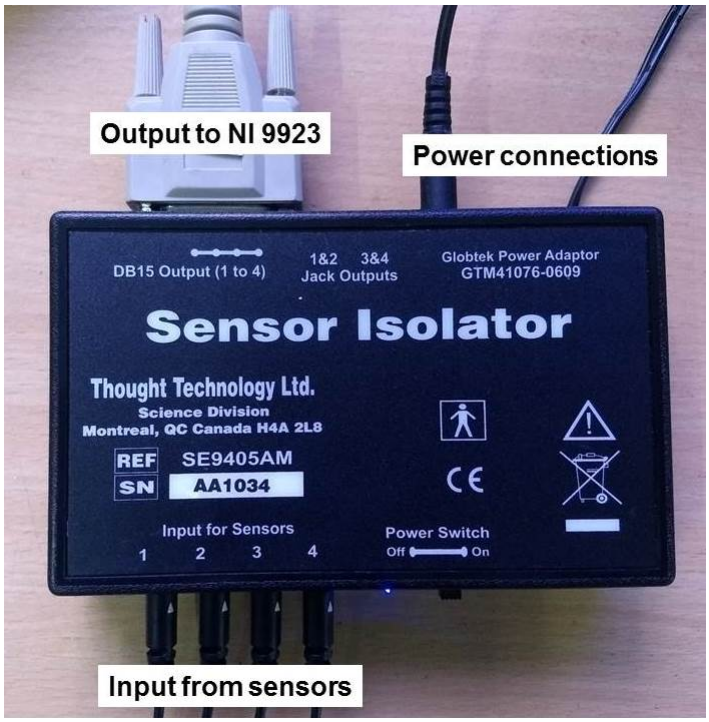


Figure 5.4: Sensor Isolator

5.2.4 NI DAQ

Data acquisition module manufactured by National Instruments is used for acquiring SEMG data. The entire module can be sub divided into 3 parts, NI 9923 - Terminal Block, NI 9205 - Voltage Input Module, and NI cDAQ 9178 - CompactDAQ Chassis. NI 9923 Terminal Block is used as an interface between the incoming connector wires from the sensor isolator and the NI 9205 by facilitating screw-terminal connectivity. It is shown in Figure 5.5.

NI 9205 - Voltage Input Module facilitates the acquisition of single ended or

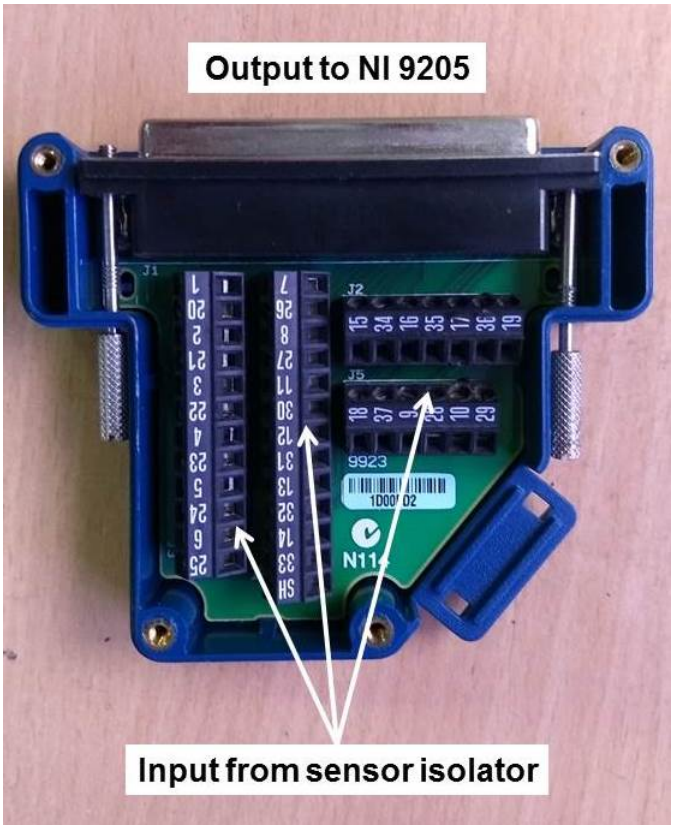


Figure 5.5: NI 9923 - terminal block

differential analog inputs, each having 4 programmable input ranges. The output of NI 9205 is given to the NI cDAQ 9178 - CompactDAQ Chassis. NI 9205 - Voltage Input Module is shown in Figure 5.6.

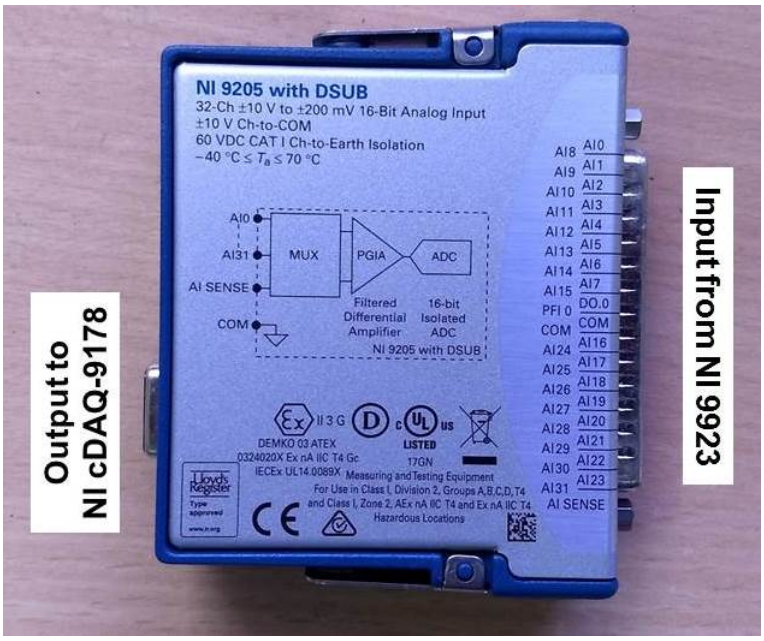


Figure 5.6: NI 9205 module

NI cDAQ 9178 - CompactDAQ Chassis is developed for small, portable systems for sensor measurement. It provides the simple plug and play functionality of USB in the application area of sensing electrical measurements. The synchronisation, timing, and data transfer between I/O modules and an external computer is also controlled by the chassis. The figure of the chassis is given in Figure 5.7.

The full diagram showing the entire DAQ connections is presented in Figure 5.8.

5.2.5 Computer

The specifications of the SEMG data acquisition computer used here is given below. The software requirement for the acquisition is also provided. Processor : Intel(R) Core(TM) i7-4770 @3.40 GHz
RAM : 24.00 GB
System Type : 64-bit Operating System, x64-based processor
Software : LabView

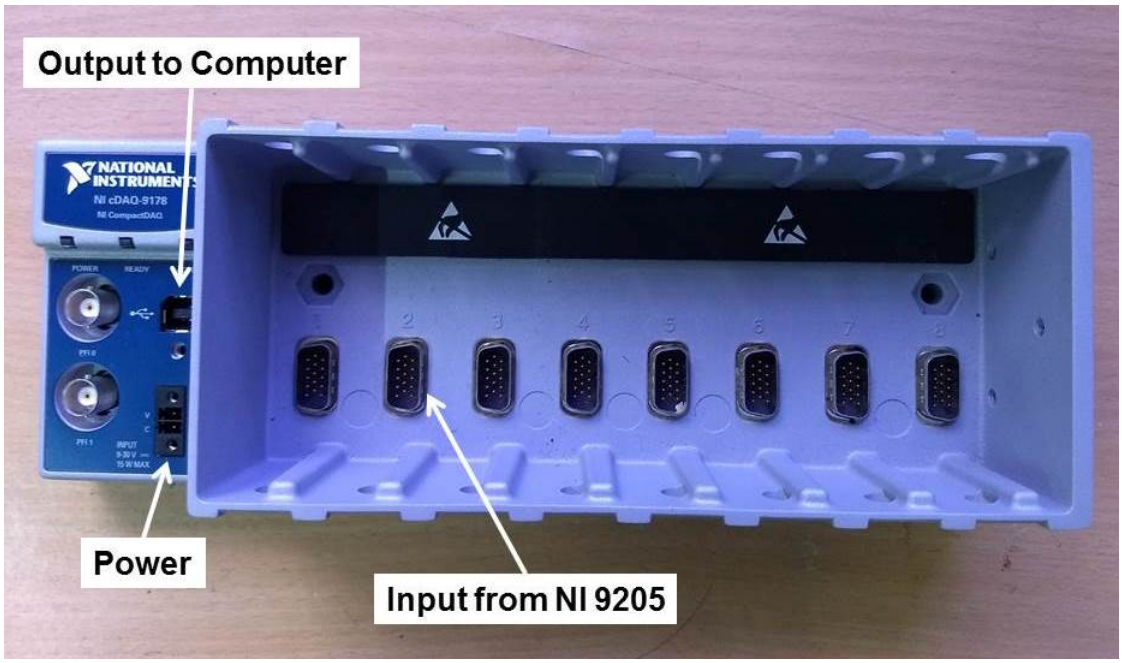


Figure 5.7: NI cDAQ 9178 - compactDAQ chassis

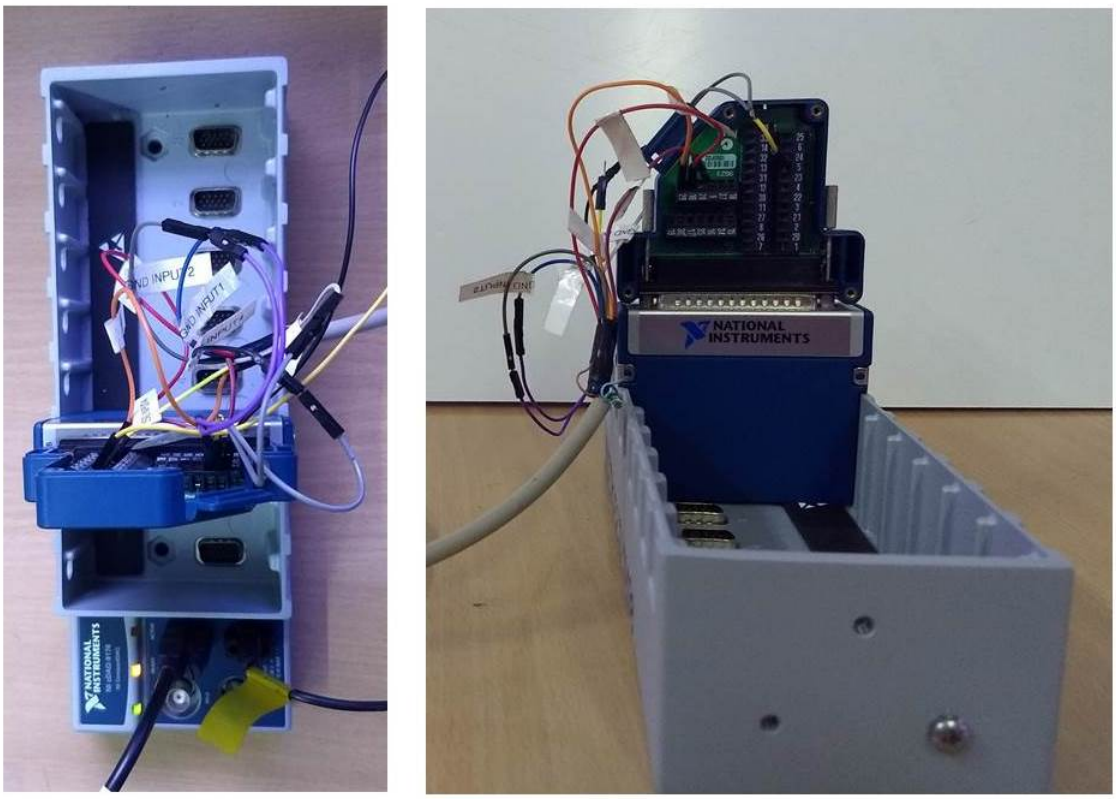


Figure 5.8: NI DAQ connections

5.3 Hardware Components for Video/Audio

A Basler acA720-290gc GigE camera having a Sony IMX287 CMOS sensor is used for video acquisition and it is able to provide 291 frames/second with VGA resolution. It has an HxV resolution of 720px x 540px with a pixel size of 6.9 µm x 6.9 µm. A C series lens with a fixed focal length of 16mm is used in the camera. The detailed technical specifications of the camera and the lens is given in Appendix. A tripod stand is used to maintain the height of the camera such that it is in level with the subject's face.

A Boya ByM1 omnidirectional lavalier condenser microphone is used for acoustic data acquisition. It is a clip on microphone with an auxiliary connectivity and is ideal for use along with video.

5.4 Data Acquisition Methodology

Data acquisition methodology refers to the steps and protocols that needs to be followed for the successful acquisition of data, especially when it is concerning human beings. The basic objective is to acquire SEMG data of silently uttered speech. Audible speech as well as video recordings of the speech articulation is also acquired for supplementing the SEMG data as well as for future research prospects. The detailed methodology associated with all the three modes of data acquisition is described in this section.

5.4.1 SEMG Data

The acquisition of SEMG data has several challenges associated with it. So proper procedures have to be ensured during all related activities, i.e. while setting up the hardware, during data recording, and during the processing of acquired raw data.

Important Points to Note while Setting up the Hardware

1. The subjects involved in the experiment should not have facial hair during data acquisition. Facial hair can cause issues with the attachment of electrodes thereby creating errors in the data acquired.

2. During electrode positioning on the face, it has to be ensured that there is sufficient distance between the electrodes in order to minimise cross talk

between different facial muscles. Also the distance should not be more such that the target muscle is missed.

3. It has to be ensured that all reference electrodes (this includes the reference electrodes of unipolar electrodes as well as the reference electrode for the entire data acquisition process) are positioned on any part that is relatively stationary as compared to the speech muscles.

4. It is better to recruit subjects who are not native speakers of English so as to account for different accents.

Important Steps Involved in Data Acquisition

1. The acquisition of SEMG has to be synchronised with parallel audio/video data capture.

2. The sentences to be read are to be displayed in front of the subject without causing run time inconvenience to the subject. This has to be done without hindering video capture as well.

3. The SEMG acquisition has to be monitored in real time to ensure that the process is running properly.

5.4.2 Audio Data

Important Points to Note while Acquiring Audio Data

1. A close talking microphone can be employed to acquire the acoustic data. The data can be acquired in stereo format.

2. The synchronisation of the audio data with the SEMG data can be done by using a hardware marker signal that is stored as the last channel for the SEMG data and as the second channel (stereo) for the audio data. For SEMG data, the marker is a binary signal, and for acoustic data it is an analog signal. The marker's first peak in both signals has to be used for synchronisation, which marks the same time point in both signals. The synchronisation location of both signals can also be pre computed in terms of samples for easier usage.

3. The data acquisition has to be done in a quiet room. There is no need for providing electrical shielding since a real life situation is preferred instead of having a specialised room for recording.

5.4.3 Video Data

The purpose of acquiring video of speech utterances is to use for future research prospects. As seen in literature several new methods have been developed where visual data is used to develop hybrid silent speech recognition models.

Important Points to Note while Acquiring Video Data

1. A background screen has to be placed behind the subject such that the acquired facial video is clear. The colour of the screen is preferred to be white

2. Proper lighting has to be ensured in the recording room to have good quality video recordings that is devoid of any issues caused due to shadows, dark areas, and reflections.

3. The synchronisation of the video data with the SEMG data can be done using the same hardware marker signal that is stored as the last channel for the SEMG data. The first peak of the SEMG marker signal and the first facial speech motion of the recorded video can be used for synchronization.

5.4.4 Data Synchronisation

The SEMG data and acoustic data can be synchronised using a hardware marker signal that is stored as the second channel (stereo) of the audio and as the seventh channel in the SEMG data. The marker signal comes as an analog signal for acoustic data and as a binary signal for SEMG data. For synchronisation, the first peak of the signal is used.

The synchronisation between the audio data and video data can be performed using simultaneous data acquisition of both using the same computer. Since the acoustic data and SEMG are synchronised using a hardware marker signal, it is then easy to synchronise the video data with SEMG data.

5.4.5 Data Alignment Methodology

The data alignment for the acoustic based recordings (audible speech) can either be hand made or by the forced alignment of the audio recordings of the dataset with a standard audible speech recognizer (S.-C. Jou et al., 2006). However, on the silent speech data, the alignments can be computed using SEMG based recognizers that are session dependent and are trained on corresponding acoustic data. An

approach of cross modal labelling described in (Janke, Wand, & Schultz, 2010) can be employed for this purpose.

5.5 Sample Data Collection

A sample dataset is obtained using the hardware setup discussed in this chapter. The protocols laid out in the data acquisition methodology were followed during the process. Two subjects, one male and one female, were involved in the data collection. Both subjects were recruited from the National Institute of Technology Karnataka student community and are non native speakers of English language. But it was ensured that the words were correctly pronounced. A total of 10 English sentences were uttered by both subjects. Each sentence was repeated five times, out of which two times audible speech was performed and in the remaining three utterances silent articulation was done. The video data corresponding to all the utterances was recorded for future research purposes. The overall data acquisition setup is given in Figure 5.9.

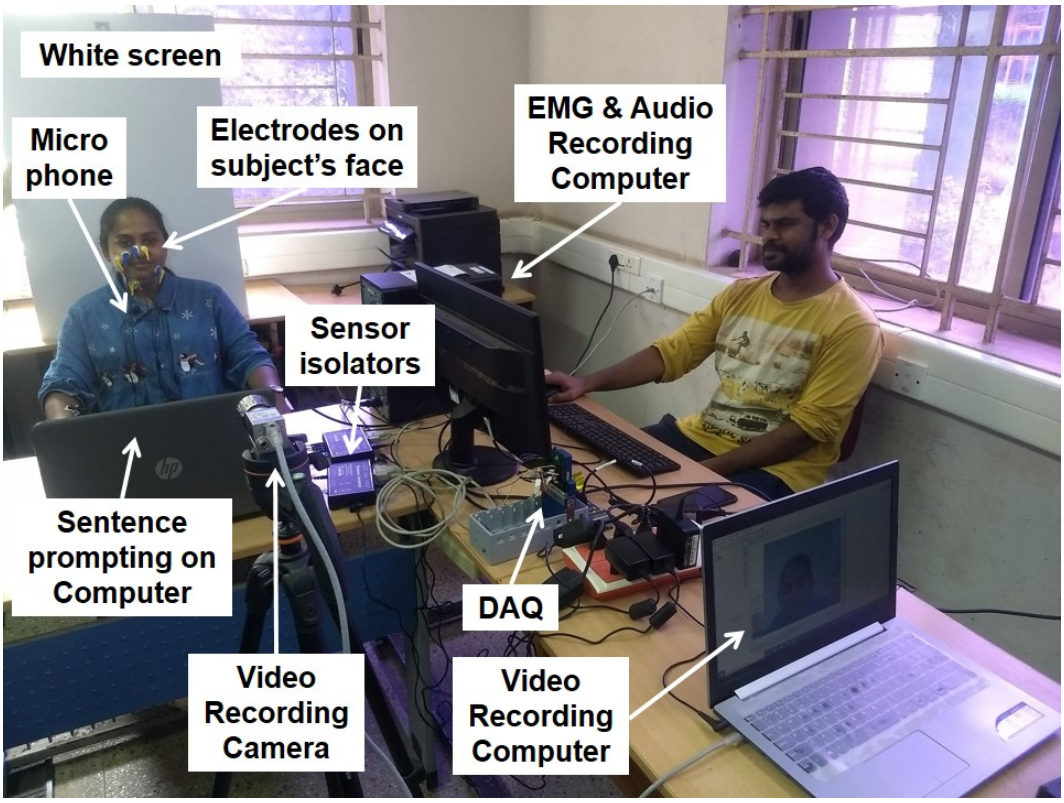


Figure 5.9: Data Acquisition Setup

5.5.1 SEMG Data Acquisition

A five channel (including reference) electrode setup is used for the sample data collection. All channels were unipolar and each channel has three electrodes ('+','-', and 'ground'). The ground electrodes of all channels were attached to the subject's wrist. The reference channel electrodes were placed on the subject's nose to ensure that there was no relative motion of the reference channel electrodes. The electrode positioning for the sample data acquisition is shown in Figure 5.10. The SEMG signals from six articulatory muscles - anterior belly of digastric (channel 1), zygomaticus major (channel 2), levator anguli oris (channel 2), depressor anguli oris (channel 5), tongue (channel 1,6), platysma (channel 5) - were captured.

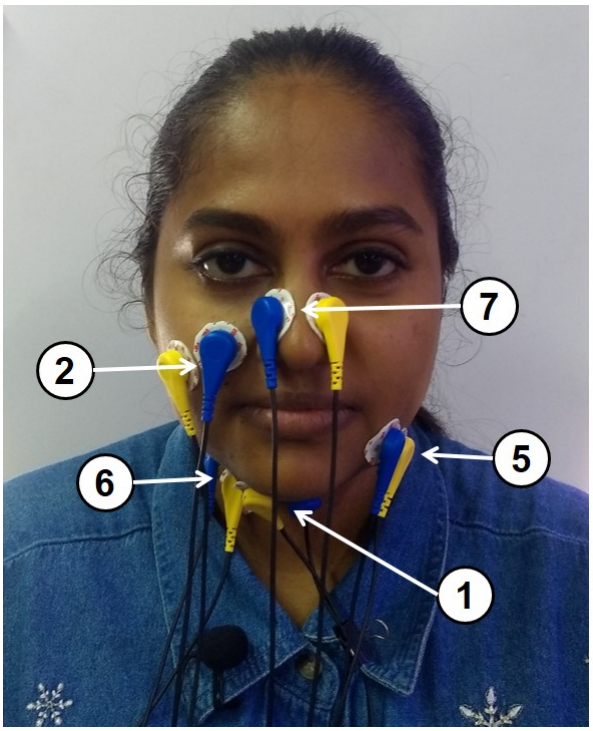


Figure 5.10: Electrode Positioning

5.5.2 Audio Data Acquisition

The audio data associated with the sample data collection was performed using a standard clip on microphone. A close talking microphone was avoided primarily for two reasons. One, to relieve the subject of an additional equipment placed on the head. Two, to minimize objects that hinder smooth recording of video.

5.5.3 Video Data Acquisition

The video data was recorded using a Basler acA720-290gc GigE camera having a resolution of 720 px x 540 px. It used a C series lens with a fixed focal length of 16 mm. The camera was placed at a slightly higher level than the subject's face in order to prevent the video getting blocked due to the presence of the sentence prompting screen. The angle of the camera was adjusted in such a way that the facial movements are accurately captured.

5.5.4 Data Alignment

An example is provide in Figure 5.11 to demonstrate how the alignment of EMG data is done. The sentence under consideration is "HE MEANS THAT LAST PART". The EMG signal consists of 1932 samples. The sampling rate used is 600 Hz, therefore this comes to approximately 3.22 seconds. The marker positions are given by two offset values, one at the starting and the next at the end. The offset values of the example under consideration are 152 and 1921. The starting points are computed from the marker signal with an additional shift of 0.2 seconds to cut out the marker itself. The ending points are calculated such that the acoustic and EMG signals have same length. In the given example there are 1769 samples remaining for the EMG signal.

The word level alignment of this sentence obtained with the help of corresponding acoustic utterance from the same session is given as follows.

0	61	$
62	78	HE
79	121	MEANS
122	152	THAT
153	187	LAST
188	226	PART
227	302	$

It has to be noted here that the alignment boundaries are denoted in frames using a frame shift of 10ms. The signal part coming between the offset values are considered. Here in the example, there are a total of 302 frames amounting to 3.02 seconds. '$' denotes a silent part where there are no relevant facial muscle activity (in the sense of speech utterance).

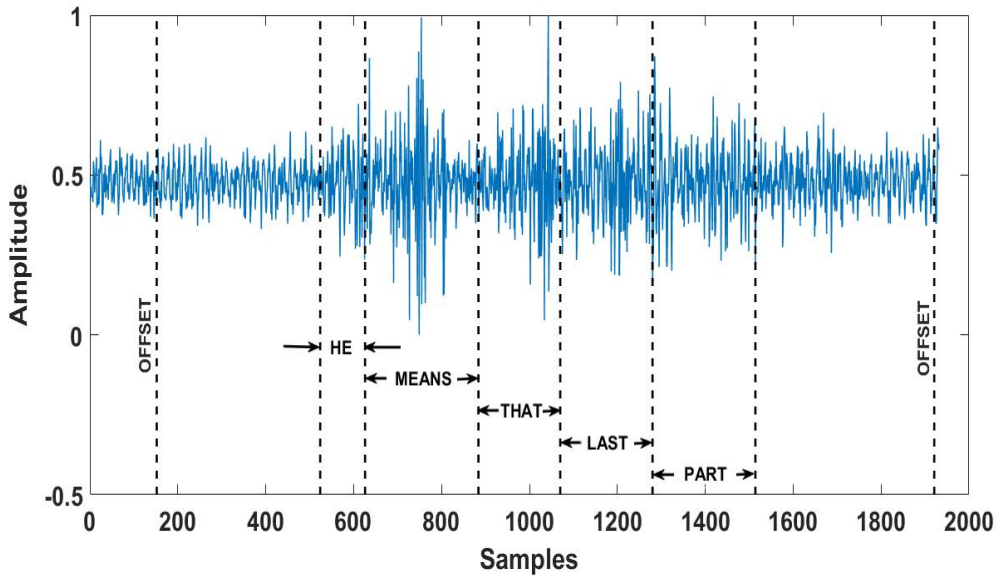


Figure 5.11: Data Alignment Example

5.6 Summary

The basic purpose of this chapter was to provide an idea about the various steps involved in the creation of an extensive silent speech database. The process starting from equipment assembly to sample data collection is explained in detail. The steps associated with audio and video data recording is also included along with SEMG data acquisition. The sample dataset acquired also covers all three modes of data. The chapter can act as a guide for carrying out data acquisition for larger vocabularies and multiple sessions of recording in the future.

Chapter 6

Conclusion and Future Work

The major outcomes of the thesis are summarized in this chapter. The recommendations for further research are also proposed based on the findings reported in this thesis.

6.1 Conclusion

Speech recognition is an important research domain in the area of human computer interaction research. However the area is dominated by acoustic speech recognition methods and comparatively less attempts are made in the area of silent speech recognition. Further, the area of silent speech recognition is mainly focused on deep learning based techniques that are complex and computationally expensive. The research presented in this thesis is an attempt to investigate more into possible feature extraction methods that can help in the realization of a computationally less expensive silent speech recognition model. The data pertaining to the silent speech used in this work is based on facial electromyography. The strategy of word based silent speech recognition was adopted instead of a phoneme based approach in order to achieve a simplified model with less computational expense.

The introduction of Detrended Fluctuation Analysis (DFA) as a feature in facial electromyography based silent speech recognition turned out to be successful in improving the word recognition accuracy of the model. The research area of silent speech recognition is dominated by time domain features and the frequency domain features like Mel Frequency Cepstral Coefficient (MFCC) found application only in the area of acoustic speech recognition. DFA being a feature that utilizes the merits of both time domain and time-frequency domain, has opened up hopes for active investigations into similar features to be used in the area of

silent speech recognition. The most promising aspect is that the use of DFA can be investigated in other research domains of human computer interaction involving similar bio signals. The improvement in word recognition accuracy as validated in both classifiers is significant to establish the superiority of DFA feature. The significance of the results are validated using appropriate statistical tests which is presented in the previous chapters.

It is also important to note that an efficient channel reduced model could be developed with the help of DFA. The attempts made to achieve channel reduction using the state-of-the-art techniques described in the literature could not offer much in terms of model accuracy. But with the addition of DFA feature along with the state-of-the-art time domain features, the model accuracy was greatly improved not only in the full channel model but also in the channel reduced model. The anchoring effect of DFA in terms of accuracy in the event of loss of some channels demonstrates the capability of the feature to perform well even if there is less data. Hence it can be a viable option in computationally less expensive pattern recognition algorithms as compared to data hungry deep learning models. The success in the implementation of an effective channel reduced model can provide a lot of relief for the subjects, especially for the laryngectomy patients. It has also opened up hopes for further investigations of better channel optimisation using techniques like 'array of electrodes', and 'hybrid-data models'.

It was also envisioned to initiate the steps for creating an effective silent speech database using SEMG. The hardware purchase and assembly, drafting of data acquisition methodology, and sample data collection were successfully completed in that direction. An extensive dataset developed by the researchers of Karlsruhe University (Wand et al., 2014) (EMG UKA Data Corpus) was purchased earlier for silent speech recognition research. The materials and methods described in the data corpus helped a lot in the whole process of data acquisition discussed in this thesis. Audio and video recordings of the sentence utterances were also obtained for future research. The data acquisition methodology enumerates the steps pertaining to each mode of data collection (SEMG, audio, and video) in detail which is helpful for future researchers working in the same domain. The sample data recorded covers all three modes of data acquisition. This can be used as the baseline for building up an extensive database.

6.2 Future Scope of Work

On the foundation of the research outcomes detailed in this thesis, the following future scope of research are presented:

1. Investigation of other time-frequency domain features to further improve the classification accuracy. A thorough literature survey in the whole area of biomedical signals and speech recognition techniques can offer several potential candidates for feature extraction. It can be useful for the whole area of human computer interaction research.

2. Further reduction of the number of channels without affecting the accuracy standards. A reduced number of electrodes can offer significant advantages such as less invasive electrode setup for the user, reduced model complexity, and faster computation.

3. Research on systems that use an array of electrodes instead of distinct electrodes on the face. Thus the processes such as fixing electrodes on the face and maintaining the electrodes can be simplified. The use of better features like DFA can compensate (in terms of model accuracy) for the loss of channels.

4. Investigations on the applicability of DFA in other areas of biomedical signal based pattern recognition problems. This can be particularly explored in the case of brain signal based speech recognition methods.

Appendix A : Technical Specifications of SEMG Data Acquisition Equipment

The technical specifications of the major components used for SEMG data acquisition in this research work is given here.

1. $3M^{TM}RedDot^{TM}$ Multi Purpose Monitoring Electrodes
2. MyoScan EMG Sensor (SA9503Z)
3. NI 9923 terminal block
4. Compact DAQ (NI 9205 module and NI cDAQ 9178 chassis)

Table 6.1: Specifications of $3M^{TM}RedDot^{TM}$ Multi Purpose Monitoring Electrodes

NPC Code		FDK001
Indication for Use		Long Term Monitoring
Dimensions	Electrode Size: Skin Contact Size: Adhesive Area: Height excluding connector:	4.06cm x 3.45cm 4.06cm x 3.45cm 10.51 sq cm 0.19cm
Electrode Materials	Backing Material: Backing Material Adhesive: Connector: Release Liner: Sensor Material:	Foam 3M™ Durapore™ Adhesive Stainless Steel Snap Si coated paper Ag/AgCl coated plastic
Sensor	Gel system: Gel area: Sensor area:	Sticky gel 3.48 sq cm 3.48 sq cm
Lifetime	Recommended max application time: Sealed pouch:	5 days 2 days
X-Ray and MRI	X-Ray: MRI:	No No
Environmental Issues	PVC-free electrode: Latex-free electrode: PVC-free packaging:	Yes Yes Yes

Table 6.2: Specifications of MyoScan EMG Sensor (SA9503Z)

Size (approx.)	37mm x 37mm x 12mm (1.45" x 1.45" x 0.45")
Weight	15g (0.5 oz)
Input impedance	$\geq$ 10GΩin parallel with 10pF
Input range	0 – 2000μVRMS
Sensitivity	<0.1μVRMS
CMRR	>130dB
Channel bandwidth	10Hz – 1kHz
Signal output range	0 – 1.0VRMS
Input / output gain	500
Supply voltage	7.26V ($\pm$ 0.02V)
Current consumption	0.7mA ($\pm$ 0.25mA)
Accuracy	$\pm$ 0.3μVRMS $\pm$ 4% of reading @ 25°C to 30°C

Table 6.3: Specifications of Sensor Isolator (SE9405AM)

Size	5.7 x 3.6 x 1.2 in (14.5 x 9 x 3 cm)
Weight	180g
Isolation Voltage	4.5kVrms
Voltage Input Range	2.8V ± 1.5V
Bandwidth	0 – 1kHz
Voltage Output Range, normal	2.8V ± 1.5V
Voltage Output Range (possible)	0 – 9V (connected device should tolerate this range)
Input impedance	1.81MΩ
Output impedance	110Ω
Accuracy	Gain: ±0.1% Offset <1mV
Noise	<100μV RMS
Temperature range (operating)	10 - 40 °C
Crosstalk	<-90dB or better
Power supply	Isolated area Examinee-applied part: 9V Alkaline battery (6LR61) Battery Life: 10 hours typical Low battery threshold: 7.25V
	Mains-connected part: 9V AC adapter

Table 6.4: Specifications of NI 9923 terminal block

Maximum voltage		30 Vrms/42 Vpk/60 VDC
Operating temperature		–40 °C to 70 °C
Maximum current rating by temperature	Pins 9, 10, 19, 28, 29 -40 °C to 55 °C: 56 °C to 70 °C:	≤5 A ≤4 A
	Pins 1–8, 11–18, 20–27, 30–37 –40 °C to 55 °C : 56 °C to 70 °C:	≤1 A ≤1 A
Screw-terminal wiring		16 AWG to 26 AWG copper conductor wire with 4.5 mm (0.18 in.) of insulation stripped from the end
Torque for screw terminals		0.4 N · m max (3.4 lb · in. max)
Weight		82 g (2.9 oz)

Table 6.5: Specifications of Compact DAQ

Number of channels	16 differential/32 single-ended channels
ADC resolution	16 bits
DNL	No missing codes guaranteed
Conversion time (maximum sampling rate) CompactRIO & CompactDAQ chassis: R Series Expansion chassis:	 4.00 μs (250 kS/s) 4.50 μs (222 kS/s)
Input coupling	DC
Nominal input ranges	$\pm$10 V, $\pm$5 V, $\pm$1 V, $\pm$0.2 V
Minimum overrange, $\pm$10 V range	4%
Maximum working voltage for analog inputs (signal + common mode)	Each channel must remain within $\pm$10.4 V of COM
Input impedance (AI-to-COM) Powered on: Powered off/overload:	 >10 GΩin parallel with 100 pF 4.7 kΩminimum
Input bias current	$\pm$100 pA
Crosstalk, at 100 kHz Adjacent channels: Non-adjacent channels:	 -65 dB -70 dB
Analog bandwidth	370 kHz
Overvoltage protection AI channel, 0 to 31: AISENSE:	 $\pm$30 V, one channel only $\pm$30 V
Settling time for multichannel measurements, accuracy, all ranges $\pm$120 ppm of full-scale step, $\pm$8 LSB: $\pm$30 ppm of full-scale step, $\pm$2 LSB:	 4 μs convert interval 8 μs convert interval
Analog triggers Number of triggers: Resolution: Bandwidth, -3 dB: Accuracy:	 1 10 bits, 1 in 1,024 370 kHz $\pm$1% of full scale
Scaling coefficients $\pm$10 V range: $\pm$5 V range: $\pm$1 V range: $\pm$0.2 V range:	 328 μV/LSB 164 μV/LSB 32.8 μV/LSB 6.57 μV/LSB
CMRR, DC to 60 Hz	100 dB

Appendix B: Technical Specifications of Video/Audio Acquisition Equipment

The technical specifications of the components used for Video/Audio acquisition in this research work is given here.

Table 6.6: Specifications of Basler acA720-290gc GigE camera

Sensor Vendor	Sony
Sensor	IMX287
Shutter	Global Shutter
Sensor Format	1/2.9”
Sensor Type	CMOS
Sensor Size	5 mm x 3.7 mm
Resolution (HxV)	720 px x 540 px
Resolution	VGA
Pixel Size (H x V)	6.9 µm x 6.9 µm
Frame Rate	291 fps
Mono/Color	Color
Interface	GigE
Pixel Bit Depth	10 / 12 bits
Synchronization	Ethernet connection, hardware trigger, free-run
Exposure Control	programmable via the camera API, hardware trigger
Digital Input	1
Digital Output	1
General Purpose I/O	1
Power Supply	PoE or 12-24 VDC
Power Requirements (typical)	2.9 W
Power Consumption PoE	3.2 W
Housing Type	Box
Housing Size (L x W x H)	42 mm x 29 mm x 29 mm
Lens Mount	C-mount
Operating Temperature	0 - 50°C
Weight (typical)	90g
Conformity	RoHS, FCC Class B, CE, IP30, GenICam, GigE Vision, IEEE802.3af(PoE), UL, KC, EAC

Table 6.7: Specifications of C series lens

Iris Option:	Variable
Length (mm):	40.50
Horizontal Field of View, 1/2" Sensor:	22.7°
Horizontal Field of View, 2/3" Sensor:	31.0°
Filter Thread:	M25.5 x 0.50 (Female)
Maximum Diameter (mm):	33
Outer Diameter (mm):	33.0
Weight (g):	74
Horizontal Field of View, 1/3" Sensor:	17.1°
Maximum Rear Protrusion (mm):	1
Maximum Image Circle (mm):	11.00
Numerical Aperture NA, Object Side:	0.040
Number of Elements (Groups):	7 (6)
Horizontal Field of View, 1/1.8" Sensor:	25.5°
Horizontal Field of View, 1/2.5" Sensor:	20.6°
Type:	Fixed Focal Length Lens
Focal Length FL (mm):	16.00
Primary Magnification PMAG:	0.145
Maximum Sensor Format:	2/3"
Working Distance (mm):	100 - ∞
Mount:	C-Mount
Aperture (f/#):	f/1.6 - f/16
Distortion (%):	<1.25
Entrance Pupil Position (mm):	19.20
Horizontal Field of View, 1/4" Sensor:	12.8°
Object Space Principal Plane (mm):	21.31
Image Space Principal Plane (mm):	-3.00
Field of View at Max Sensor Format:	Horizontal: 31 Vertical: 23.4 Diagonal: 38.3
Maximum Distortion (%):	-0.8
Exit Pupil Position (mm):	-5.43
Lens Wavelength Range:	VIS
Imaging Lens Type:	Compact Lens
Storage Temperature (°C):	-20 to +60

Table 6.8: Specifications of Boya ByM1 Auxiliary Omnidirectional Lavalier Condenser Microphone

Type	Clip on
Product Dimensions	12.7 x 5.08 x 10.16 cm 2.5 Grams
Batteries	1 LR44 batteries required.
Item model number	BYM1
Colour	Black
Compatible Devices	Laptop, DSLR, Camcorder, Tablet, Smartphone
Connector	3.5 mm Jack, 6.35 mm Jack
Size	BYM1
Battery Type	Nickel-Zinc
Hardware Platform	PC, Audio Recorder, Smartphone
Frequency Range	65Hz 18KHz
Power Source	Battery Powered
Item Weight	2.5 g

Bibliography

Abdullah, A., & Chemmangat, K. (2020). A computationally efficient semg based silent speech interface using channel reduction and decision tree based classification. *Procedia Computer Science*, *171*, 120–129.

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... Almojil, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, *9*, 131858–131876.

Ambroz, C., Scott, A., Ambroz, A., & Talbott, E. O. (2000). Chronic low back pain assessment using surface electromyography. *Journal of occupational and environmental medicine*, 660–669.

Amini, M., Pedram, M. M., Moradi, A., Ouchani, M., et al. (2021). Diagnosis of alzheimer's disease by time-dependent power spectrum descriptors and convolutional neural network using eeg signal. *Computational and Mathematical Methods in Medicine*, *2021*.

Bachmann, D., Weichert, F., & Rinkenauer, G. (2018). Review of three-dimensional human-computer interaction with focus on the leap motion controller. *Sensors*, *18*(7), 2194.

Birkholz, P., & Neuschaefer-Rube, C. (2011). Combined optical distance sensing and electropalatography to measure articulation. In *Twelfth annual conference of the international speech communication association.*

Brody, G., Scott, R., & Balasubramanian, R. (1974). A model for myoelectric signal generation. *Medical and biological engineering*, *12*, 29–41.

Ceachir, O., Hainarosie, R., & Zainea, V. (2014). Total laryngectomy–past, present, future. *Maedica*, *9*(2), 210.

Çerçi, Ç., & Temeltaş, H. (2018). Feature extraction of emg signals, classification with ann and knn algorithms. In *2018 26th signal processing and communications applications conference (siu)* (pp. 1–4).

Chan, A. D., Englehart, K., Hudgins, B., & Lovely, D. F. (2001). Myo-electric signals to augment speech recognition. *Medical and Biological Engineering and Computing*, *39*, 500–504.

Chan, A. D. C. (2003). *Multi-expert automatic speech recognition system using myoelectric signals* (Unpublished doctoral dissertation). Ph. D. dissertation, 2003, aAINQ87627.

Chandrasekhar, V., Vazhayil, V., & Rao, M. (2020). Design of a real time portable low-cost multi-channel surface electromyography system to aid neuromuscular disorder and post stroke rehabilitation patients. In *2020 42nd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 4138–4142).

Chatterjee, S., Pratiher, S., & Bose, R. (2017). Multifractal detrended fluctuation analysis based novel feature extraction technique for automated detection of focal and non-focal electroencephalogram signals. *IET Science, Measurement & Technology*, *11*(8), 1014–1021.

Criswell, E. (2010). *Cram's introduction to surface electromyography.* Jones & Bartlett Publishers.

Crone, N. E., Miglioretti, D. L., Gordon, B., & Lesser, R. P. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. ii. event-related synchronization in the gamma band. *Brain: a journal of neurology*, *121*(12), 2301–2315.

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

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, *52*(4), 270–287.

Diener, L., Felsch, G., Angrick, M., & Schultz, T. (2018). Session-independent array-based emg-to-speech conversion using convolutional neural networks. In *Speech communication; 13th itg-symposium* (pp. 1–5).

Diener, L., Herff, C., Janke, M., & Schultz, T. (2016). An initial investigation into the real-time conversion of facial surface emg signals to audible speech. In *2016 38th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 888–891).

Diener, L., Janke, M., & Schultz, T. (2015). Direct conversion from facial myoelectric signals to speech using deep neural networks. In *2015 international joint conference on neural networks (ijcnn)* (pp. 1–7).

Diener, L., & Schultz, T. (2018). Investigating objective intelligibility in real-time emg-to-speech conversion. In *Interspeech* (pp. 3162–3166).

Dikshit, P. S., & Schubert, R. W. (1995). Electroglottograph as an additional source of information in isolated word recognition. In *Proceedings of the*

1995 fourteenth southern biomedical engineering conference (pp. 1–4).

Dressing, A., Kaller, C. P., Nitschke, K., Beume, L.-A., Kuemmerer, D., Schmidt, C. S., ... others (2019). Neural correlates of acute apraxia: Evidence from lesion data and functional mri in stroke patients. *Cortex*, *120*, 1–21.

D'Zmura, M., Deng, S., Lappas, T., Thorpe, S. G., & Srinivasan, R. (2009). Toward eeg sensing of imagined speech. In *Hci (1)* (pp. 40–48).

Fridriksson, J., Bonilha, L., Baker, J. M., Moser, D., & Rorden, C. (2010). Activity in preserved left hemisphere regions predicts anomia severity in aphasia. *Cerebral cortex*, *20*(5), 1013–1019.

Gilbert, J. M., Rybchenko, S. I., Hofe, R., Ell, S. R., Fagan, M. J., Moore, R. K., & Green, P. (2010). Isolated word recognition of silent speech using magnetic implants and sensors. *Medical engineering & physics*, *32*(10), 1189–1197.

Goncharova, I. I., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2003). Emg contamination of eeg: spectral and topographical characteristics. *Clinical neurophysiology*, *114*(9), 1580–1593.

Gupta, S., Patil, A. T., Purohit, M., Parmar, M., Patel, M., Patil, H. A., & Guido, R. C. (2021). Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, *139*, 105–117.

Hashimoto, M., Takahashi, K., & Shimada, M. (2009). Wheelchair control using an eog-and emg-based gesture interface. In *2009 ieee/asme international conference on advanced intelligent mechatronics* (pp. 1212–1217).

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Heracleous, P., Badin, P., Bailly, G., & Hagita, N. (2011). A pilot study on augmented speech communication based on electro-magnetic articulography. *Pattern Recognition Letters*, *32*(8), 1119–1125.

Hilfiker, P., & Meyer, M. (1984). Normal and myopathic propagation of surface motor unit action potentials. *Electroencephalography and clinical neurophysiology*, *57*(1), 21–31.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832–844.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural*

computation, *9*(8), 1735–1780.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, *52*(4), 288–300.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech recognition using mfcc. In *International conference on computer graphics, simulation and modeling* (pp. 135–138).

Janke, M., & Diener, L. (2017). Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2375–2385.

Janke, M., Wand, M., & Schultz, T. (2010). A spectral mapping method for emg-based recognition of silent speech. In *B-interface* (pp. 22–31).

Jeffrey, L. E., et al. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Jorgensen, C., & Binsted, K. (2005). Web browser control using emg based sub vocal speech recognition. In *Proceedings of the 38th annual hawaii international conference on system sciences* (pp. 294c–294c).

Jorgensen, C., Lee, D. D., & Agabont, S. (2003). Sub auditory speech recognition based on emg signals. In *Proceedings of the international joint conference on neural networks, 2003.* (Vol. 4, pp. 3128–3133).

Jou, S., Schultz, T., & Waibel, A. (2007). Multi-stream articulatory feature classifiers for surface electromyographic continuous speech recognition. In *Internat. conf. on acoustics, speech, and signal processing. ieee, honolulu, hawaii.*

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., & Waibel, A. (2006). Towards continuous speech recognition using surface electromyography. In *Ninth international conference on spoken language processing.*

Kapila, M., Deore, N., Palav, R., Kazi, R., Shah, R., & Jagade, M. (2011). A brief review of voice restoration following total laryngectomy. *Indian journal of cancer*, *48*(1), 99–104.

Khadivi, A., Nazarpour, K., & Zadeh, H. S. (2005). Semg classification for upper-limb prosthesis control using higher order statistics. In *Proceedings.(icassp'05). ieee international conference on acoustics, speech, and signal processing, 2005.* (Vol. 5, pp. v–385).

Khushaba, R. N., Takruri, M., Miro, J. V., & Kodagoda, S. (2014). Towards limb position invariant myoelectric pattern recognition using time-dependent

spectral features. *Neural Networks*, *55*, 42–58.

Kim, M., Cao, B., Mau, T., & Wang, J. (2017). Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network. *IEEE/ACM transactions on audio, speech, and language processing*, *25*(12), 2323–2336.

Krishna, G., Tran, C., Han, Y., Carnahan, M., & Tewfik, A. H. (2020). Speech synthesis using eeg. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1235–1238).

Larson, E., Terry, H. P., & Stepp, C. E. (2012). Audio-visual feedback for electromyographic control of vowel synthesis. In *2012 annual international conference of the ieee engineering in medicine and biology society* (pp. 3600–3603).

Ma, S., Jin, D., Zhang, M., Zhang, B., Wang, Y., Li, G., & Yang, M. (2019). Silent speech recognition based on surface electromyography. In *2019 chinese automation congress (cac)* (pp. 4497–4501).

Maier-Hein, L., Metze, F., Schultz, T., & Waibel, A. (2005). Session independent non-audible speech recognition using surface electromyography. In *Ieee workshop on automatic speech recognition and understanding, 2005.* (pp. 331–336).

Makin, J. G., Moses, D. A., & Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, *23*(4), 575–582.

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, *80*, 9411–9457.

Manabe, H., Hiraiwa, A., & Sugimura, T. (2003). Unvoiced speech recognition using emg-mime speech recognition. In *Chi'03 extended abstracts on human factors in computing systems* (pp. 794–795).

Manabe, H., & Zhang, Z. (2004). Multi-stream hmm for emg-based speech recognition. In *The 26th annual international conference of the ieee engineering in medicine and biology society* (Vol. 2, pp. 4389–4392).

Meekins, G. D., So, Y., & Quan, D. (2008). American association of neuromuscular & electrodiagnostic medicine evidenced-based review: Use of surface electromyography in the diagnosis and study of neuromuscular disorders. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, *38*(4), 1219–1224.

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., & Kline, J. C.

(2017). Silent speech recognition as an alternative communication device for persons with laryngectomy. *IEEE/ACM transactions on audio, speech, and language processing*, *25*(12), 2386–2398.

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., & Kline, J. C. (2018). Development of semg sensors and algorithms for silent speech recognition. *Journal of neural engineering*, *15*(4), 046031.

Miyagi, M., Nishida, M., Horiuchi, Y., & Ichikawa, A. (2006). Analysis of prosody in finger braille using electromyography. In *2006 international conference of the ieee engineering in medicine and biology society* (pp. 4901–4904).

Morse, M., Gopalan, Y., & Wright, M. (1991). Speech recognition using myoelectric signals with neural networks. In *Proceedings of the annual international conference of the ieee engineering in medicine and biology society volume 13: 1991* (pp. 1877–1878).

Morse, M. S., Day, S. H., Trull, B., & Morse, H. (1989). Use of myoelectric signals to recognize speech. In *Images of the twenty-first century. proceedings of the annual international engineering in medicine and biology society,* (pp. 1793–1794).

Morse, M. S., & O'Brien, E. M. (1986). Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. *Computers in biology and medicine*, *16*(6), 399–410.

Mumtaz, R., Preuß, S., Neuschaefer-Rube, C., Hey, C., Sader, R., & Birkholz, P. (2014). Tongue contour reconstruction from optical and electrical palatography. *IEEE Signal Processing Letters*, *21*(6), 658–662.

Muro, M., Nagata, A., Murakami, K., & Moritani, T. (1982). Surface emg power spectral analysis of neuro-muscular disorders during isometric and isotonic contractions. *American Journal of Physical Medicine & Rehabilitation*, *61*(5), 244–254.

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of eeg.* Oxford University Press, USA.

Ouisper. (2006-2009). *Oral ultrasound synthetic speech source.* National Research Agency (ANR), France.

Petajan, E. (1984). Automatic lip reading to enhance speech recognition [ph. d. dissertation]. *Illinois: University of Illinois at Urbana-Champaign*.

Phinyomark, A., Phukpattaranont, P., & Limsakul, C. (2012). Fractal analysis features for weak and single-channel upper-limb emg signals. *Expert Systems with Applications*, *39*(12), 11156–11163.

Phinyomark, A., Quaine, F., Charbonnier, S., Serviere, C., Tarpin-Bernard, F., & Laurillau, Y. (2013). Emg feature evaluation for improving myoelectric pattern recognition robustness. *Expert Systems with applications*, *40*(12), 4832–4840.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, *91*(9), 1306–1326.

Potamianos, G., Neti, C., Huang, J., Connell, J. H., Chu, S., Libal, V., ... Jiang, J. (2004). Towards practical deployment of audio-visual speech recognition. In *2004 ieee international conference on acoustics, speech, and signal processing* (Vol. 3, pp. iii–777).

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., ... others (2011). The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, *25*(2), 404–439.

Powar, O. S., & Chemmangat, K. (2019). Dynamic time warping for reducing the effect of force variation on myoelectric control of hand prostheses. *Journal of Electromyography and Kinesiology*, *48*, 152–160.

Powar, O. S., & Chemmangat, K. (2020). Reducing the effect of wrist variation on pattern recognition of myoelectric hand prostheses control through dynamic time warping. *Biomedical Signal Processing and Control*, *55*, 101626.

Powar, O. S., Chemmangat, K., & Figarado, S. (2018). A novel pre-processing procedure for enhanced feature extraction and characterization of electromyogram signals. *Biomedical Signal Processing and Control*, *42*, 277–286.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T. (1994). *Human-computer interaction*. Addison-Wesley Longman Ltd.

Ramaekers, V. T., Disselhorst-Klug, C., Schneider, J., Silny, J., Forst, J., Forst, R., ... Rau, G. (1993). Clinical application of a noninvasive multi-electrode array emg for the recording of single motor unit activity. *Neuropediatrics*, *24*(03), 134–138.

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69). World scientific.

Rush, P. A. (2013). Voice rehabilitation following laryngectomy. In S. E. Kountakis (Ed.), *Encyclopedia of otolaryngology, head and neck surgery* (pp. 3046–3055). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-23499-6_126 doi: 10.1007/978-3-642-23499-6_126

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B.

(1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, *31*(1), 26–35.

Schultz, T., & Wand, M. (2010). Modeling coarticulation in emg-based continuous speech recognition. *Speech Communication*, *52*(4), 341–353.

Schünke, M., & Schulte, E. S. (2006). *U.[2006] prometheus lernatlas der anatomie. kopf und neuroanatomie.* Georg Thieme Verlag KG, Stuttgart.

Sebkhi, N., Sahadat, N., Hersek, S., Bhavsar, A., Siahpoushan, S., Ghoovanloo, M., & Inan, O. T. (2019). A deep neural network-based permanent magnet localization for tongue tracking. *IEEE Sensors Journal*, *19*(20), 9324–9331.

Stepp, C. E., Heaton, J. T., Rolland, R. G., & Hillman, R. E. (2009). Neck and face surface electromyography for prosthetic voice control after total laryngectomy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *17*(2), 146–155.

Stepp, C. E., Hillman, R. E., & Heaton, J. T. (2010). Use of neck strap muscle intermuscular coherence as an indicator of vocal hyperfunction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *18*(3), 329–335.

Stone, S., & Birkholz, P. (2016a). Angle correction in optopalatographic tongue distance measurements. *IEEE Sensors Journal*, *17*(2), 459–468.

Stone, S., & Birkholz, P. (2016b). Silent-speech command word recognition using electro-optical stomatography. In *Interspeech* (pp. 2350–2351).

Stone, S., & Birkholz, P. (2020). Cross-speaker silent-speech command word recognition using electro-optical stomatography. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7849–7853).

Sugie, N., & Tsunoda, K. (1985). A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production. *IEEE Transactions on Biomedical Engineering*(7), 485–490.

Suppes, P., Lu, Z.-L., & Han, B. (1997). Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, *94*(26), 14965–14969.

Toutios, A., & Margaritis, K. (2008). Estimating electropalatographic patterns from the speech signal. *Computer Speech & Language*, *22*(4), 346–359.

Vigotsky, A. D., Halperin, I., Lehman, G. J., Trajano, G. S., & Vieira, T. M. (2018). Interpreting signal amplitudes in surface electromyography studies in sport and rehabilitation sciences. *Frontiers in physiology*, 985.

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., & Waibel, A. (2006). Sub-word unit based non-audible speech recognition using surface electromyography.

In *Ninth international conference on spoken language processing.*

Wand, M., Janke, M., & Schultz, T. (2014). The emg-uka corpus for electromyographic speech processing. In *Fifteenth annual conference of the international speech communication association.*

Wand, M., & Schmidhuber, J. (2016). Deep neural network frontend for continuous emg-based speech recognition. In *Interspeech* (pp. 3032–3036).

Wand, M., Schulte, C., Janke, M., & Schultz, T. (2013). Array-based electromyographic silent speech interface. In *Biosignals* (pp. 89–96).

Wand, M., & Schultz, T. (2009). Speaker-adaptive speech recognition based on surface electromyography. In *International joint conference on biomedical engineering systems and technologies* (pp. 271–285).

Wand, M., & Schultz, T. (2014). Towards real-life application of emg-based speech recognition by using unsupervised adaptation. In *Fifteenth annual conference of the international speech communication association.*

Wang, J., Balasubramanian, A., de la Vega, L. G. M., Green, J. R., Samal, A., & Prabhakaran, B. (2013). Word recognition from continuous articulatory movement time-series data using symbolic representations. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies* (pp. 119–127).

Wang, X., Zhu, M., Cui, H., Yang, Z., Wang, X., Zhang, H., ... Li, G. (2020). The effects of channel number on classification performance for semg-based speech recognition. In *2020 42nd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 3102–3105).

Wang, Y., Zhang, M., Wu, R., Gao, H., Yang, M., Luo, Z., & Li, G. (2020). Silent speech decoding using spectrogram features based on neuromuscular activities. *Brain sciences*, *10*(7), 442.

Webster, J. G. (2009). *Medical instrumentation: application and design.* John Wiley & Sons.

Westerink, J. H., Van Den Broek, E. L., Schut, M. H., Van Herk, J., & Tuinenbreijer, K. (2008). Computing emotion awareness through galvanic skin response and facial electromyography. *Probing experience: From assessment of user emotions and behaviour to development of products*, 149–162.

Wolpaw, J., Birbaumer, N., McFarland, D., Pfurtscheller, G., & Vaughan, T. (2002). *Brain-computer interfaces for communication and control. clinocal neurophysiology, 113, 767-791.*

Wrench, A. A., McIntosh, A. D., Watson, C., & Hardcastle, W. J. (1998). Optopalatograph: Real-time feedback of tongue movement in 3d. In *Proceedings*

of the 5th international conference in spoken language processing.

Yau, W. C., Arjunan, S. P., & Kumar, D. K. (2008). Classification of voiceless speech using facial muscle activity and vision based techniques. In *Tencon 2008-2008 ieee region 10 conference* (pp. 1–6).

Yu, D., & Deng, L. (2016). *Automatic speech recognition* (Vol. 1). Springer.

Zhang, M., Wang, Y., Zhang, W., Yang, M., Luo, Z., & Li, G. (2020). Inductive conformal prediction for silent speech recognition. *Journal of neural engineering*, *17*(6), 066019.

Publications based on the thesis

Papers in refereed journal

1. Asif Abdullah and Krishnan Chemmangat, "A Computationally Efficient sEMG based Silent Speech Interface using Channel Reduction and Decision Tree based Classification", *Procedia Computer Science*, 2020, 171:120–129.

2. Asif Abdullah, Omkar S Powar, and Krishnan Chemmangat. "Application of fractal analysis based feature extractor for channel reduction of silent speech interface using facial electromyography", *International Journal of Intelligent Engineering and Systems*, 2023, 16(3):428-439.

3. Asif Abdullah, Omkar S Powar and Krishnan Chemmangat, "Evaluation of fractal analysis as feature extractors for silent speech interface using facial electromyography" (Under Review).

Bio-Data

Name : Asif Abdullah
Date of birth: 17-10-1992
Nationality: Indian
Marital status: Married
E-mail: asifabdullah92@gmail.com
Mobile: +91-8147515174

Address:
Kaleeckal House
Opp MSM College
Kayamkulam
Alappuzha, Kerala, India, PIN-690502

Software Skills: C, C++, MATLAB/SIMULINK

Education:

- M.Tech. in Power and Energy Systems from National Institute of Technology, Surathkal, Karnataka with a C.G.P.A. of 7.35/10 during 2015-17.
- B.Tech. in Electrical and Electronics Engineering from College of Engineering Chengannur, Alappuzha, Kerala with a percentage of 72.01 during 2010-14.
- SSCE (12th) in Science from Jawahar Navodaya Vidyalaya Chennithala, Alappuzha, Kerala with a percentage of 92.8 in the year 2010.
- SSE (10th) in Science from Jawahar Navodaya Vidyalaya Chennithala, Alappuzha, Kerala with a percentage of 96.2 in the year 2008.