# VISVESVARAYA TECHNOLOGICAL UNIVERSITY

**JNANA SANGAMA, BELAGAVI - 590018**



**TECHNICAL SEMINAR REPORT (18ECS84)**

**ON**

## “VISUAL SPEECH RECOGNITION”

*Submitted in partial fulfilment of the requirements for the award of the degree of*

**BACHELOR OF ENGINEERING**

**IN**

**ELECTRONICS AND COMMUNICATION**

**For the Academic year 2022-2023**

*Report Submitted by*

**GUDIKAL SAI VAMSI (1MV19EC041)**

***Under the guidance of***

**Mrs. KRISHNAPRIYA S SHARMA**

Associate Professor / Department of ECE



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING SIR M. VISVESVARAYA INSTITUTE OF TECHNOLOGY**

Bengaluru -562157

## MARCH 2023

**SIR M. VISVESVARAYA INSTITUTE OF TECHNOLOGY**

Bengaluru -562157

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**



### **CERTIFICATE**

This is to certify that the Technical Seminar (18ECS84) on **“VISUAL SPEECH RECOGNITION”** prepared by **GUDIKAL SAI VAMSI (1MV19EC041)**, a bonafide student of **SIR M. VISVESVARAYA INSTITUTE OF TECHNOLOGY**. The report is in partial fulfilment of the requirements for the award of the degree of “**Bachelor of Engineering**” in **Electronics and Communication Engineering.** From the **Visvesvaraya Technological University**, Belagavi, Karnataka, India, during the academic year 2022-2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report submitted to the Department. The Seminar report has been approved as it satisfies the academic requirement in respect to the work prescribed for the said Degree.

------------------------------- ------------------------------ ----------------------------Signature of the Guide Signature of HOD Signature of Principal

**Mrs. Krishnapriya S Sharma Dr. V. G. Supriya Prof. Rakesh**

Associate Professor Professor & Head Principal

Dept of ECE Dept of ECE Sir MVIT, Bengaluru

**DECLARATION**

I hereby declare that the Seminar Report on **“VISUAL SPEECH RECOGNITION”** undertaken has been presented under the guidance of Mrs. Krishnapriya S Sharma, Associate Professor, Department of Electronics and Communication Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru. This topic has not been submitted previously in the Dept. of ECE and any other Departments of Sir MVIT.

Place: Bengaluru GUDIKAL SAI VAMSI

Date: 28-03-2023 1MV19EC041

# ACKNOWLEDGEMENT

A technical seminar is incomplete if it fails to thank all those instrumental in the successful completion of the report.

I welcome this opportunity to convey my regards, gratitude, respect, decorations, and a lot of thanks to them who inspired me to complete this report in the stipulated period of time provided.

It’s a great privilege to place on record my deep sense of gratitude to the Management and **Prof. Rakesh S.G.**, Principal, Sir M. Visvesvaraya Institute of Technology, who patronized throughout our career & for the facilities provided to carry out this work successfully.

I would like to extend my heartfelt gratitude to **Dr. V. G. Supriya**, Professor and Head, Dept. of ECE, Sir M. Visvesvaraya Institute of Technology, for her constant support and encouragement.

I would like to thank our project guide **Mrs. Krishnapriya S Sharma**, Associate Professor**,** Dept. of ECE, Sir M. Visvesvaraya Institute of Technology, for her valuable guidance and support in the completion of this seminar.

I thank the teaching and non-teaching staff members who have helped me directly or indirectly in completion of this seminar.

Finally, I would also like to thank my parents and friends who rendered me active support for the completion of this seminar report. I acknowledge them with a lot of gratitude and regard and without all of the above this report may not be easily possible.

**GUDIKAL SAI VAMSI**

**(1MV19EC041)**

**ABSTRACT**

Everybody has the experience of talking aloud in the cell phone in the midst of the disturbance while travelling in trains or buses. There is no need of shouting anymore for this purpose. Visual Speech Recognition is the answer for this problem. The Visual Speech Recognition is an amazing solution for those who had lost their voice but wish to speak over phone. It is developed at the Karlsruhe Institute of Technology and you can expect to see it in the near future. When demonstrated, it seems to detect every lip movement and internally converts the electrical pulses into sounds signals and sends them neglecting all other surrounding noise. It is definitely going to be a good solution for those feeling annoyed when other speak loud over phone.It aims to notice every movement of the lips and transform them into sounds, which could help people who lose voices to speak, and allow people to make silent calls without bothering others. Rather than making any sounds, your handset would decipher the movements your mouth makes by measuring muscle activity, then convert this into speech that the person on the other end of the call can hear. So, basically, it reads your lips. This new technology will be very helpful whenever a person loses his voice while speaking or allow people to make silent calls without disturbing others, even we can tell our PIN number to a trusted friend or relative without eavesdropping. At the other end, the listener can hear a clear voice. The awesome feature added to this technology is that "it is an instant polyglot" can be immediately transformed into the language of the user's choice. This translation works for languages like English, French & German.

**CONTENTS**

*Certificate* i

*Declaration* ii

*Acknowledgement* iii

*Abstract* iv

*Contents* v

*List of figures*  vi

**Chapters**  **Titles**  **Page No.**

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE SURVEY 3

CHAPTER 3: PROCESS OF SPEAKING 5

CHAPTER 4 METHODS 6

CHAPTER 5: MERITS AND DEMERITS 11

CHAPTER 6: FUTURE SCOPE 12

CHAPTER 7: CONCLUSION 13

*References* 14

**LIST OF FIGURES**

**Fig No Figure Name Page No.**

Fig 1.1 Many people talking at one 2

Fig 3.1 Process of speaking 5

Fig 4.1 Electromyography signal generation 6

Fig 4.2 Electromyography Instruments 7

Fig 4.3 Block diagram of EMG 7

Fig 4.4 Interfacing with electromyographer and body 8

**CHAPTER 1**

**INTRODUCTION**

Visual speech recognition alludes to the detailed feature-based analysis on the lips and its surrounding environment. It includes various aspects of feature extraction due to the need for consideration of the exterior environment and the details that play an important role in prediction. The main technology behind this Visual speech recognition is referred as Silent sound technology.

You are in a movie theatre or noisy restaurant or a bus etc where there is lot of noise around is big issue while talking on a mobile phone. But in the future this problem is eliminated with Silent sound technology, a new technology that transforms lip movements into a computer-generated voice for the listener at the other end of the phone. It is a technology that helps you to transmit information without using your vocal cords. This technology aims to notice lip movements & transform them into a computer-generated sound that can be transmitted over a phone. Hence person on other end of phone receives the information in audio.

The device developed by the Karlsruhe Institute of Technology (KIT), uses electromyography, monitoring tiny muscular movements that occur when we speak and converting them into electrical pulses that can then be turned into speech, without a sound uttered.

This technology aims to notice every movement of the lips and transform them into sounds, which could help people who lose voices to speak, and allow people to make silent calls without bothering others. Rather than making any sounds, your handset would decipher the movements your mouth makes by measuring muscle activity, then convert this into speech that the person on the other end of the call can hear. So, basically, it reads your lips.

The technology opens up a host of applications, from helping people who have lost their voice due to illness or accident to telling a trusted friend your PIN number over the phone without anyone eavesdropping — assuming no lip-readers are around. The technology can also turn you into an instant polyglot. Because the electrical pulses are universal, they can be immediately transformed into the language of the user’s choice.

Native speakers can silently utter a sentence in their language, and the receivers hear the translated sentence in their language. It appears as if the native speaker produced speech in a foreign language. The translation technology works for languages like English, French and German, but for languages like Chinese, where different tones can hold many different meanings, poses a problem.



Fig 1.1 Many people talking at one place

Silent Sound Technology will put an end to embarrassed situation such as-

• A person answering his silent, but vibrating cell phone in a meeting, lecture or performance, and whispering loudly, ‘I can’t talk to you right now’.

• In the case of an urgent call, apologetically rushing out of the room in order to answer or call the person back.

• The technology opens up a host of applications, from helping people who have lost their voice due to illness or accident to telling a trusted friend your PIN number over the phone without anyone eavesdropping — assuming no lip-readers are around. Native speakers can silently utter a sentence in their language, and the receivers hear the translated sentence in their language. It appears as if the native speaker produced speech in a foreign language.

**CHAPTER 2**

**LITERATURE SURVEY**

[1] Human beings identify speeches of a speaker using information that can be characterized by several different modes of communication. Visual data such as lip and tongue expression, in addition to speech audio, can also help to understand speech. By using visual data that detects lip gestures and understands what a subject is saying, is referred to as lip reading. Lip reading can help to understand speeches not only in a silent environment but also in an environment with a lot of noise. Previous human-machine interaction studies have determined that visual information can improve the accuracy of speech recognition, particularly in noisy environments. In this paper [1], the authors present a model for predicting words from video data. While current models have already succeeded in integrating visual information into speech recognition, they all contained one or the other deficiency. To resolve this, they have pre-processed the data by using the VGG model to detect and crop around the subject’s face in all frames of the video data and then use the frame sequence as input to the model and train a LSTM classifier to obtain a word-level classification. The proposed model aims at recognizing terms and phrases from a broad dataset effectively.

[2] Visual speech recognition or automated lipreading is a field of growing attention. Video data proved its usefulness in multimodal speech recognition, especially when acoustic data is heavily noised or even inaccessible. In this paper [2], the authors present a novel method for visual speech recognition and benchmark it on the famous LRW lip-reading dataset by outperforming the existing approaches. After a comprehensive evaluation, they adapt the developed method and test it on the collected RUSAVIC corpus we recorded in-the-wild for vehicle driver. The results obtained demonstrate not only the high performance of the proposed method, but also the fundamental possibility of recognizing speech only by using video modality, even in such difficult natural conditions as driving.

[3]. Silent speech recognition (SSR) can decode the activities of electromyography (EMG) from surface of articulatory muscles to the according brain information. As the essentials in SSR, feature extraction and recognition are directly related to the final decoding effects, which requires careful research, especially on small datasets. This paper [3] collected the surface EMG (sEMG) of 10 isolated Chinese words by 6 surface electrodes on face and around, and based on which sEMG features were extracted by wavelet packet transform, principal component analysis and XGBoost (an implementation of gradient boosted decision trees) respectively. Then Support Vector Machine, K-Nearest Neighbor and Random Forest were explored for recognition on the 10 words. The results indicated that the highest accuracy of 72% was achieved by the method of Random Forest based on XGBoost.

[4]. A machine being able to perform lip-reading would have been deemed impossible a few decades ago. However, the exponential growth of machine learning in the past few years has made it possible for a machine to understand human speech based on visual inputs alone. Numerous research studies infer that a very less percentage of the English language can be comprehended through visual data alone, i.e., lip reading. Visual speech recognition experts can only infer about 3–4% of words spoken through lip-reading after viewing videos (without audio) multiple times. These experts also examine other parameters such as body language, facial cues, habits, and context to some extent. This task is very tedious (or exhausting). The proposed visual speech recognition approach has used the concept of deep learning to perform word-level classification. ResNet architecture is used with 3D convolution layers as the encoder and Gated Recurrent Units (GRU) as the decoder. The whole video sequence was used as an input in this approach. The results of the proposed approach are satisfactory. It achieves 90% accuracy on the BBC data set and 88% on the custom video data set. The proposed approach is limited to word-level only and can easily be extended to short phrases or sentences.

**CHAPTER 3**

**PROCESS OF SPEAKING**

* The air passes through the larynx and the tongue and the words are formed with the help of the articulator muscles in the mouth and the jaw.
* The articulator muscles are activated irrespective of the fact that air passes through them or not.
* The weak signals are sent from the brain to the speech muscle. These signals are collectively known as the electromyograms.

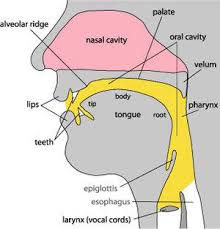
****

Fig 3.1 Process of speaking

**CHAPTER 4**

**METHODS**

Silent Sound Technology is processed through some ways or methods. They are

• Electromyography (EMG)

• Image Processing

**ELECTROMYOGRAPHY**

Electromyography (EMG) is a technique for evaluating and recording the electrical activity produced by skeletal muscles. EMG is performed using an instrument called an electromyograph, to produce a record called an electromyogram. An electromyograph detects the electrical potential generated by muscle cells when these cells are electrically or neurologically activated. The signals can be analysed to detect medical abnormalities, activation level, recruitment order or to analyse the biomechanics of human or animal movement.

• The Silent Sound Technology uses electromyography, monitoring tiny muscular movements that occur when we speak.

• Monitored signals are converted into electrical pulses that can then be turned into speech, without a sound uttered.

• Electromyography (EMG) is a technique for evaluating and recording the electrical activity produced by skeletal muscles.

• An electromyography detects the electrical potential generated by muscle cells, when these cells are electrically or neurologically activated.

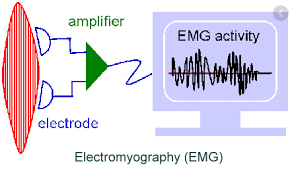
****

Fig 4.1 Electromyography signal generation

There are many applications for the use of EMG. EMG is used clinically for the diagnosis of neurological and neuromuscular problems.

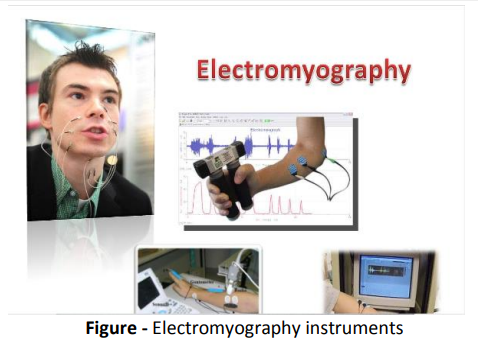


Fig 4.2 Electromyography Instruments

The block diagram clearly explains the implementation of silent sound technology through electromyography. EMG is performed using instrument called an electromyograph, to produce a record called an electromyogram. An electromyograph detects the electrical potential generated by muscle cells when these cells are electrically or neurologically activated.

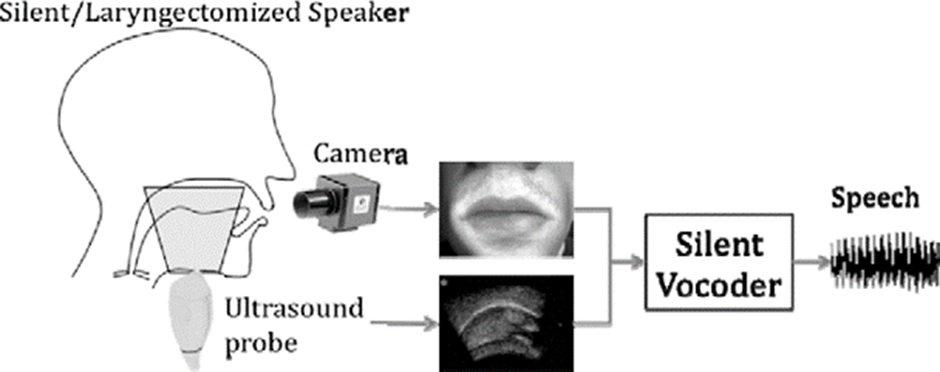


Fig 4.3 Block diagram of EMG

**WORKING**

There are two kinds of EMG in widespread use: surface EMG and intramuscular (needle and fine-wire) EMG. To perform intramuscular EMG, a needle electrode or a needle containing two fine-wire electrodes is inserted through the skin into the muscle tissue. A trained professional (such as a neurologist, physiatrist, or physical therapist) observes the electrical activity while inserting the electrode. The insertional activity provides valuable information about the state of the muscle and its innervating nerve. Normal muscles at rest make certain, normal electrical signals when the needle is inserted into them. Then the electrical activity when the muscle is at rest is studied. Abnormal spontaneous activity might indicate some nerve and/or muscle damage. Then the patient is asked to contract the muscle smoothly. The shape, size, and frequency of the resulting motor unit potentials are judged. Then the electrode is retracted a few millimetres, and again the activity is analyzed until at least 10–20 units have been collected. Each electrode track gives only a very local picture of the activity of the whole muscle. Because skeletal muscles differ in the inner structure, the electrode has to be placed at various locations to obtain an accurate study.

Intramuscular EMG may be considered too invasive or unnecessary in some cases. Instead, a surface electrode may be used to monitor the general picture of muscle activation, as opposed to the activity of only a few fibres as observed using an intramuscular EMG. This technique is used in a number of settings; for example, in the physiotherapy clinic, muscle activation is monitored using surface EMG and patients have an auditory or visual stimulus to help them know when they are activating the muscle (biofeedback).

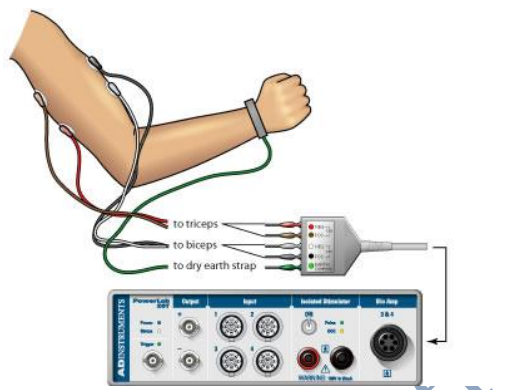
****

Fig 4.4 Interfacing with electromyographer and body

A motor unit is defined as one motor neuron and all of the muscle fibres it innervates. When a motor unit fires, the impulse (called an action potential) is carried down the motor neuron to the muscle. The area where the nerve contacts the muscle is called the neuromuscular junction, or the motor end plate. After the action potential is transmitted across the neuromuscular junction, an action potential is elicited in all of the innervated muscle fibres of that particular motor unit. The sum of all this electrical activity is known as a motor unit action potential (MUAP). This electrophysiologic activity from multiple motor units is the signal typically evaluated during an EMG. The composition of the motor unit, the number of muscle fibres per motor unit, the metabolic type of muscle fibres and many other factors affect the shape of the motor unit potentials in the myogram. Nerve conduction testing is also often done at the same time as an EMG to diagnose neurological diseases. Some patients can find the procedure somewhat painful, whereas others experience only a small amount of discomfort when the needle is inserted. The muscle or muscles being tested may be slightly sore for a day or two after the procedure.

**IMAGE PROCESSING**

The simplest form of digital image processing converts the digital data tape into a film image with minimal corrections and calibrations. Then large mainframe computers are employed for sophisticated interactive manipulation of the data. In the present context, overhead prospective are employed to analyse the picture. In electrical engineering and computer science, image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or, a set of characteristics or parameters related to the image. Most image processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it.

Analysis of remotely sensed data is done using various image processing techniques There two types of image processing: -

• Analog image processing

• Digital image processing

**ANALOG IMAGE PROCESSING**

Analog processing technique is applied to hard copy data such as photographs or printouts. It adopts certain elements of interpretation, such as primary element, spatial arrangement etc. With the combination of multi-concept of examining remotely sensed data it allows us to make a verdict not only as to what an object is but also its importance. Apart from these it also includes optical photogrammetric techniques allowing for precise measurement of the height, width, location, etc. of an object.

**DIGITAL IMAGE PROCESSING**

Digital image processing is been used in silent sound technology, is the processing of converting the image into image, video or audio but in Silent sound Technology (talking without talking) the output is audio with minimal corrections and calibration. The interface will contain ultrasound transducer, high resolution optical camera, lips reader and silent vocal. Ultra sound device which couples with high resolutions optical camera which will capture image of the lips and tongue movement. The image will be sent to lips reader and the lips reader compare the earlier spoken words with the present lips and tongue movement and the matched Image of lips and tongue will generate a visual speech signal.

To overcome the flaws and deficiencies in order to get the originality of the data, it needs to undergo several steps of processing.

1. Pre-processing

2. Display and enhancement

3. Information extraction

**CHAPTER 5**

**MERITS AND DEMERITS**

**MERITS**

* Very useful for those people who lost their voice and has been rendered mute due to accident.
* Very good technology for noise cancellation technique.
* Helps in making phone calls in noisy environment.
* Very useful for sharing confidential information like secret PIN number on phone at public place.
* Very useful for astronaut.

**DEMERITS**

* This technology works in many languages of user’s choice like English, French & German, etc. But, for the languages like Chinese is difficult because different tones can hold many different meanings.

**APPLICATIONS**

* It will help people who have lost their voice as a result of accident or cannot speak loudly again as result of old age.
* It can be used in military for communication of secrete or sensitive information.
* It is applicable if you want to make a call when you are in a conference meeting or library without disturbing the others.
* Speaker can speak his native language like German and listener can listen to it in his native language like English.
* It is applicable for those who want to make a call in nosily environment e.g., people working in train station, movie theatre, market etc.
* As we know in space there is no medium for sound to travel therefore this technology can be best utilized by astronauts.

**CHAPTER 6**

**FUTURE SCOPE**

Visual speech recognition gives way to a bright future to speech recognition technology from simple voice commands to memorandum dictated over the phone all this is fairly possible in noisy public places. Without having electrodes hanging all around your face, these electrodes will be incorporated into cell phones. Nano technology will be a mentionable step towards making the device handy.

In the future this technology can be incorporated into mobile phone or headset which headset would decipher the movement of the lips and jaws and received electrical impulse which will be convert into sound signal before transmitted.

**CHAPTER 7**

**CONCLUSION**

Thus, Visual Speech Recognition, one of the recent trends in the field of information technology implements” Talking Without Talking”. It will be one of the innovation and useful technology and in mere future this technology will be use in our day-to-day life. ‘Silent Sound’ technology aims to notice every movement of the lips and transform them into sounds, which could help people who lose voices to speak, and allow people to make silent calls without bothering others. Rather than making any sounds, your handset would decipher the movements your mouth makes by measuring muscle activity, then convert this into speech that the person on the other end of the call can hear. So, basically, it reads your lips.

**REFERENCES**

**[1]. Sulochana Devi, Siddhi Chokshi, Kritika Kotian, Juili Warwatkar,** “Visual Speech Recognition”, International Conference on Nascent Technologies in Engineering (ICNTE), IEEE 2021.

**[2]. Denis Ivanko, Dmitry Ryumin, Alexey Kashevnik, Alexandr Axyonov, Alexey Karpov,** “Visual Speech Recognition in a Driver Assistance System”, IEEE 2022.

**[3]. Siyuan Ma, Dantong Jin , Ming Zhang, Bixuan Zhang, You Wang, Guang Li, Meng Yang,** “Silent Speech Recognition Based on Surface Electromyography”, IEEE ,2019.

**[4]. Navin Kumar Mudaliar, Kavita Hegde, Anand Ramesh, Dr. Varsha Patil,** “Visual Speech Recognition: A Deep Learning Approach”, in International Conference on Communication and Electronics Systems (ICCES 2020), IEEE 2020.