BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

FACULTY OF ELECTRICAL AND COMPUTER ENGINEERING

MSc. in Computer Engineering

Amharic Language Speech (Lip Motion) Recognition Using Deep Learning

By:

Dagmawi Samuel

July, 2022
Bahir Dar, Ethiopia

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

FACULTY OF ELECTRICAL AND COMPUTER ENGINEERING

Amharic Language Speech (Lip Motion) Recognition Using Deep Learning

By:

Dagmawi Samuel

A thesis submitted
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Engineering

Advisor: Abrham Debasu (Ass. Professor)

July, 2022
Bahir Dar, Ethiopia

Bahir Dar Institute of Technology-Bahir Dar University

School Of Research and Graduate Studies

Faculty of Electrical and Computer Engineering

THESIS

Student:

Dagmawi Samuel		28/07/2022
Name	Signature	Date

The following graduate faculty members certify that this student has successfully presented the necessary written thesis proposal and oral presentation of this proposal for partial fulfillment of the thesis-option requirements for the Degree of Master of Science in Computer Engineering.

Approved:

Advisor:

Abrham Debasu (Ass. Prof.)		28/07/2022
Name	Signature	Date

External Examiner:

Dr. Beakal Gizachew (Ass. Prof)		August 28,2022
Name	Signature	Date

Internal Examiner:

Dr. Birhanu Hailu (Ass. Prof)		25/08/2022
Name	Signature	Date

Chair Person:

Mr. Molla Atanaw		
Name	Signature	Date

Faculty Dean:

Mr. Tewodros Gera (As. Prof)		
Name	Signature	Date

Acknowledgement

First of all we want to say thanks and appreciate Bahir Dar Institute of Technology-Bahir Dar University and information Technology for bring very willing to equip its students well theoretically and most of all practically. Secondly, we are very grateful to Abrham Debasu (Ass. Professor) who is our advisor at Bahir Dar University School of Electrical and Computer Engineering for his encouragement, excellent guidance; creative suggestions and critical comments have greatly contributed to the Industrial Internship work.

During our work, special cooperation and continual support has been given by Ethiopian Satellite Television (ESAT) Addis Ababa, Ethiopia main studio, and all who responded to our questionnaires, helped us in a great deal in our practical study. Finally, we thank our parents for their encouragement, individual and continual support.

Table of Contents

Declaration

This is to certify that the thesis entitled Amharic Language Speech (Lip Motion) Recognition using Deep Learning, submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Engineering under Faculty of Electrical and Computer Engineering, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during the course of this investigation have been duly acknowledged.

Dagmawi Samuel		28/07/2022
Name	Signature	Date

Abbreviations

AAM	Active Appearance Models
ANN	Artificial Neural Network
ASR	Audio Speech Recognition
AV-ASR	Audiovisual Automatic Speech Recognition
AVSR	Audiovisual Speech Recognition
BiLSTM	Bidirectional Long Short Term Memory
BiRNN	Bidirectional Recurrent Neural Network
CER	Character Error Rate
CHMM	Coupled Hidden Markov Model
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DNN	Deep Neural Network
DWT	Discrete Wavelet Transform
FFNN	Feed-forward Neural Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Units
HCI	Human Computer Interface
HMM	Hidden Markov Model
ISWYS	I See What You Say
LDA	Linear Discriminant Analysis
LSTM	Long Short Term Memory
MFCC	Mel-frequency Cepstral Coefficients
MSE	Mean Square Error
MSHMM	Multi-stream Hidden Markov Model
PLAVD	Phrase-level Amharic Visual Dataset
PLP	Perceptual Linear Prediction
RNN	Recurrent Neural Network
ROI	Region of Interest
SNR	Signal to Noise Ratio

STCNN	Spatiotemporal Convolutional Neural Network
SVM	Support Vector Machine
VASR	Visual Automated Speech Recognition
VSR	Visual Speech Recognition
WER	Word Error Rate
WRE	Word Recognition Error

List of Table

List of Figures

Abstract

Visual speech recognition, often known as lip reading, is a technique for interpreting a speaker's words by seeing his or her mouth movement. People are shown straining to understand what the speaker is saying in numerous settings. Video data that has been corrupted, such as sound (audio data) that has been intentionally or unintentionally deleted or distorted, video data captured by surveillance cameras (most security camera videos captured from afar have no audio data or the audio is unusable), hearing impaired people who can't hear the voice, and other situations try to understand the speech by looking at the speaker's lip movement. Because visual voice recognition is such a difficult task for a human to perform, it must be automated. Several scholars from around the world have conducted various studies on programmed visual speech recognition for various dialects. For the word and digit level, previous visual discourse recognition for Amharic is proposed. In addition, they take into account solely the video data from the front side. Using a deep learning computation, we presented phrase-level visual discourse recognition for the Amharic dialect in this study. Amharic is one of Ethiopia's most widely spoken dialects. A claim dataset comprising a few test expressions was gathered from a few Amharic language speakers. Because a video is made up of a series of sequential images, the video data will be encased in a pattern of picture outlines. The gathered video data is preprocessed: the video data is converted to outlines, and the outlines are then passed through a series of picture preprocessing forms. Finally, we put our photo data together and give it a name. The preprocessed data is then used to extract highlights, which is followed by categorization for preprocessing, we use OpenCV (Open Source Computer Vision), Matplotlib, Viola and Jones (to identify the face and the lip), and other python tools. We use a Convolutional Neural Network (CNN) to extract highlights. In the system 70% of the dataset were used for training and 30% of the datasets were used for testing. We extract 22 frames for a single phrase from a single video. For categorization, we used a Recurrent Neural Network (RNN), especially the Bidirectional Long Short Term Memory (BiLSTM) and we meet an accuracy of 92% for our study.

Keywords: Amharic; Lip-reading; Deep Learning, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN)

CHAPTER ONE: INTRODUCTION

1.1 Background

Speech is a way of communication through any communicable means that express thoughts and feeling through articulate sounds or visual speech which is transferring information by moving the body parts. In order to speak with one another, the communicating parties must understand the speech by hearing or visually understanding it. During this particular work, we'll be considering Visual Speech Recognition (VSR). We've investigated the task of speech recognition from various videos', and not from audio.

Lip-reading can be interpreted as a technique for understanding speech that involves visually representing the movements of the lips, face, and tongue. Lip-reading is still a really challenging task, even for the trained expert. Through the advancements of recent technological changes in computer vision, pattern recognition, and signal processing, there's a dynamic shift to automating the difficult task of lip reading.

Visual Speech Recognition (VSR) is critical in Audiovisual Speech Recognition (AVSR) like, security camera videos without voices, human-computer interaction, the hearing-impaired community, signing recognition, and translation systems, digital entertainment, and other areas. Visual speech recognition, also mentioned as automatic speech recognition, is used in applications that need an understanding of spoken words from visual signals while the speaker is speaking. There are numerous Visual Speech Recognition (VSR) systems available today, both which are combined with Audio Speech Recognition (ASR) and as standalone systems that are deployed starting from earlier researchers that proposed lip-reading systems traditionally [1] [2], [3] and it's advancement to using Deep Neural Networks (DNN) [4].

Improvements has been recorded on latest studies done on Visual Speech Recognition (VSR) and going further on to recognition accuracy. The foremost reason for this is often the utilization of Deep Neural Networks (DNN). The most object of this research is to rearrange an Amharic lip reading dataset and use Deep Neural Networks (DNN) to develop Visual Speech Recognition (VSR) for the Amharic language. Automatic Speech Recognition (ASR) could even be assigned as a technology that permits a computer to interpret words that a speaker speaks into a microphone or telephone and convert it to a transcription. Its

performance has yet to realize the extent required for speech to become a very pervasive interface. Indeed, even in “clean” acoustic environments, state-of-the-art Automatic Speech Recognition (ASR) system performance lags human auditory perception by up to an order of magnitude, whereas its lack of robustness to channel and environment noise continues to be a big difficulty.

Nowadays there is a drift to form communication and interaction between humans and their artificial partners making it easier and more natural. Speech recognition technology has reached a maximum performance and algorithm for building speech recognizers. However, the main problem of background noise and reverberations which are caused by the environment are still insurmountable. Therefore, inspecting other sources, apart from sound, for complementary information which could alleviate these problems, could also be a necessity [5]. It’s documented that both human speaking and perception are bimodal process in nature. Visual observation of the lips, teeth and tongue offers important information about the place of pronunciation articulation. A certain individual listener can use visual cues, like lip and tongue movements, to strengthen the extent of speech understanding. The tactic of using sight is typically mentioned as lip-reading which is to make sense of what someone is saying by watching the movement of his lips [6] . A clear Visual Speech Recognition (VSR) system refers to a system which utilizes the visual information of the movement of the speech articulators such as the lips, teeth and somehow tongue of the speaker. The advantages are that such a system isn't sensitive to ambient noise and alter in acoustic conditions, doesn't require the user to make a sound, and provides the user with a natural feel of speech and dexterity of the mouth.

The existence of speech command based systems are useful as a natural interface for users to interact and control computers. These types of systems provide more flexibility as compared to the traditional interfaces like keyboard and mouse. However, most of those systems are supported audio signals and are sensitive to signal strength, ambient noise and acoustic conditions [7]. To overcome this limitation, speech data that's orthogonal to the audio signals like visual speech information are often used. Audiovisual Speech Recognition (AVSR) system are those systems that combine the audio and visual modalities to identify utterances in a speech. There are two ways of approaching phonetics, one approach studies

the physiological mechanisms of speaking and this is often referred to as articulatory phonetics and the other form of approaching phonetics is known as acoustic phonetics this approach cares with measuring and analyzing the physical properties of the sound waves that is produced while we speak. Along with articulatory phonetics, organs of articulation are divided into movable articulators and stationary articulators. Movable articulator is that the articulator that does all or most of the moving during a speech gesture. The movable articulator is typically the lower lip, some part of the tongue and jaws. A stationary articulator is that the articulator that creates little or no movement during a speech gesture. Stationary articulators include the upper lip, the upper teeth, the varies parts of the upper surface of the mouth, and therefore the back wall of the pharynx [8], [9].

1.2 Motivation of the Study

As communication is now and then required: For the communicating parties to understand each other they must have to hear what the speaker is talking or know what he/she means. In any case, there are a few individuals who misplaced their hearing capabilities by diverse circumstances. If an individual is unable to tune in the voice of the speaker, he/she got to use the visual data of the speaker to get what the speaker is talking about. And also communication for people with difficulties is mandatory yet could have boundaries for their interaction with community that either caused by certain disease temporarily or those with permanent problems that caused difficulty to speak or much more equivalent problems. So lip-reading is an inevitable task to reconstruct the communication process and save the interaction with society.

In our day to day action, we have diverse activities that need the interaction between different individuals for the purpose of work or any other cause, even as a community there are various people that have their own roles such as writers who report something that happens some place or taxi drivers and taxi users who communicate each other, Bimodality is a typical feature of human speech. Both audible and visual components, such as lip synchronization and facial expressions, influence a person's perception of speech. Visual perspectives of communication can compensate for a misfortune in sound-related highlights of speech in noisy situations. Sound-related and visual discourse recognition in combination is more exact than either sound or visual speech recognition alone. Speech perception can be advanced by combining several data sources, such as sound and video views of dialogue [10].

And the greatest inspiration for us to work on lip-reading radiates from seeing this situation from diverse places and analyzing different inquiries about an automatic lip-reading through the use of video arrangements of the speaker's mouth which have pulled in lots of interests. Since automatic lip-reading under noisy environments is very effective in compensation for the decrease of speech recognition rate with an Audio Speech Recognition (ASR) system with bimodal based on audio-visual data is noted to being an important portion of the Human Computer Interface (HCI). So we have given more weighting esteem to visual data than to audio one for those having bad Signal to Noise Ratio (SNR) but, on the other side, more to sound information than to visual one for those voice signal having a clean Signal to Noise Ratio (SNR) [11]. Beneath noisy circumstances, this bimodal approach has been a great selection as and it shows superior recognition rate to Audio Speech Recognition (ASR) frameworks.

1.3 Statement of the Problem

Even if visual information records are getting progressively prevalent audio features are still the main contribution and play a more important role, than visual features but visual features will have a domain role in some cases exist like a difficulty to extract useful information from the audio. There are many applications during which it's necessary to recognize speech under extremely adverse acoustic environments.

A few video records are accessible without the audio records being synchronized in the video. This may happen for a number of reasons. It's a progressing step for the video that is captured from a remote position, like athletes inside the field, from the safety cameras without a sound recorder, the sound record could be mutilated intentionally or unintentionally, the sound and thus the video records are captured independently and there's no synchronization, like in a few motion pictures.

Lip-reading is also a necessary task for hearing impaired people. With the help of reading the movement of the lip they can understand what others are talking about. In addition to understanding lip movement the system can be integrated to sign language recognition and can be used to understand sign language more clearly and to improve the performance of sign language recognition and translation systems. Hearing-impaired people can only achieve a word-level accuracy of 17% [13] [14]. To overcome the problem of reading the movement of the

lip, it's better to automate the lip-reading task. There has been tons of research done on lip-reading for various languages like English, Indian, Arabic, and so on. However, there has been little or no work done on the Amharic language. [12] Proposed an audiovisual speech recognition system for the Amharic language. This work fills the gap of the above mentioned identified problems for the Amharic language through lip-reading system.

However, within the previous works, the lip-reading is on digits and phonemes that didn't work for longer variation on each sequence of images. Our proposed system goes to figure on the word and phrase level which will handle little longer variation of length on each sequence of images. To the present end, this research answers the following research questions:

- How to extract frontal face of the speakers?
- Which learning function of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) is suitable for visual speech recognition for the Amharic language?
- How to develop a phrase-level visual speech recognition model for the Amharic language?

1.4 Objectives of the Study

1.4.1 General Objective

The general objective of the study is to design a speech (lip motion) recognition model for the Amharic language.

1.4.2 Specific Objective

For the achievement of the general objective of our study, we used the following specific tasks

- Reviewing different literatures
- Collecting and preparing the dataset for training and testing our model
- Identify appropriate learning function of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for Amharic Visual Speech Recognition (VSR)
- Design visual speech recognition model for Amharic language
- Develop phrase-level visual speech recognition model for the Amharic language
- Testing and evaluating the model and analyze the result

1.5 Methodology

Experimental research is the type of research methodology that we used. We completed actions such as literature review, dataset preparation, tool and technique use, testing, and assessment as part of the whole study process.

Literature review: The numerous sorts of approaches, methods, and tools used in implementing the visual-based (lip-reading) speech recognition system will be examined in an exhaustive literature study. Reviewing the existing literature will aid us in identifying effective image processing, feature extraction, and classification methods for our research. Literature on the Amharic language will also be examined in order to gain a better understanding of the language's phonetic and viseme qualities.

Dataset: In a controlled environment, the data is captured in video format. The video files are collected from the front view, to make our dataset more balanced (the films may not always be accessible with a front view of the speaker). The participants (speakers) on the data preparation team are of various genders and ages. Since mouth localization has been conducted on images, a person with a mustache makes the work more complex. As a result, speakers with and without mustaches (beards) have been considered.

Techniques and Tools: The majority of lip-reading studies focus on five main tasks: video framing, face localization, lip localization, feature extraction, and classification. According to the survey [15] [16], the majority of classic lip reading approaches used Active Appearance Models (AAM) to extract features and Hidden Markov Models (HMM) to classify them. The usage of a Deep Neural Network (DNN) has been shown to improve the performance of lip-reading algorithms in recent studies. The majority of research following Deep Neural Network (DNN) have achieved good recognition accuracy. With this data, we were able to do end-to-end Deep Neural Network (DNN) learning. For both feature extraction and classification algorithms, we evaluated Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN). We approach Ethiopian Satellite Television (ESAT) to collect a recorded news video. The most recent version of Python was utilized to create the development environment. The OpenCV library is an open-source image processing and computer vision toolkit that can be readily integrated with Python and used to conduct image processing and computer vision operations. As a result, we'll use the OpenCV library for video framing and image processing, as well as prototyping.

Evaluation: We used a variety of evaluation metrics to assess our model, including confusion matrices, recognition accuracy and F1 score, precision, and recall.

1.6 Scope and Limitation

The goal of our suggested thesis is to create an Amharic-language model for automatic Visual Speech Recognition (VSR) or lip motion reading. For the Amharic language, our research focuses on lip movement identification at the word and phrase level. The following tasks were not included in our proposed research:

- It is only for the Amharic language; it didn't recognize words or phrases for other languages.
- The video files captured other than front views are not included.
- It didn't recognize the end-to-end sentence level.

1.7 Significance of the Study

The construction of a phrase-level lip reading model that analyzes video shot from one point of view as a phrase level is the study's significant contribution to the scientific community, particularly in the field of lip-reading. Other contributions to this research include hearing-impaired people, Audiovisual Speech Recognition (AVSR), security, and so on.

Audiovisual Speech Recognition (AVSR): Speech recognition systems can benefit from Visual Speech Recognition (VSR). In a noisy setting, Audiovisual Speech Recognition (AVSR) can be used to recognize speech. Visual Speech Recognition (VSR) is a technique for improving the accuracy of automatic speech recognition systems.

Hearing-impaired people: Hearing-impaired people are unable to hear the speaker's sound. To communicate with one other, these persons must either read the speaker's lip movements or utilize sign language. Reading lip movements, on the other hand, is an extremely challenging assignment for a person. As a result, automatic lip-reading is a potential research project for the hearing handicapped.

Security: Lip-reading is important in security situations, such as when a surveillance camera is used and it is difficult to hear what the person on the camera is saying because the video is collected from a far distance.

1.8 Application of Results

Lip-reading can be viewed as a stand-alone procedure or as a supplement to speech recognition. The applications for lip reading as a stand-alone application are diverse: multimedia phones for the hearing impaired, mobile phone model interface spaces (e.g., phone models that use lip-reading are already being designed at the time of this writing), person identification, speech recovery from deteriorated or mute movie clips, and perhaps the most promoted application, security by video surveillance.

1.9 Organization of the Thesis

The remaining thesis are arranged as follows:

The second chapter tackles video and image processing, as well as the Amharic language, feature extraction, video categorization, and other relevant topics. The third chapter outlines our proposed work's approach and design, as well as the general system structure, materials employed, and experiment. The experimental results, as well as their analysis, appraisal, and discussion, are presented in Chapter Four. Finally the conclusion part and recommendation for future work takes place in Chapter Five.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter includes a survey of relevant literature as well as an assessment of the present state of the field. This chapter reviews the literature on computer vision, image preprocessing (segmentation, localization, and other techniques), video preprocessing, feature extraction, classification, and other topics. These, as well as all other possibilities linked to the title we mentioned, are being investigated on a daily basis. We will explore various literature that is being researched by various experts in order to discover all of the benefits of these theories. The performance of knowledge of computer vision, machine learning, and other preprocessing phases will improve as the volume of literature grows. As a result, this chapter will give a survey of the various literature on Visual Speech Recognition (VSR).

2.2 Computer Vision

Both C. M. Dana H and C. E. Hang T., Vandoni on Computer vision topics that study how computers can interpret digital data at a high level (image or video) [17] . The gathering, processing, analysis, and comprehension of digital data, as well as the extraction of high-dimensional data from the actual world to produce numerical or symbolic information, are all covered by computer vision. Digital data is increasing in both amount and size these days. Medical photographs, agricultural studies, security cameras, traffic police law enforcement, scientific research, satellite images, videos, and other sources of digital information can be collected. These digital files from various sources may require additional processing in order to improve their quality or make the data more intelligible. As a result, computer vision plays a big role in this kind of thing. The most closely related terms to computer vision are image processing, image analysis, and machine vision. This encompasses a diverse set of methodologies and applications, some of which overlap. This shows that the underlying principles used and developed in different fields are similar, implying that there is only one area with multiple names. Computer graphics generate image data from 3D models. Computer vision routinely uses image data to create 3D models [18]. Computer vision activities will be carried out on picture and video files. When doing computer vision tasks on video data, they must be transformed into continuous image frames. Different image preprocessing operations will be done on the image file after the video file has been

converted into picture frames. Then, characteristics will be extracted from the image files, and each image's transition will be evaluated to anticipate the video file's content.

Video processing: An approach that uses a series of fast moving image sequences is known as video processing. When we talk about video processing, we're usually talking about image processing. As a result, in order to classify photos, anticipate images, increase image quality, or process images further, image processing needs follow specific protocols.

2.3 Lip-reading

Individuals mostly communicate through speech. The majority of the time, audio communications are used to communicate between people. In most cases, simply conversing via audio signal is adequate. Visual speech understanding may be the sole option if the audio file is lost or the person who is supposed to hear the sound is unable to do so.

Lip reading is not a new human invention; it has been used since at least 1500 AD, and probably before [19]. Pietro Ponce, a Spanish Benedictine monk who died in 1588 AD, was the first successful lip reading teacher [14]. Despite the fact that only half of the ten utterances are visible, the reader must guess or estimate the words he or she has missed. According to [19] the Jena method "trains the eye and exercises the mind." Regardless of the lip-reading techniques used, they all rely on the ability of the reader to identify lip movement.

The act of interpreting or reading speech by a machine utilizing visual information gathered from the image of the mouth, lip, face, chin, and any other areas linked to the face is known as automatic lip-reading or visual speech recognition [14]. Because of the relevance of the issue, academics are very interested in automatic lip-reading or visual speech recognition. [14]: Automatic lip reading can:

- **Aid for voice recognition** - While auditory noise or unwelcome sound from behind can disrupt speech recognition systems, visual speech recognition (VSR) is unaffected since it relies on the visual signal rather than the audio signal. A comprehensive system with sign language recognition is utilized to aid hearing impaired people.
- Be used in human-computer interaction.
- Assist in the recognition of different face expressions.
- Be used in talking heads systems, it can be used to synthesis visual speech.

- Be used in surveillance systems for security as well as to identify and authenticate the speaker.

Preprocessing: There will be some preprocessing duties after the video has been collected. First and foremost, the video's irrelevant segments have been deleted. The video file must then be converted into a frame sequence. Following the framing of the video files, the picture files should be processed to locate the speaker's face region, a process known as face localization. Following face localization, the next stage will be to locate the speaker's Region of Interest (ROI), which is the lip section of the speaker's face.

Face Localization: Because faces can be presented at varying scales, orientations, locations, and attitudes, face detection in static images or videos is an essential yet difficult problem in computer vision [20] , All facial tasks, such as face localization, facial feature detection, face identification, face verification, and facial emotion recognition, need face detection [21]. There are 4 (four) major sorts of techniques to face localization. These types of methods used in face localization are: knowledge-based, feature-based, appearance-based, and template-based.

Knowledge-based Approach: A face is represented by a set of human coding rules in this technique, which is a top-down approach. In these approach the regulations state that "the average intensity values of the center and higher portions are significantly different," "the difference between the average intensity values of the center and higher parts is significant," and so on are the regulations [22].

The following are some of the benefits of this strategy [22]:

- It's straight forward to come up with simple principles to define facial characteristics and their interactions.
- In a static background, it is a better strategy for face localization.
- Facial candidates are determined using coded criteria once face features are retrieved from an input image.

The drawbacks are as follows [22]:

- It's difficult to convert human knowledge into rules that are accurate: particular rules can't recognize faces, and generic rules might produce a lot of false positives.

- This method is difficult to adapt to detect face components in varied positions since all possible reasons cannot be identified.

Feature-based Methods: This is a bottom-up technique, in which face features are detected first. Edge, form, texture, intensity, color, and other pixel attributes are used in this method. The goal of this strategy is to find invariant features. The advantage of this strategy is that features are unaffected by changes in position and orientation. The downsides of this method include difficulties detecting features in a complicated background and trouble locating facial features owing to changes in noise or illumination [22].

Template-based Methods: This approach recognizes a face by comparing it to a template. The template can be predefined, for example, based on edges or areas, or changeable, for example, based on facial shapes. The templates need to be saved in a database. To locate facial parts, the correlation approach is employed. The average area intensity values were employed instead of absolute pixel values in this method, which used relative pair-wise ratios of the brightness of the facial region [22].

- This approach has the advantage of being quite simple to utilize. Templates must be initialized near the face pictures, and enumerating templates for multiple poses is complex (similar to knowledge-based techniques).

Appearance-based Methods: A classifier is trained utilizing some positive and negative examples of faces in this method [22]. Representation (holistic and block-based approaches), pre-processing, train a classifier, search Strategy, post-processing, view-based (facial features are detected without any knowledge of geometry), Neural Networks (used to detect front and non-front parts of the face with variation in poses), Principal Component Analysis Support Vector Machine, Hidden Markov Model, and others are among the techniques used in this method.

2.3.1 Finding the Region of Interest (ROI)

The method for detecting the Region of Interest (ROI) with human lips was developed using the face detection approach. Because the Region of Interest (ROI) in the lips is smaller than the original picture of the human face, methods for detecting the lips can take longer. In the first video images in the video stream, the human speaker is presumed to have spoken lips squeezed together. The Region of Interest (ROI) around the lips must be detected sequence of images in each frame

of the movie. Fitting a variety of color models to the image is usually the first step, followed by face detection and extraction of the Region of Interest (ROI) surrounding the lips [23]. Lip segmentation was previously performed by wrapping lipstick around the speaker's lips. Even in a confined context, the application of lipstick enables picture data recognition of the lip area straight forward and precise, but this method is unsuccessful because it does not bring comfort to consumers [24]. As a result, rather than painting such marks on the speakers' lips, it is preferable to examine extracting features.

2.3.2 Feature Extraction:

The process of lowering the dimensionality of an initial collection of raw input data into more manageable groups of patterns for processing is known as the extraction feature. A vast number of variables necessitates a lot of computer resources to process these massive data sets. Feature extraction refers to methods for selecting and/or combining variables into features in order to limit the amount of data that must be processed while still properly and comprehensively defining the original data set [25]. It's a difficult and interesting endeavor to extract characteristics for visual speech recognition. When it comes to creating visual features, the large amount of data in video sequences is a major challenge, and it's a problem that needs to be solved. A feature vector of between 10 and 100 items must be collected from each video frame, which contains hundreds of pixels [26]. The Hidden Markov Model (HMM) was created in the mid-1980s to make visual speech detection easier. Since then, numerous approaches to visual speech recognition have been developed. The Active Shape Model (ASM) [27], the Active Appearance Model (AAM) [28], and Principal Component Analysis are used in traditional lip-reading systems for feature extraction (PCA). Since the introduction of deep learning, extracting features has become a lot easier.

Classification: The process of determining which class a given piece of data belongs in is known as classification. Labels/ targets or categories are occasionally used to identify classes. The process of comparing and estimating a mapping function from input variables to output variables is known as classification predictive modeling. Different researchers utilize different algorithms for visual speech recognition.

2.4 Visemes

In the visual realm, a viseme is a mouth shape or appearance, or a series of mouth dynamics required to form a phoneme. In other terms, a viseme is the visual representation of a phoneme. As a result, phonemes are the most basic units of speech recognition, and a viseme is the smallest visual feature that describes a phoneme or a set of phonemes.

Although visemes can be used to detect spoken words in theory, they only cover a small part of the word manifold [23]. Visemes are too small to adequately define the total world information since transitions between visemes are not employed in the viseme-based speech representation. When using visemes in visual speech recognition systems, there may be complications, such as:

- For the visemes and phonemes classes, there is no absolute (standard) number. Some research discovered ten, thirteen, fourteen, or sixteen visemes, as well as 44 or 53 phonemes [29]. For visemes, there is no globally agreed norm. Every researcher has his or her own set of criteria.
- Even if the number of visemes in two trials is the same, having the same visemic set in both cases, a visemic set is not required [29].
- Some phonemes have the same viseme, which means that the viseme is shared by multiple phonemes [29].
- Some visemes share a phoneme, which means that one phoneme can be mapped to multiple visemes depending on the following phoneme. Consider the phoneme n in the words "banana" and "Nottingham," try to pronounce both words, and use a mirror to observe your mouth. You'll observe that when articulating n in "banana," the mouth opened wider because it is affected by the next 28 phonemes, whereas when articulating in "Nottingham," the mouth shape was more circular because it is affected by the next phoneme "o." [29].
- The visual depiction of some phonemes, particularly those articulated from the inside of the mouth, such as the glottal, is inadequate (h) [29].
- Viseme representation is provided by the fact that a major amount of the word manifold is not used in the identification process (i.e. transitions between visemes). This method is insufficient since defining each Visemes feature space with greater regions would

necessitate more instances of the same viseme derived from various terms, resulting in significant feature space overlap when describing unique visemes. [23].

- The downside of viseme-based representation is that in video sequences that graphically reflect spoken words, some visemes may be significantly distorted or even disappear.

These restrictions suggest that mapping the viseme to lip movements is challenging and that the information may be corrupted as a result. When the recognition process is continuous (rather than static), recognition might become increasingly challenging, hence the viseme technique is not appropriate for continuous lip motion recognition.

2.4.1 Amharic Language

Ethiopians speak about 80 different languages. However, Amharic is the most widely spoken and dominating language [30]. Amharic (also known as Abyssinian, Amarinya, Amarigna, and Ethiopian language) is Ethiopia's official language. Since the 13th century, it has been the language of the court and the majority of the people in Highland Ethiopia. Ethiopia's official language is Amharic, which is spoken in government, courts, and on all official documents [31]. Amharic is written in a script derived from the Ge'ez alphabet. Each consonant vowel combination has seven forms or modifications, totaling 33 basic letters. The language is written from left to right, unlike North Semitic languages like Arabic, Hebrew, and Syrian. Amharic uses a script that is based on the Ge'ez alphabet. ሀ, ለ, ሐ, መ, ሰ, ረ, ሸ are examples of the Amharic language symbols and consists of 33 fundamental letters, with seven forms or variants for each consonant vowel combination. Take the initial letter of the Amharic alphabet, which has seven variations (ሀ, ሁ, ሂ, ሃ, ሄ, ሆ). This means that the Amharic language has a total of 238 different symbols. There are forty (40) special characters to characterize labialization in addition to these primary symbols, such as ቋ, ኗ, ሟ, ሯ, ሷ, ሯ etc. Unlike other North Semitic languages like Arabic, Hebrew, and Syrian, the language is written from left to right. As a Semitic language, Amharic is distantly connected to Hebrew and Arabic; as a result, Amharic is probably as near to Spanish and Portuguese as Spanish and Portuguese are to English [32].

2.4.2 Amharic Visemes

Like any other language, Amharic has its own unique characteristics. Amharic, for example, has a unique collection of speech sounds not found in other languages such as English. These are glottalized plosives with a harsh click-like quality, similar to ጵ, ጥ, ጭ, ቅ, ጽ [33].

There are 38 phonemes, seven vowels, and thirty-one consonants in the Amharic language's overall sound inventory. Stops, fricatives, nasals, liquids, and semivowels are among the Amharic consonants [33].

Speech reading, also known as lip-reading or visual speech recognition (VSR), is a method of understanding speech that relies on visual clues gathered from photographs of the mouth, lips, chin, and other relevant parts of the face. Speech reading is essentially a type of audiovisual speech recognition in which the visual modality takes precedence over the acoustic. By observing the configuration and motion of the speaker's visible articulators, a person proficient in speech reading can infer the meaning of spoken sentences. Although it is commonly referred to as lip-reading, speech information is not solely derived from labial configurations because the tongue and tooth position also provide information. However, it is widely agreed that the mouth Region of Interest (ROI) provides the majority of information about visual speech [34] .

In addition to speech reading, humans may utilize their vision to aid in aural communication in a variety of ways. When the listener has problems understanding the acoustic speech, visible speech gives an additional information source. Listeners may also have difficulty understanding acoustic speech in instances where they are unfamiliar with the speaker, for as when listening to a foreign language or a speaker with an accent. When confronted with noisy situations, visual information can significantly improve human recognition. Visual speech's complementary and supplemental character can be exploited in speech processing applications such as automatic speech detection, augmentation, and coding in acoustically loud environments. [34]**.**

2.5 Machine Learning

The study of computer algorithms that improve themselves automatically over time is known as machine learning. Datamining methods that uncover general patterns in large data sets to information filtering systems that learn users' preferences automatically [35] and this are some of the applications. It's considered a part of artificial intelligence. Machine learning algorithms build

a model out of sample data called "training data" in order to make predictions or choices without being explicitly taught to do so [36]. When constructing standard algorithms to perform the essential tasks would be difficult or impossible, machine learning algorithms are used in a wide range of applications, including email filtering, medicine, speech recognition, and computer vision.

Machine learning is linked to computational statistics, which focuses on using computers to make predictions; however, not all machine learning is statistical learning. Mathematical optimization research contributes to the science of machine learning by providing tools, theory, and application fields. Data mining is a related field of study that focuses on exploratory data analysis using unsupervised learning [37].When applied to commercial concerns, machine learning is frequently referred to as predictive analytics.

Machine learning is the process through which computers learn to perform tasks without being explicitly instructed. It entails computers learning from data provided in order to do specific jobs. It is possible to develop algorithms that direct the machine on how to complete all steps required to solve the problem at hand for fundamental jobs committed to computers; no learning is required on the computer's part. It may be difficult for a human to manually create the algorithms required for increasingly complex jobs. In actuality, assisting the computer in constructing its own algorithm rather than having human programmers explain each essential step may be more advantageous.

Deep Learning: Deep learning (deep structured learning) is a machine learning technique that employs artificial neural networks and representation learning [38]. The information processing and distributed communication nodes of biological systems inspired Artificial Neural Networks (ANNs). In a variety of ways, Artificial Neural Networks (ANNs) differ from biological brains. The organic brain of most living animals is dynamic (plastic) and analogue, whereas neural networks are static and symbolic. In deep learning, the term "deep" refers to the network's use of multiple layers. A linear perceptron cannot be a universal classifier, according to early research, but a network with a non-polynomial activation function and one unbounded width hidden layer can. Deep learning is a modern variant that uses an infinite number of bounded-size layers to enable for practical application and efficient implementation while maintaining theoretical universality under moderate conditions.

2.6 Related Works

B. B. and D. B. . Eric Petajan introducing the concept of automatic lip-reading or Visual Speech Recognition (VSR) in 1988, [39]. Since then, a great deal of research has been conducted in several languages around the world using various methodologies. Visual Speech Recognition (VSR) or automatic lip-reading, as various researchers have stated, is one of the areas that has many application areas, such as automatic sign language recognition and translation systems, Audiovisual Speech Recognition (AVSR) systems, security cameras (surveillance), sporting videos, and many other areas. Researchers are very interested in automatic lip-reading or visual speech recognition because of its many applications. Since the commencement of the research, there has been research in the field. It is impossible to describe every study on automatic lip-reading here; nevertheless, some of the studies can be discussed to acquire a general grasp of the subject's evolution.

They developed a better automatic lip-reading system to improve speech recognition. This is an updated version of their previous 1986 work. The earlier voice recognition technology worked effectively in both noisy and calm contexts with tiny vocabulary. With a vocabulary of more than 100 words, previous research has not been able to achieve accurate acoustic speech recognition in noise. This paper presents the results of 19 Visual Speech Recognition (VSR) tests performed on a variety of speakers in ideal conditions. Lip reading is only second to the auditory system as a source of speech information, according to the writers of this article. It supports the acoustic speech signal but is less variable; the acoustic signal is so dependent on lip, teeth, and tongue position that lip-reading alone can provide important phonetic information. The purpose of this work was to locate and then follow the nostrils from frame to frame in order to recognize the mouth region in a full-face image. By region matching the values for each nostril's region parameter against a nostril template, the nostrils were discovered. The distance between the nostrils was also utilized to validate the position of the nostril regions, both horizontally and vertically. Using a mouse, a semi-automated nostril template was created for each speaker by manually defining a bounding area around them in a single facial shot. The entire face had to be scrutinized to detect the nostrils in the first frame of an utterance, but in following frames, the search was reduced to a little rectangular nostril window. The panes in each frame were altered to center the nostrils once the nostrils were located. The nostril window had to be large enough to contain the nostrils even if the

frame to frame movement was as large as feasible since the position of the nostril window in one frame was used as the position of the nostril search window in the next. The mouth areas were encased in a rectangular window that was manually adjusted in relation to the nose window. Because the nostrils and mouth are skeletally inflexible in relation to one another, each speaker's windows for the nostrils and mouth can be specified only once. Inside the mouth window, any dark patches were assumed to be mouth regions. Because it decreased residual modifications caused by global movements of the head and body that were not limited by head mounting the camera, nostril tracking was efficient, robust, and essential for successful automated lip-reading. When collecting a picture sequence during an utterance, the contour coder continually saved data in a 200 video frame circular buffer. After that, area codes were created from the contour-coded visual speech templates. Finally, the binary picture sequences of the mouth were obtained using nostril tracking and recorded with the area of the mouth areas in preparation for vector quantization. As a representative sample of mouth pictures for a specific speaker, two examples of each uttered alphabetic letter were employed. The mouth photographs were used in an iterative clustering process with a preset mouth image distance threshold to separate the groups. After then, the number of clusters was compared to 255. The distance was then calculated using a sequential approximation technique, yielding a cluster count that was close to, but not greater than, 255. The clustering method was restarted with the adjusted distance threshold for each iteration of the succeeding approximation. As a result, after around 255 clusters were created, in order to vector quantize visual speech templates, a representation from each cluster was added into the mouth picture codebook (mouth image sequences). The binary mouth image distance measure D12 is a function of three factors, as shown in the equation below: the area of the entire dark region in the two images A1, A2, and the Hamming distance H12 between images, where the Hamming distance is the area of the exclusive OR of the panels in the two binary images.

(A1 +A2) D 12 =255H12/ 1

The Hamming distance alone is insufficient to distinguish images with small mouth apertures that are linguistically distant. The value of D12 has been standardized to 0=D12=255. D12 is set to zero when Al=A2=0 is detected. If Al and A2 are both zero, D12 equals 255. Because the closed mouth or blank frame occurs only during bilabial plosives or potentially throughout at least some silences, it should be placed in its own blank frame cluster. There is also a linguistic distinction

between complete mouth closure and modest mouth openings. An inter cluster distance database is produced once the mouth image codebook is created, allowing for very fast vector quantized template matching. The distance between the two clusters is the average distance between every frame in one cluster and every frame in the other cluster. By comparing each image of each template to the codebook and substituting the mouth photos with the index or vector of the closest picture in the codebook, the visual speech templates may now be vector quantized.

The letter pairs A-K, B-P, C-Z, D-T, S-X, and Q-U are visually indistinguishable, as they indicated in their results section. The exam vocabulary was picked from the spoken alphabet and the digits zero through nine. When compared to auditory recognition alone, the results of the integrated recognition system imply a reduced constraint on attainable performance improvements. Finally, the authors conclude that vector quantization of the visual speech templates to one of only 256 mouth images generated adequate performance and should make the lip-reading system more practical to execute in real time.

I See What You Say (ISWYS) is an Arabic lip reading system [2]. This is a research-based Arabic-language voice recognition system that converts lip gestures into understandable text. It's done by studying a video of lips movements that resemble speech and then using video analysis and motion estimation to turn them into legible characters. By removing successive frames, their system separates the video into n frames and generates n-1 picture frames. The purpose of this paper is to create a lip-reading algorithm that recognizes Arabic alphabets exclusively through lip gestures. This can be accomplished by using a camera to record lip movements and then converting them into text. The system is separated into two subsystems: the first, which administrators use, is the training subsystem (trainers). The recognition subsystem, which is used by all users, is the other. The method used in the training sub-system is based on a motion estimation technique, and it was used to extract various features from the video, such as the first, second, and third seconds. The error function was calculated and the result was displayed for inspection.

The ISWYS was created to aid hearing-impaired people in communicating. This system focuses on a single mode of communication: converting speech from a hearing person into readable text for a hearing-impaired person. This is performed by tracking and analyzing the movements of the lips using video analysis and motion estimation techniques to derive the matched letter. A training area for administrators and a recognition section for hearing-impaired users make up the system.

The administrator is in charge of system training. A database of movies for each viseme is created throughout the training phase, which the hearing-impaired user can later use in the recognition process.

Motion estimation, a technique used to evaluate MPEG movies, is used by the ISWYS system. It's also known as a pixel displacement method since it measures the difference between two consecutive frames to track pixel movement. Direct motion can be estimated using a variety of methods, including optical flow and block motion estimation. Optical flow or optic flow is the velocity vector field of apparent motion of brightness patterns in a sequence of photographs. It specifies how much each picture pixel moves between successive photos. In order to calculate the pixel motion estimate, a two-frame differential approach is used. The most fundamental method of motion estimate is block-based motion estimation. Splitting the frame into pixel parts is how this method works. After that, the motion estimate for the current frame block is done by finding a matching block in a reference frame. The difference between these blocks is used to determine the displacement. The ISWYS system's algorithm is divided into two phases: training and recognition. In both segments, the process for acquiring video characteristics is the same. The difference is in how the system uses the data. During the training phase, the algorithm saves all of the attributes of a video shot by the trainer (admin) and uploads it to the ISWYS database. The algorithm records a video for the user during the recognition phase and displays the result on the screen.

Automatic lip-reading system for Dutch [3]. This study goes into great depth regarding the work being done at Delft University of Technology to develop a reliable automatic lip-reading system. They have described the construction and characteristics of the data corpus for the Dutch language that they have created for this study. They took into account a wide range of visual data, including motion description based on optical flow, form description based on key point detection and a statistical technique, and appearance description aspects. This study discusses the use of Active Appearance Models to recognize landmarks on the speaker's face for lip reading. The Hidden Markov Models technique of the HTK Toolkit is used to make the inference.

Each video frame is utilized to extract the position of specific points on the face using Active Appearance Models (AAMs). A statistical model of shape and texture variation is created using the Active Appearance Models (AAM) approach. A training set of some example shapes is used

to calculate the average shape. The sample shapes are aligned using the Generalized Procrustes Analysis. Each face sample's control points are then bent to match those in the meaning shape. The search starts with the mean model and iteratively adjusts the model parameters within the learned range, reducing the difference in appearance between the original picture and the image synthesized using the new model. The number of parameters required is calculated using PCA. A good initial assumption, as with the majority of search strategies, helps to speed up the process. It was discovered that using a face (mouth detection) tracking algorithm as a pre-processing step considerably accelerated the search for shape parameters during Active Appearance Models (AAM) based processing. This improvement allowed for a real-time implementation of the method. The Active Appearance Models (AAM) employs shape information extracted from a face image to calculate a set of relevant parameters that describe the look of facial characteristics.

Data parameterization is determined by the aspect of the data that has to be represented; for speech, it should characterize the curvature of the mouth as exactly as possible and capture the transformations that the mouth undergoes. Different image processing algorithms allow different parts of the data to be captured. They employed the Hidden Markov Model (HMM) Toolkit developed at Cambridge University for real recognition after extracting visual data such as geometric features, LGE features, and optical flow features. Visemes were chosen as the recognition units. In the visual world, visemes are the phonemes that correlate to meaning units of speech. Most researchers employ a set of 40 (forty) phonemes for the Dutch language, according to the authors.

They decided to start with a speaker's dependent method in their result section due to the significant amount of labor required to process all of the data. They had previously analyzed over 2 million frames in this example, from which features had to be retrieved. They were able to reduce the processing time for Active Appearance Models (AAM) and LGE to almost real-time by incorporating a mouth detection/tracking device. Optical flow, on the other hand, necessitates a large number of resources. Researchers looked at a variety of activities ranging in difficulty, including connected digits, connected letters, connected words, and random sentences. The latter is really equivalent with continuous lip reading, which is exactly what they hoped to do. They've also taken into account tri-viseme (equivalent to tri-phone) formulations with an inner word context where possible. However, due to the corpus's small size, many tri-visemes went unnoticed.

Finally, they stated that this was the result of their work on all of the aforementioned phoneme classes.

LIPNET: sentence-level Lip reading [40]. This is an end-to-end trained model, according to the article, that maps a variable duration series of video frames to text using spatiotemporal convolutions, Long Short Term Memory (LSTM) recurrent network, and the connectionist temporal classification loss. This is the first model of lip reading that works at the sentence level, according to the authors. To simultaneously train spatiotemporal visual data and a sequence model, they used a single end-to-end speaker-independent deep model. LipNet is a lip-reading neural network that transforms variable-length video frame sequences to text sequences. One of LipNet's basic elements is spatial temporal convolutions.

Convolutional Neural Network (CNNs), which use layered convolutions to operate spatially over an image, have played a key role in improving performance in computer vision applications that require an image as input, such as object recognition. A basic 2D convolution layer (without a bias and with unit stride) computes from C channels to C' channels.

$$[\mathrm{conv}(\mathbf{x}, \mathbf{w})]_{c'ij} = \sum_{c=1}^{C} \sum_{i'=1}^{k_w} \sum_{j'=1}^{k_h} w_{c'ci'j'} x_{c,i+i',j+j'},$$

For input x and weights w RC' C kw kh [40], where Icij=0 for I j out of bound By convolving over time and space, Spatiotemporal Convolutional Neural Networks (STCNNs) can handle video data. Long Short Term Memory (LSTM) is also included in LipNet. This is a Recurrent Neural Network (RNN) that adds cells and gates for propagating information over successive time-steps and learning to control the flow of information to prior Recurrent Neural Network (RNNs). They utilized the conventional Long Short Term Memory (LSTM) formulation with forget gates in this study:

$$\begin{aligned}
[\tilde{\mathbf{i}}_t, \tilde{\mathbf{f}}_t, \tilde{\mathbf{o}}_t, \tilde{\mathbf{g}}_t]^T &= \mathbf{W}_x \mathbf{z}_t + \mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{b} \\
\mathbf{i}_t = \mathrm{sigm}(\tilde{\mathbf{i}}_t) \quad \mathbf{f}_t &= \mathrm{sigm}(\tilde{\mathbf{f}}_t) \\
\mathbf{o}_t = \mathrm{sigm}(\tilde{\mathbf{o}}_t) \quad \mathbf{g}_t &= \tanh(\tilde{\mathbf{g}}_t) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t),
\end{aligned}$$

The input sequence to the LSTM is z:=z1,...,zT, O represents element-wise multiplication, and sigma(r)=1/(1+exp(-r)). The bidirectional Long Short Term Memory (LSTM was used by them Bidirectional Long Short Term Memory (BiLSTM).

LipNet's third and final building block is Connectionist Temporal Classification (CTC). The Connectionist Temporal Classification (CTC) loss is commonly used in current voice recognition because it eliminates the need for training data that aligns inputs to intended outputs. Connectionist Temporal Classification (CTC) calculates the probability of a sequence by marginalizing all sequences specified as equivalent to it, given a model that produces a series of discrete distributions over the token classes (vocabulary) augmented with a special "blank" token. This avoids the requirement for alignments when working with variable-length sequences.

The authors calculated the Word Error Rate (WER) and Character Error Rate (CER), which are common indicators of ASR model performance, to assess LipNet and the baselines' accuracy. LipNet's estimated maximum probability forecasts are obtained using Connectionist Temporal Classification (CTC) beam search. The minimum number of word (or character) insertions, substitutions, and deletions required to convert the prediction to the ground truth, divided by the number of words (or characters) in the ground truth, is known as Word Error Rate (WER) and Character Error Rate (CER), Word Error Rate (WER) is often equal to a classification error when the predicted sentence and the ground truth have the same number of words, which is especially true in our circumstance because nearly all mistakes are substitution errors. Individually, the three hearing-impaired adults scored 57.3, 50.4 %, and 35.5 % Word Error Rate (WER), for a total of 47.7 % Word Error Rate (WER). The baseline Long Short Term Memory (LSTM) performs marginally better than in hearing-impaired people, with a Word Error Rate (WER) of 31.4 %. When compared to hearing-impaired people, Baseline-2D and LipNet achieve roughly 4.3 times and 7.2 times reduced Word Error Rate (WER), respectively. Baseline-2D has a Word Error Rate (WER) of 10.9 %, while LipNet has a Word Error Rate (WER) of 6.6 %, highlighting the necessity of integrating Spatiotemporal Convolutional Neural Network (STCNNs) with Long Short Term Memory (LSTMs). This difference in performance supports the idea that extracting spatiotemporal information with Spatiotemporal Convolutional Neural Network (STCNN) is preferable than aggregating spatial features.

Audio Visual Speech Recognition (AVSR) for Amharic Language Using LIP Movement [14]**.** The goal of this research is to identify lip movements and define their relationships with speech sounds in order to improve speech recognition systems' performance, especially in noisy environments. For face and mouth detection, the authors used a Viola-Jones object recognizer called HaarCascade face detection and HaarCascade mouth detection, with the mouth Region of Interest (ROI) extracted after detection.

Visual feature extraction approaches use the extracted area of Region of Interest (ROI) as an input. Visual feature extraction is done with Discrete Wavelet Transform (DWT), and visual feature vector reduction is done with Linear Discriminant Analysis (LDA). They used Mel-frequency Cepstral Coefficients (MFCC) to extract audio features. In decision fusion, visual and auditory features are combined. Three separate classifiers were utilized as a result. For Audio Speech Recognition (ASR), the Hidden Markov Model (HMM) classifier is utilized, whereas the Coupled Hidden Markov Model (CHHM) is used for audiovisual integration and the Hidden Markov Model (HMM) classifier is used for Visual Speech Recognition (VSR). A microphone captures an audio speech signal from a speaker, while a camera captures video frames of the speaker's face. After that, the audio and video streams can be examined at the signal level. This system has two streams of speech information: an audio stream and a visual stream, in contrast to Audio Speech Recognition (ASR) systems, which have only an audio stream of information. The audio and visual streams are divided and analyzed separately in this architecture to extract key properties. Once the critical information from each signal has been extracted, it is combined and used to improve speech recognition. This system has two streams of speech information: an auditory stream and a visual stream. To extract significant characteristics, audio and visual streams are divided and analyzed separately in this architecture. Once the important information from each signal has been recovered, it is combined and used to improve speech recognition.

Because speaker videos contain information unrelated to the speech, such as the speaker's identify and background, the visual front-end component must remove this irrelevant data, leaving only the speech-related data. Prior to the extraction of visual speech characteristics, the mouth region of the speakers is recognized and a Region of Interest (ROI) is separated. The preprocessing of video inputs is the first step in front-end processing, followed by feature extraction, which converts speech into a parameter vector suitable for further processing. The visual front-end recognizes the

most speech information-rich area of the speaker's face following the mouth and extracts it in a parametric fashion that the recognizer can handle. Visual preprocessing, area of Region of Interest (ROI) extraction, and feature extraction are the three subtasks in the frontend component.

They employed Hidden Markov Model (HMM) Visual feature categorization as the final stage of visual speech recognition for this research. They chose Hidden Markov Model (HMM) for the classification technique because of its widespread use, which has resulted in multiple successful applications in statistical modeling of audible speech and the availability of Hidden Markov Model (HMM) implementation packages in Python programs (e.g., hmmlearn, scikit-learn).

Recognition is the process of choosing the most likely Hidden Markov Model (HMM) from a set of Hidden Markov Model's (HMM) that can produce a particular observed sequence (created during the training phase). They used forward and backward algorithms to calculate the likelihood of a model being constructed for a specific observation sequence in this method. The probability may be calculated successfully using dynamic programming by employing a forward method that reproduces the observation using Hidden Markov Model (HMM) and a backward way that back-traces the observation using Hidden Markov Model (HMM). To build a collection of feature vectors that represent the frames, the same processes as in the training phase will be used, as well as a vector quantization method that allows them to replace a series of feature vector values for a set of unique symbols. For a variety of words or phones (Amharic phones), each trained Hidden Markov Model (HMM) will be compared to a specific image sequence, and the model with the best probability will be picked. The remainder of this work concentrates on audio processing and, finally, merging the previously completed video and audio processing.

The system was put to the test by recording videos and audio for testing and training purposes. They employed two basic evaluation criteria for phone (vowel) and word recognition: speakers dependent and speakers independent. Based on the first evaluation criteria, word recognition was 60.42 % for visual only, 65.31 % for audio solo, and 70.1 % for audio-visual (speaker-dependent). They discovered that overall vowel (phone) recognition was 71.45 % for visual only, 76.34 % for audio solo, and 83.92 % for audio-visual speech based on the speakers' dependent assessment criteria.

A computer vision technique to automatic word detection for Amharic speech based on lip motion [41]. Using the information accessible in lip movements, this work proposes a solution for

automatic lip motion recognition for the Amharic language spoken by speakers by recognizing lip motions and defining their association with uttered words. The method used in this work is a computer vision technique that uses shape information from lip features to determine distinct aspects of the lip contour. The Artificial Neural Network (ANN) and Support Vector Machine (SVM) are used to detect facial parts, extract features using shape information, and recognize them Support Vector Machine (SVM). A computer vision technique to automatic word detection for Amharic speech based on lip motion [41]. Using the information accessible in lip movements, this work proposes a solution for automatic lip motion recognition for the Amharic language spoken by speakers by recognizing lip motions and defining their association with uttered words. The method used in this work is a computer vision technique that uses shape information from lip features to determine distinct aspects of the lip contour. The Artificial Neural Network (ANN) and Support Vector Machine are used to detect facial parts, extract features using shape information, and recognize them Support Vector Machine (SVM).

The recognition of Amharic visemes at the word level of visual speech is the subject of this study. The proposed method studied the lip motion shape extraction model by creating three primary methodologies, according to the authors. The Viola-Jones object detection technique was used to detect the mouth region, and the YIQ color space was employed to offer appropriate discrimination between the lip region and the face area, and lip counter shape information was provided as input to the classifier. On each classification parameter, they compared the classification methodologies of Artificial Neural Network (ANN) and Support Vector Machine (SVM) classifiers with the average shape information features. Finally, they concluded that their research reveals that the Support Vector Machine (SVM) classifier outperforms the Artificial Neural Network (ANN) in terms of classification performance. Artificial Neural Network (ANN) and Support Vector Machine (SVM) were used to achieve classification accuracies of 65.71 % and 66.43 %, respectively.

Audio-Visual Speech Recognition for other Languages

Other languages using Audiovisual Speech Recognition (AVSR) [42] created a speech recognition system for the French language that employs both acoustic and visual speech data to increase recognition performance in noisy conditions. Their system is made up of three parts: A visual module, an acoustic module, and a sensor fusion module are all included. The visual module

locates and tracks a speaker's lip movements, extracting relevant speech features. This task is carried out with the help of an appearance-based lip model that has been learned from examples. The contour information of the lips and the grey-level information of the mouth area are used to depict visual speech elements. The acoustic module extracts features from the audio signal that are noise-resistant. The multi-stream method allows alternative temporal topologies and levels of stream integration to be defined, allowing for more realistic modeling of temporal dependencies than traditional approaches. They show how to learn the asynchrony between the two modalities and how to incorporate it into multi-stream models using two alternative ways. The suggested system's better performance is demonstrated using a huge multi-speaker database of constantly uttered digits. Acoustic Perceptual Linear Prediction (PLP) features had a 56% error rate on a recognition task with a 15 dB acoustic Signal to Noise Ratio (SNR), noise robust RASTA-PLP (Relative Spectra) acoustic features had a 7.2% error rate, and combined noise robust acoustic features and visual features had a 2.5% error rate.

The creation of a modular system for flexible human–computer interaction via voice was presented by [43]. The speech recognition component combines acoustic and visual information (automated lip-reading) to improve overall recognition, particularly in loud situations. The lip location module extracts the image of the speaker's lips, which serves as the visual input, automatically from a camera image of the speaker's face. Finally, the face tracker sub-system automatically detects and tracks the speaker's face. The first bi-modal speech recognizer is the result of combining the three functionalities, enabling the speaker acceptable freedom of movement inside a potentially noisy room while continuing to communicate with the computer via voice. For varied signal/noise settings, the combined system achieves a 20 to 50% mistake rate decrease when compared to audio-only recognition. They also demonstrated the components of a lip-reading/speech recognition system that obtains the essential visual information non-invasively and automatically. They work together to perform automatic lip-reading in realistic scenarios where lip motion information improves speech recognition in both good and bad acoustic environments. Simultaneously, the speaker is given reasonable mobility inside a room, with no requirement to situate himself in any certain location.

For the English language, [7] described a lip-reading technique based on motion capturing and Support Vector Machine (SVM). The experimental results show that normalization using linear

interpolation and Mean Square Error (MSE) can overcome inter and intra subject speech speed variations, and that the vertical component of optical flow can be used for speech recognition. The features are chosen using a robust feature selection technique based on non-overlapping fixed size columns. The features are classified using Support Vector Machine (SVM). The findings show that the technique described can achieve very high success rates. The overall accuracy was 95.9%, the specificity was 98.1%, and the sensitivity was 66.4%.

Rodomagoulakis [11] looked into various visual feature extraction and audio-visual integration issues in the field of Audiovisual Automatic Speech Recognition (AV-ASR). The mouth region, which is called the Region of Interest (ROI), is detected and tracked across sequential time frames using color-based detection and template matching algorithms. Following that, pixel data are transformed into "compact," descriptive features using the Discrete Cosine Transform (DCT). The author shows that various aspects of Region of Interest (ROI) detection, such as the tilt and size of the speaker's head, have an impact on the performance of both visual and audio-visual recognizers.

To counteract these impacts, the author looked at rotation correction and scaling normalization approaches. Over a baseline implementation, the improved visual front-end schema resulted in a 95% reduction in Word Recognition Error (WRE). On the other hand, the author looked at a new K-means clustering-based approach for unsupervised stream weight estimation. For classification and identification tasks, stream weight behavior and adaptation are assessed under a variety of noises at the word or phrase level. Finally, they compared the outcomes of static and adaptive weighting to evaluate their weight estimation approach, and they also measured the improvements made by utilizing an audio-visual recognizer rather than a typical audio-only recognizer. All experiments used the CUAVE AV English language database, and recognizers were created with Hidden Markov Model Toolkit (HTK).

For the Japanese language, [43] suggested an Audiovisual Speech Recognition (AVSR) system based on deep learning architectures for audio and visual feature extraction, as well as a Multi-stream Hidden Markov Model (MSHMM) for multimodal feature integration and isolated word recognition. The deep de-noising auto encoder, when compared to the original Mel-frequency Cepstral Coefficients (MFCCs), can effectively filter out the effect of noise superimposed on original clean audio inputs, and acquired de-noised audio features achieve significant noise robustness in an isolated word recognition task. Furthermore, our Convolutional Neural network

(CNN) based visual feature extraction method accurately predicted the phoneme label sequence from the mouth region image sequence. In the isolated word identification task, the acquired visual features outperformed standard image-based visual features such as PCA by a large margin. Finally, by combining the acquired audio and visual data, Multi-stream Hidden Markov Model (MSHMM) was used for an Audiovisual Speech Recognition (AVSR) task. Their findings showed that accurate results can be achieved even with a simple yet intuitive multimodal integration mechanism, it is conceivable to achieve dependable Audiovisual Speech Recognition (AVSR) performance by adaptively switching the information source from audio feature inputs to visual feature inputs in response to variations in signal input reliability. Although they were unable to achieve automatic stream weight selection, their experimental results proved the benefit of using Multi-stream Hidden Markov Model (MSHMM) as an Audiovisual Speech Recognition (AVSR) mechanism. The next main goal of their research is to see if we can use our existing technique to create useful, real-world applications. Future work will include a study to see how the Visual Speech Recognition (VSR) approach, which uses Convolutional Neural Network (CNN) acquired translation, rotation, and scaling invariant visual features, contributes to robust speech recognition performance in a real-world environment with dynamic changes like reverberation, illumination, and facial orientation.

2.7 Speech Recognition for Amharic Language

Asratu Aemiro (2015) [44] created two types of Amharic voice recognition systems: canonical speech recognizers and enhanced speech recognizers. The canonical Audio Speech Recognition (ASR) system is built on the canonical pronunciation model, which includes a canonical pronunciation dictionary as well as a decision tree. Each different word in the vocabularies is represented by a single pronunciation in the canonical pronunciation dictionary. The canonical decision tree is built by solely considering the location of phoneme articulations, as was done by earlier Amharic Audio Speech Recognition (ASR) researchers. On the other hand, the development of an enhanced speech recognition system employs an enhanced pronunciation model, which consists of an enhanced pronunciation dictionary and an enhanced decision tree, both of which are designed by taking into account the patterns identified based on phoneme co-articulation effects.

Solomon Berhanu (2001) [45] conducted automatic speech recognition for Amharic in 2001. The author created an isolated consonant-vowel syllable Amharic recognition system that uses (Hidden Markov Model Toolkit (HTK) to recognize a subset of isolated consonant vowel syllables. The author chose 41 CV syllables of Amharic language out of 234 and captured speech data from 4 (four) males and 4 (four) females ranging in age from 20 to 33 years. The average recognition accuracies for speaker dependent and independent systems were 87.68% and 72.75%, respectively.

Kinfe Tadesse (2002) [46] used Hidden Markov Model Toolkit (HTK) to create standalone Amharic word recognition systems based on sub-words. Phones, triphones, and Consonant-Vowel syllables (CV-syllables) were employed as sub-word units in this experiment, and the system was developed using 20 phones out of 37 and 104 Consonant-Vowel syllables (CV-syllables). Speech data was collected from 15 speakers for training and 5 speakers for testing. For speaker-dependent phone-based and triphone-based systems, respectively, average recognition accuracies of 83.07% and 78% were found. In terms of speaker independent systems, phone and triphone-based speaker independent systems had average recognition accuracies of 72% and 68.4%, respectively.

Tamrie, Z. (2021) [47] *in* order to extract the features, they used Convolutional Neural Networks (CNN), Histograms of Oriented Gradients, and their combination approaches. To recognize the spoken word, we feed these features to random forest singly and in combination. Each of these attributes was assessed using precision, recall, and fl-score classifiers to measure our model's performance and compare its accuracy to earlier relevant efforts. On HOG, Convolutional Neural Networks (CNN), and mixed features, our model system achieves 66.03%, 75.24%, and 76.51% accuracy, respectively.

Befkadu Belete (2017) [48] Hidden Markov Model (HMM) classifiers are employed for Audio Speech Recognition (ASR), Coupled Hidden Markov Model (CHMM) classifiers for audiovisual integration, and Hidden Markov Model (HMM) classifiers for Visual Speech Recognition (VSR) The method starts by capturing an audio speech signal from a speaker using a microphone and video frames of the speaker's face using a camera. After that, the audio and video streams can be examined at the signal level. Unlike Audio Speech Recognition (AVR) systems, which only have an audio stream of data, this system has two streams of speech data: an audio stream and a visual stream. Audio and visual streams are divided and analyzed independently in this architecture to

extract relevant properties. Once the critical information from each signal has been extracted, it is combined to improve speech recognition.

The Coupled Hidden Markov Model (CHMM) classifier is utilized for audiovisual integration, while the Hidden Markov Model (HMM) classifier is used for Visual Speech Recognition (VSR). The system begins by capturing an audio speech signal from a speaker through a microphone and video frames of the speaker's face through a camera. The audio and video streams can then be signal-level examined. This system, unlike Audio Speech Recognition (ASR) systems, has two streams of speech information: an audio stream and a visual stream. The audio and visual streams are divided and analyzed independently in this architecture to extract relevant properties. After obtaining the crucial data from each signal, it is combined to improve speech recognition.

Because speaker videos contain information that is unrelated to the speech, such as the speaker's identification and background, the visual front-end component must remove this unneeded information, leaving only the speech-related information. Prior to extracting visual speech characteristics, the mouth region of the speakers is recognized and a Region of Interest (ROI) is separated. Front-end processing begins with video input preprocessing, which is followed by feature extraction, which converts speech into a parameter vector suitable for subsequent processing. The visual front-end identifies the area of the speaker's face following the mouth that has the most speech information and extracts it in a parametric format that the recognizer can handle. The frontend component has three subtasks: visual preprocessing, Region of Interest (ROI) extraction, and feature extraction.

The system was tested using movies and audio recordings for testing and training. They employed two basic evaluation criteria for phone (vowel) and word recognition: speaker dependent and speaker independent. Based on the first evaluation criteria, overall word recognition was 60.42% for visual only, 65.31% for audio-only, and 70.1% for audio-visual (speaker-dependent). According to the speakers' dependent assessment criteria, overall vowel (phone) recognition was 71.45% for visual only, 76.34% for audio solo, and 83.92% for audio-visual speech.

Muluken B. (2017), [49] using the information accessible in lip movements, this work proposes a solution for automatic lip motion recognition by recognizing lip motions and defining their association with spoken words for the Amharic language spoken by speakers. The method used in this study is a computer vision technique that uses shape information from lip characteristics to

distinguish different elements of the lip contour. Face parts are detected, features are extracted using shape information, and recognition is performed using the Artificial Neural Network (ANN) and Support Vector Machine (SVM). To detect the mouth region, the Viola-Jones object identification technique was employed, and the original mouth image was transformed into YIQ color space and applied to the saturation components to detect the lip image from the face area.

This study looks towards recognizing the word level of visual speech for Amharic visemes. According to the authors, the proposed technique studied the lip motion shape extraction model by developing three key methodologies. The model uses the Viola-Jones object identification technique to detect the mouth region, the YIQ color space to offer sufficient discrimination between the lip region and the face area, and lip counter shape information as input to the classifier. They compared the classification procedures of Artificial Neural Network (ANN) and Support Vector Machine (SVM) classifiers with the average shape information features on each classification parameter. Finally, they stated that their experiment demonstrated that the Support Vector Machine (SVM) classifier outperformed the Artificial Neural Network (ANN) in classification performance. Artificial Neural Network (ANN) and Support Vector Machine (SVM) produce classification accuracy of 65.71% and 66.43%, respectively.

2.8 Summary

Previous works for the Amharic language have exclusively used audio to recognize speech. Traditional acoustic-based speech processing systems have recently achieved high levels of performance, but their effectiveness is strongly reliant on a match between training and test conditions. The performance of acoustic speech processing applications might decline dramatically in the presence of mismatched conditions (e.g., acoustic noise). The visual speech modality is unaffected by the majority of auditory modality degradations. Because of its independence, as well as the bimodal character of communication, the visual speech modality can naturally operate as a complement to the acoustic speech paradigm.

We have discussed and enumerated the literature we read and papers connected to our research in this chapter. Computer vision is concerned with the collection, processing, analysis, and comprehension of digital data, as well as the extraction of high-dimensional data from the real world in order to generate numerical or symbolic information. When doing computer vision tasks

on video data, they must be transformed into continuous image frames. Different image preprocessing operations will be done on the image file after the video file has been converted into picture frames. Then, characteristics will be extracted from the image files, and each image's transition will be evaluated to anticipate the video file's content. Video processing is a technique that involves the use of a series of rapidly moving image sequences. We're generally talking about image processing when we talk about video processing. As a result, image processing should follow particular processes in order to identify an image, predict an image, improve an image's quality, or process an image further. Visual speech understanding or lip reading is one of the applications of computer vision. Lip-reading is the process of deciphering speech by visual clues such as the movement of the speaker's mouth. The visual speech recognition video file's input data. This is known as face localization. Following face localization, the next stage will be to locate the speaker's Region of Interest (ROI), which is the lip section of the speaker's face. The mouth region of the speaker is primarily of relevance to lip-reading systems. Feature extraction and classification activities will be conducted on the preprocessed frames after locating the region of interest for each of the frames. The visual character of the Amharic language is another issue we explored in this chapter. In the visual domain, a viseme is a mouth shape (or expression) or a set of mouth motions required to create a phoneme. Many challenges arise when visemes are used in visual speech recognition systems. These issues suggest that mapping the vision with the motion of the lip is challenging, and that the information may be corrupted as a result. Like any other language, Amharic has its own distinct qualities. Amharic, for instance, features a set of speech sounds that aren't found in other languages like English. These are glottalized plosives with acute click-like characteristics ጵ, ጥ, ጭ, ቅ, ጽ like every other language, Amharic has its unique viseme-phoneme mapping. In addition, we've talked about connected works. There has been a lot of study done on visual speech recognition. We were unable to discuss each paper individually, but we did attempt to do so for a few of them.

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This section covers the architecture, methods, materials, and techniques. We looked at many architectures used in the field of visual voice recognition and created our own that we believe is the best for our domain and specifications. We used video footage to do our computer vision study. The video data will be subjected to some preprocessing procedures. When it comes to approaches and techniques, it's already been decided that deep learning algorithms will be used. The term "deep learning algorithm" is, in fact, fairly broad. We've determined which form of deep learning algorithm will be used in our study. We looked at some articles to see if they may assist us narrow our selection boundary and choose the algorithm that best fits our domain. The process will be divided into two parts: feature extraction and classification. A deep learning method will be used for both feature extraction and classification. The specifics of everything mentioned above will be covered in the sections that follow.

3.2 Architecture of the System

The system architecture shows how our visual speech recognition procedure should be carried out in its entirety. There is video data recorded from the speaker right at the start. Because a video is made up of a collection of consecutive images, the video data will be framed into a series of image frames. Image preprocessing processes are applied to each of the image frames. The frames contain extra information that isn't required for our work, such as the image's background. The only information we want from the speaker is the face portion of the speaker's photograph, following which we will proceed.

This system has a video stream. Visual streams processed in this architecture to extract essential information. Once the essential data has been extracted from each signal, it is fused together and used to improve speech recognition. The design of the frontend processing system for visual, the integration of visual vectors, and the training of the recognizer are the three key steps of the system implementation.

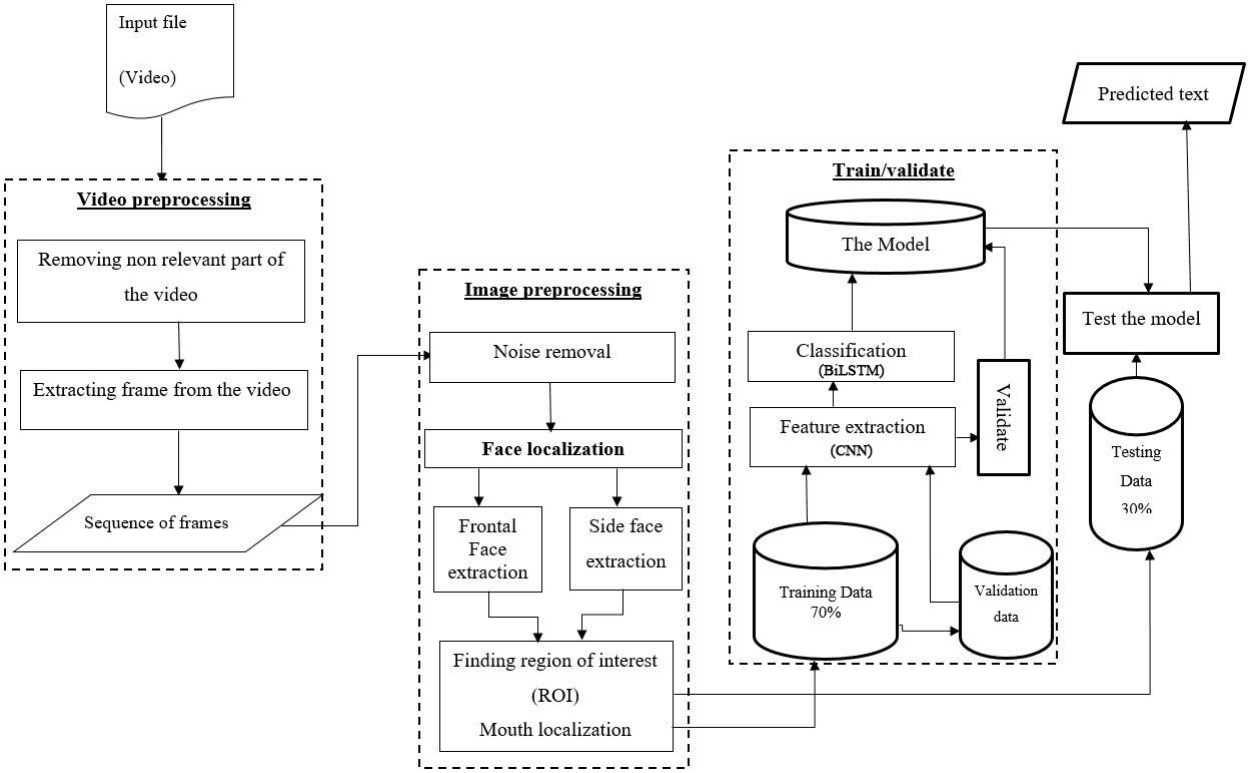

Figure 1: Overall system architecture

As shown in the figure 2 below, the frames that were collected from the video images were used for the inputs to the Convolutional Neural Network (CNN) layer and used the correct features from images inter the Long Short Term Memory (LSTM) model to have a predicted model.

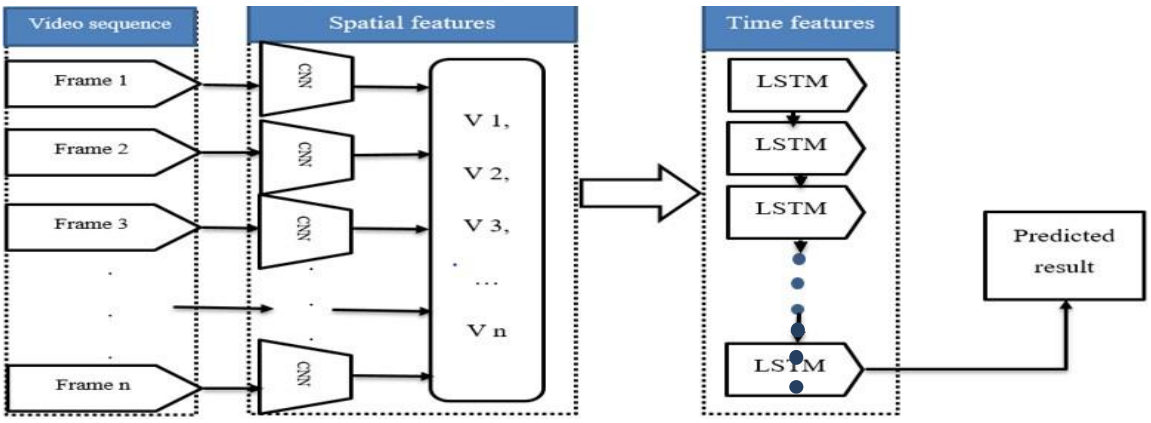

Figure 2: The architecture of the system proposed

3.2.1 Video Processing

At the Ethiopian Satellite Television (ESAT), the video is captured from the speakers using a high definition camera situated 2 (two) to 3 (three) meter away from the speakers. Many elements of the video are irrelevant to our task. The background of the speaker or other organs of the speaker that are unrelated to our work are examples of irrelevant portions. The face of the speaker is the most important component of the video. The extraction of the frame sequence and the relevant section of each frame are both part of the video preprocessing. To extract a sequence of frames from the video, we used the OpenCV library. HaarCascade is a machine learning object recognition technique that employs the feature idea proposed to recognize objects in an input image or sequence of images (video) [44].

Since we are working on frontal side video it is a must to recognize the speaker's frontal faces. To recognize the speaker's face at frontal side, we employed the CAFFE model for the human face extraction algorithm. Even if Viola and Jones (HaarCascade) architecture is superior at extracting the speaker's frontal face. However it is only capable of recognizing the speaker's face when the video is taken from the front.

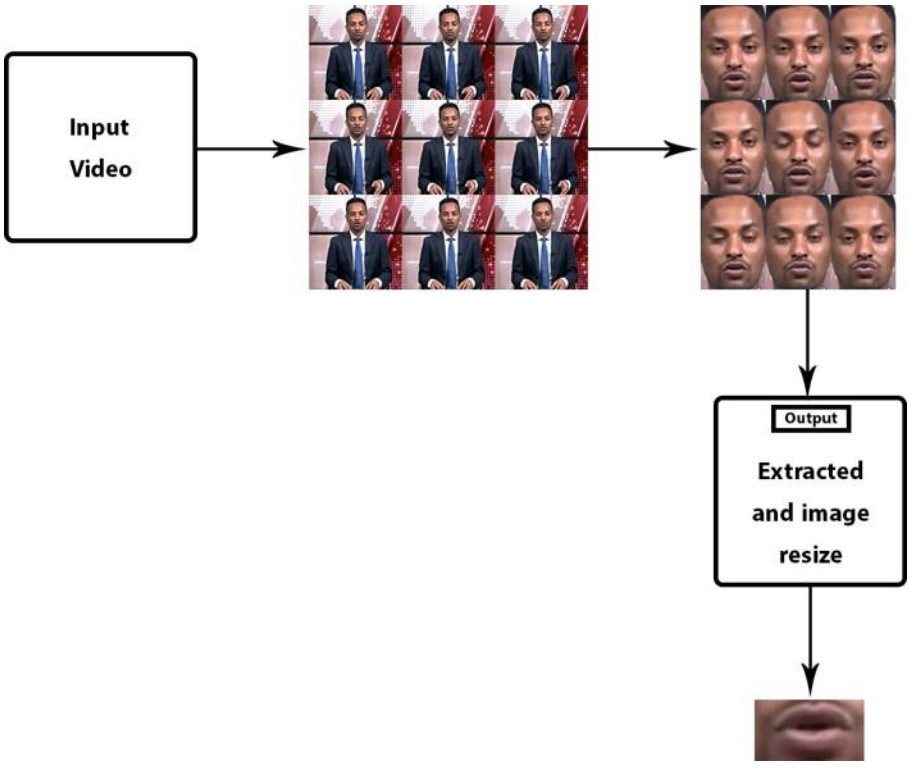

Figure 3: Video Processing

Noise Removal: The process of removing noise from a noisy image in order to restore the true image is known as image denoising [45]. Noise unavoidably degrades images produced by modern cameras, resulting in lower visual image quality. As a result, making an effort to decrease noise while keeping image attributes is necessary (edges, corners, and other sharp structures). Image denoising techniques can be divided into two categories. There are two types of picture denoising techniques: linear image denoising and non-linear image denoising. Denoising algorithms that are linear are superior for speed but have poor performance. Linear denoising approaches are more effective in terms of performance, but they are slower [46]. We employ nonlinear image denoising algorithms in this paper. There are a plethora of non-linear image noise removal techniques to choose from. We employed the median filter noise removal method for our goals. We chose this approach because it keeps edges while eliminating noise[47].

3.2.2 Finding Region of Interest (ROI)

The Region of Interest (ROI) supplies the raw input data for visual feature extraction, and hence the accuracy of Region of Interest (ROI) extraction has a significant impact on the overall performance of Visual Automated Speech Recognition (VASR) system. The high deformation of the lip shape, as well as the variation in the content of the mouth region due to the presence or absence of tongue, teeth, and the opening and closing of the lips during speaking, make identifying the ROI more challenging. Variations in illumination conditions and changes in the position and orientation of the speakers can also alter Region of interest (ROI) detection methods. Region of interest (ROI) extraction is also influenced by the presence or absence of a beard or moustache.

After extracting the face portion of the frames, the next step is to extract the frames' Region of interest (ROI). To extract a sequence of frames from the movie, we used the OpenCV library. HaarCascade is a machine learning object recognition technique that employs the feature notion proposed by Paul Viola and Michael Jones in their work "Rapid Object Detection Using a Boosted Cascade of Simple Features" to recognize objects in an input picture or video sequence [44]. We used the Viola Jones object detection technique for the majority of our model's preprocessing phases. This assignment is simple for humans, but it is difficult for computers to complete. It necessitates specific directions and limitations. To make the task more doable, Viola–Jones requires a comprehensive picture of human facial elements. Integral pictures, an intermediate

picture format, can be used to determine rectangle features reasonably and quickly. The pixels above and to the left of the, inclusive, make up the integral picture at position:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'),$$

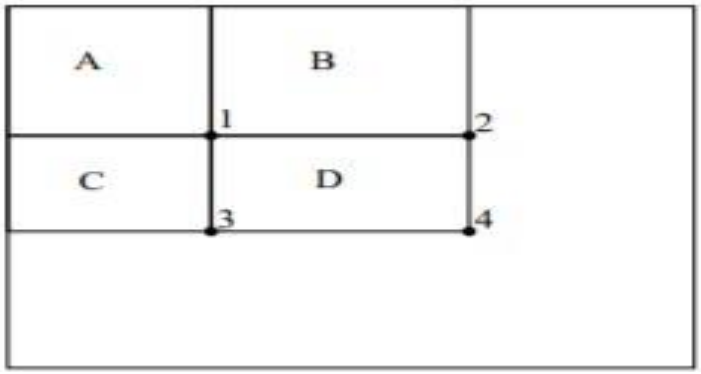

Figure 4: With four arrays, the sum of rectangular pixels may be calculated

The sum of all the pixels in the rectangle is the integral image value at location 1. A+B is the value on location 2, A+C is the value on location 3, and A+B+C+D is the value on location 4. 4+1- (2+3) can be used to calculate the total within [44]. The discovery of the smaller value is significant, and it is important to note that it is more efficient. Boosted classifiers can be configured to reject the vast majority of negative sub-windows while identifying almost all positive occurrences (i.e., the false-negative rate of a boosted classifier can be reduced to near zero). To achieve a low false-positive rate, simpler classifiers were employed to reject the majority of the sub-windows before applying more advanced classifiers. The entire detection procedure is based on a degenerating decision tree, dubbed a "Cascade" by the authors. When the first classifier returns a positive result, a second classifier is evaluated, which has also been tweaked to achieve extremely high detection rates. The third classifier is triggered by a positive result from the second, and so on. If you get a negative result at any point during the process, it means the sub window is being denied right away. The default AdaBoost threshold was created with a low error rate in mind while working with training data. A lower threshold value, in general, results in better detection but a higher false-positive rate.

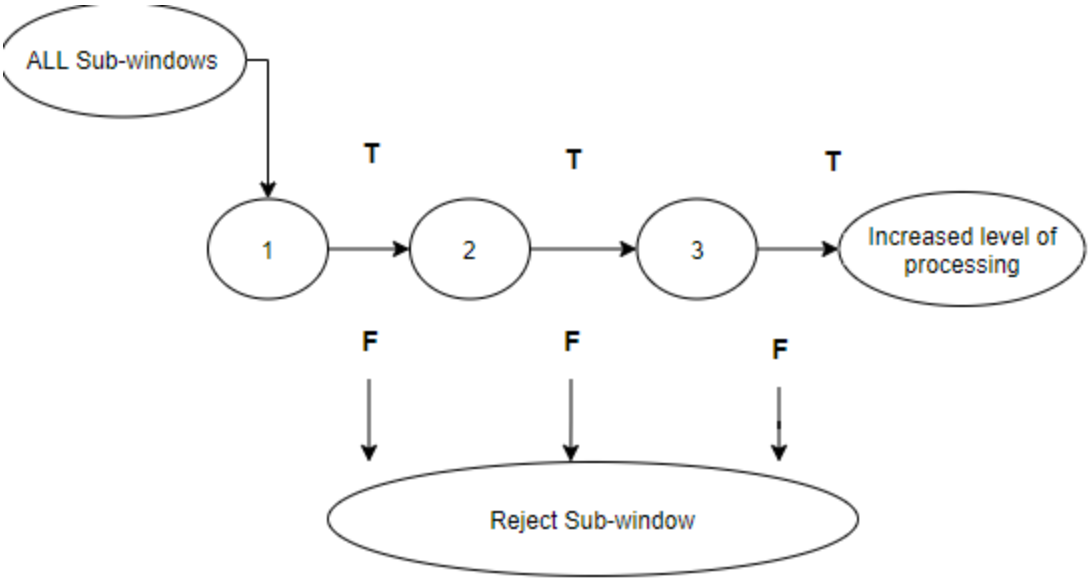

Figure 5: The detection cascade's architecture [44]

In the schematic architecture of detection matrix of [44] each sub-window is subjected to a series of classifiers. The first classifier eliminates a large number of bad cases with minimal processing. More negatives are removed in subsequent layers, but this necessitates more processing. The number of sub-windows has been significantly reduced after multiple processing processes. More phases of the cascade (as in our detection system) or a different detection method can be used for additional processing.

3.3 Feature Extraction

The goal of feature extraction is to keep as much speech-related information as possible in a minimal number of parameters from the original photographs of the speaker. A variety of transformation algorithms are employed in visual feature extraction, including Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Principal Components Analysis (PCA), and Linear Discriminant Analysis (LDA). Feature extraction, according to [48] , is an important aspect of image processing and object recognition. Feature extraction takes a set of measured data and generates derived values (features) that are intended to be useful and non-redundant, allowing for faster learning and generalization, as well as, in certain situations, better human interpretation. The extraction of features is linked to dimensionality reduction [49].

The Wavelet Transform (DWT) transform decomposes the input image into low-frequency sub bands (known as approximate images) and high-frequency sub bands (known as detailed images).

The low frequency contents of the image are contained in the LL region of the Wavelet Transform (DWT) transform, the high-frequency horizontal details are contained in the HL region, the high-frequency vertical details are contained in the LH region, and the high-frequency details for both the horizontal and vertical directions are contained in the HH region. When the Wavelet Transform (DWT) is applied to an image, it produces high-pass and low-pass filtering. Higher levels of decomposition can be used to extract more precise features from an image.

When features are picked, it is anticipated that they will contain all of the essential information from the input data for the intended purpose, allowing the desired process to be carried out using this reduced representation rather than the complete original input data. Feature extraction is the process of decreasing the amount of resources needed to describe a huge set of data. One of the key challenges that arises while reviewing sophisticated data is the large number of variables involved in the process. A large number of variables necessitates a significant amount of storage and calculation time (primarily time), and it also increases the risk of the classification algorithm overfitting training data and failing to generalize to fresh samples.

In Deep Neural Networks (DNN), the attributes of input images are expressed in an intermediate layer, which has piqued researchers' interest in recent years [50]. Deep learning using Deep Neural Networks (DNNs) has gotten a lot of attention in recent years when it comes to image processing because it has successfully demonstrated automatic organization of neurons selectively reacting to images of a cat, as well as high recognition performance in relation to difficult tasks such as general object recognition [50]. To learn about thousands of items from millions of input data, a model with a big learning capacity is necessary. Due to the huge complexity of the object recognition challenge, even a dataset as large as ImageNet is insufficient fully specify this problem, hence our model will need a lot of previous knowledge to compensate for the data we don't have [51]. One sort of model in this area is Convolutional Neural Networks (CNNs) [43]. They may modify the scope and depth of their learning, and they develop strong and mostly true assumptions about the world. They may modify the breadth and depth of their learning, and they make strong and largely valid assumptions about the nature of images (namely, statistical stationarity and pixel-dependent locality). Regular feedforward neural networks, which have similar layer sizes, have considerably fewer connections and parameters than Convolutional Neural Networks (CNN). As a result, they are less difficult to train, and their theoretically best performance is only little inferior [51].

Convolutional Neural Networks (CNN) is a sort of artificial neural network that excels at identifying and comprehending patterns. Convolutional Neural Networks (CNN) is hence suitable for picture analysis [52]. Convolutional Neural Networks (CNN) is made up of layers that are stacked in a sequential order, each performing its own set of functions on the data input to it. The input layer holds the raw data, the convolutional layer computes the output volume by performing a dot product between the image patch and all of the filters, the activation function layer applies an activation function to every element of the convolution layer's output, and the pooling layer pools the output of the pre-processing layer. A fully connected layer that receives input from the preceding layer and outputs a one-dimensional array of class scores calculated [52]. The features that will be implemented are dictated by the type of system that will be utilized to implement them. In image processing, the most popular categories are spectral and geometric.

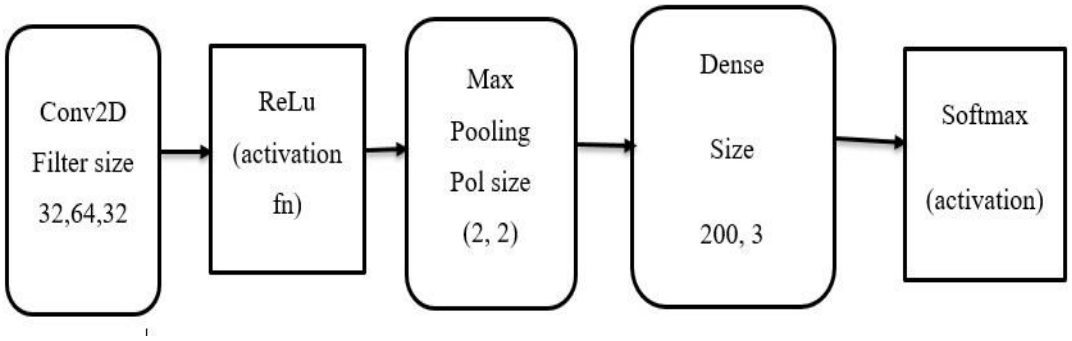

Figure 6: Convolutional Neural Network (CNN) layers

3.3.1 Convolution Layer

The process of extracting features from the image input through the convolution layer can be thought of as the process of extracting features from the image. When it comes to comprehending images in human vision and computer vision, the two are vastly different. When human eyesight recognizes a picture, for example, the image's brightness, size, and shape are used to identify it. The image is in a matrix format with simply numbers in computer vision.

When a computer learns an image, it must first extract the image's features from this matrix. This is accomplished by convolution in the picture feature extraction process. When we take a 5X5 matrix image, the 3X3 matrices that slide along the image with a step size of 1 is called a filter (or convolution kernel). Each time the filter slides, it multiplies its value by the picture matrix, and

the results are added together. The resulting matrix is a feature matrix element. The feature matrix of the image can be obtained after passing each image value.

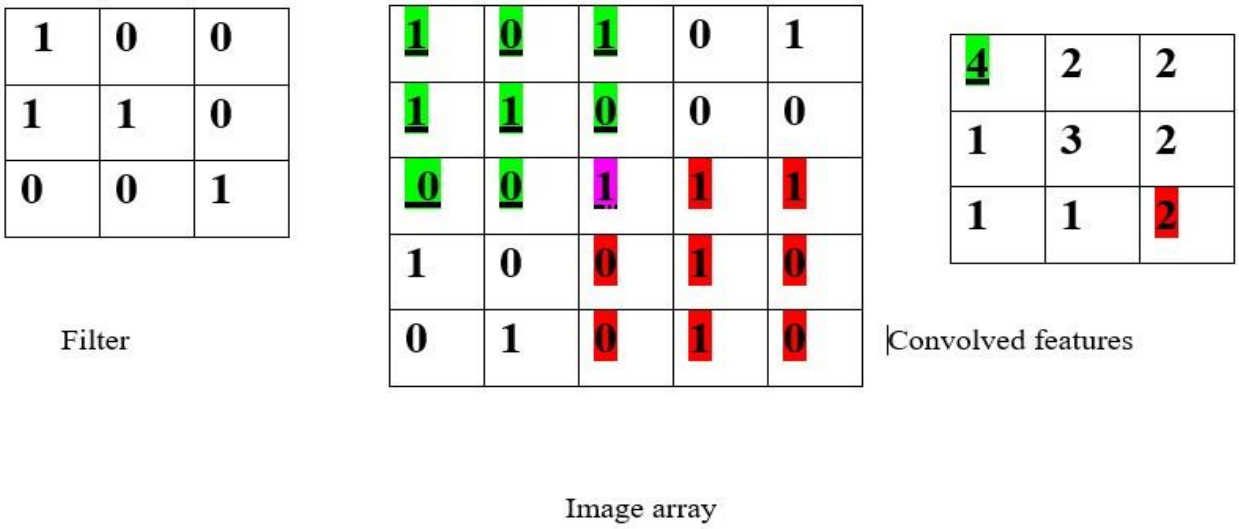

Figure 7: Convolution layer technique for feature extraction

3.3.2 Max Pooling Layer

The pooling layer is frequently inserted between the convolution layers on Convolutional Neural Networks (CNN). The pooling layer can substantially minimize the size of the parameter matrix and parameter sizes in the final fully linked layers, speeding up calculations and preventing over-fitting. The size of the trained images is too huge in image recognition and classification. To reduce the amount of training parameters, we need to add a pooling layer between the convolution layers. Pooling occurs in all depth dimensions. The image's depth remains constant. Max pooling is the most prevalent type of pooling. Specifying the sliding size or filter size then taking the maximum from the values in the sliding matrix is one process of max-pooling. For example, considering 4x4 matrix and the sliding size or the filter size is 2x2 the step size is set to 2. The maximum value in the filter region for each slide is then selected. The process is shown in the figure below.

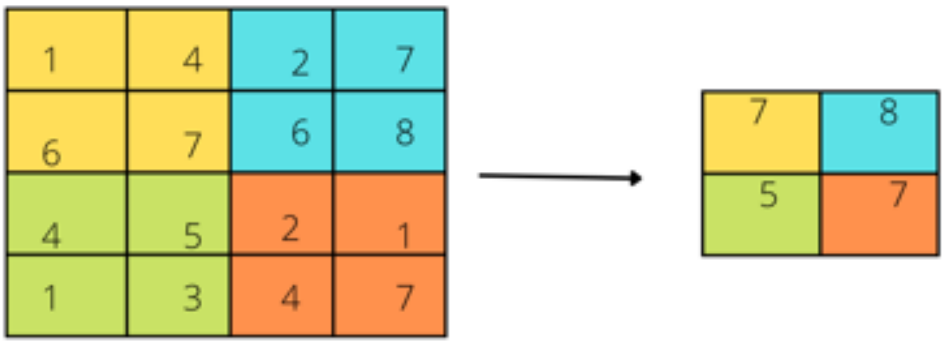

Figure 8: Max pooling with a 2 by 2 filter and a stride of 2

3.3.3 Dense Layers

Dense neural networks, which are built up of dense layers, are the most fundamental neural network architecture (also known as fully connected layers). All of the inputs and outputs to all of the neurons in each layer are connected by the dense layer. The convolutional neural network's last component. It takes the formal layers' outputs as inputs and translates them to the classification task's targets. Let's imagine we have five outputs from the formal convolution and pooling layers, and we want to categorize them into three groups. The five outputs, as we know, are the essential properties that can help us decide which category the input image belongs to, and the three categories are the classification task's targets, which are also the fully connected layer's outputs. The fully connected layer's weights and bias, along with the important characteristics, will create linear combinations to output the three categories needed to complete the classification assignment.

3.3.4 Lose Function

The lose function is an essential component of the Convolutional Neural Networks (CNN) architecture. The model's prediction quality, which tells us how well the neural network performs on a certain task, can be reflected in the loss function. During network training, the network produces predicted values through each layer of processes, and then calculates the difference between the predicted and true values using a loss function. The goal of training the neural network is to narrow the gap (loss). Cross-entropy loss, mean squared error, and hinge loss are the most well-known loss functions in deep learning algorithms. Our neural network's performance worsens when a large number of loose functions are registered. The loss should be kept to a minimum.

3.3.5 Activation Functions

The activation function is another crucial component of the network. An artificial neural network computes a "weighted sum" of its input, adds a bias, and then determines whether or not it should be "fired." This is something that the activation function takes care of.

Let us consider the equation:

$Y = \sum$ (Weight *input) + bias

Y can have any value between -infinitive and +infinitive. The neuron has no idea what the value's boundaries are. So, how do we determine if the neurons should fire or not? It is preferable to include an activation function for this purpose. To determine whether the neuron produces a value of Y and whether or not this neuron should be considered "fired" by external connections. Depending on the type of data and the amount of data being used, different activation functions might be used.

3.4 Classification

Predicting which class an item belongs to is referred to as classification. Classification algorithms can be seen in action in document categorization, email spam filtering, image recognition, speech recognition, and handwriting recognition. Therefore, a neural network is one of many machine learning methods that can be used to address categorization problems. Its strength lies in its ability to dynamically generate complex prediction functions and to consider how humans think in a way that no other algorithm can. Neural networks have produced the best results in a variety of categorization problems. Academics are considering Deep Learning as a way to open up new areas for research into the automated extraction of complex information at high levels of abstraction. Higher-level (more dynamic) highlights are specified in terms of lower-level (less differentiated) highlights, resulting in a hierarchical, tiered learning and presentation architecture [53]. Artificial intelligence inspired Deep Learning's progressive learning engineering, which mimics the highly layered learning process used by the human brain's primary sensory zones of the neocortex to separate highlights and reflections from basic information. The general architecture of deep learning is depicted in Figure below.

Image classification is one of the most basic jobs in image processing, video processing, and pattern recognition. It involves assigning, categorizing, and labeling an image into one or more classes. Low- or mid-level features are extracted to represent the image in classical image classification, and then a trainable classifier is employed to assign a label to it. In the deep learning era, the high-level feature representation of deep convolutional neural networks has proven to be considerably superior to hand-crafted low-level and mid-level features. Both feature extraction and classification networks can be merged and trained end-to-end in a deep convolutional neural network [54].

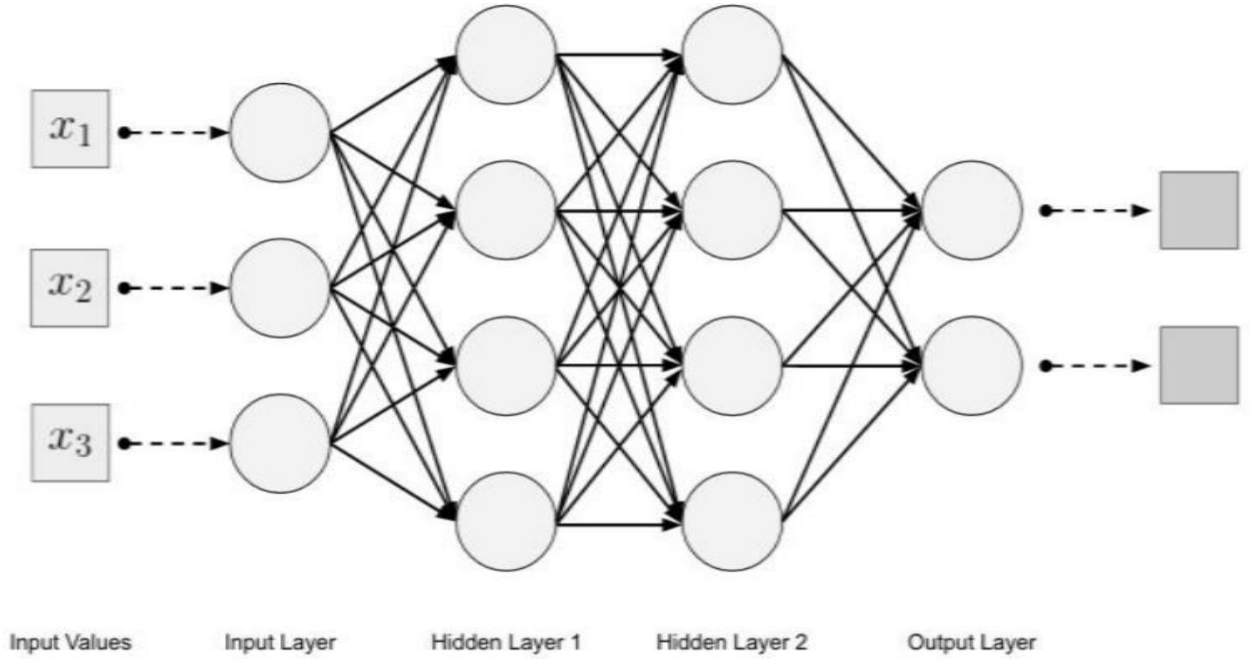

Figure 9: Dens Layer

In video classification, the algorithm must be able to recognize the sequential aspects in the visual sequence. Both general and spatiotemporal characteristics can be found in video data. The time dimension of the data is particularly important when classifying video data into different classes. The vast amount of video data necessitates the development of systems for automatically classifying video data into categories such as human actions and complicated occurrences. To account for temporal indications in movies, there is a considerable body of literature devoted to computing effective local feature descriptors from spatiotemporal volumes [55]. The Convolutional Neural Network (CNN) designs covered in the feature extraction are all Feed-forward Neural Networks (FFNNs), which are unsuitable for sequence labeling since their

connections do not form cycles. Recurrent network structures have been introduced to better grasp the temporal aspects of time series data, leading to the birth of Recurrent Neural Networks (RNNs). Recurrent Neural Networks (RNNs) allow cyclical connections to form cycles, allowing the network's internal state to retain a "memory" of prior inputs [55].

A Recurrent Neural Network (RNN) is a Neural Network in which the output of the previous step is used as an input for the current step. All the inputs and outputs were independent in prior methods (before the introduction of Recurrent Neural Network (RNN), but when scenarios like predicting the next word of a sentence are required, the previous words are necessary, hence remembering the previous word is required. To address this problem, the Recurrent Neural Network (RNN) with hidden layers was proposed. The most fundamental and significant element of Recurrent Neural Network (RNN) is the hidden state, which remembers certain information about the data's sequential structure. Understanding context is, as one might think, an important skill for tasks like sequence recognition and categorization. A Recurrent Neural Network (RNN) is replicated over time in order to evaluate a sequence of any size by flowing the contextual knowledge gained to update the corresponding weights that indicate its internal states [56]. Recurrent Neural Network (RNNs) have a "memory" that keeps all of the data about the calculations. Because it performs the same action on all of the inputs or hidden layers to generate the output, it uses the same parameters for each input. This, unlike other neural networks, decreases parameter complexity.

Consider a network having one input layer, three hidden levels, and one output layer. Each hidden layer will then contain a set of weights and biases, similar to other neural networks, such as (W1, B1) for the first layer, (W2, B2) for the second hidden layer, and (W3, B3) for the third hidden layer (W3, B3). This means that each layer in this system is self-contained and does not recall prior outputs. Recurrent Neural Network (RNN) then completes the following task:

- By applying the same weights and biases to all of the layers in the network, independent activations can be turned to dependent activations, reducing the complexity of raising parameters and remembering each previous output by feeding each output into the next hidden layer.
- Because all three hidden layers have the identical weights and biases, they may be combined into a single recurrent layer.

When we train a recurrent neural network, the following will happen:

- The network is given one time step of the input.
- Its present state is determined by a combination of current input and past state.
- For the next time step, the current state ht is changed to ht-1.
- As the nature of the problem dictates, we can go back in time and connect the data from all of the prior situations.
- The output is calculated using the final current state after all of the time steps have been completed.
- The error is then formed by comparing the output to the actual output, i.e. the desired output.
- Then feed the error back to the network, which will update the weights, allowing us to train the Recurrent Neural Network (RNN).

Standard Recurrent Neural Network (RNNs) are well known for having a narrow contextual information range. Long-term contextual dependencies are difficult to grasp. The problem stems from the amount of effect a given input receives in the buried layer [56]. Regrettably, ordinary Recurrent Neural Network (RNNs) can only access a limited amount of data in practice. When an input cycle passes through the network's recurrent connections, its effect on the hidden layer, and therefore on network output, either decays or increases exponentially. This limitation (known as the vanishing gradient problem in the literature) prevents an Recurrent Neural Network (RNN) from bridging time gaps of more than 10 steps between relevant input and target events [57]. The short-term memory of Recurrent Neural Network (RNN) is usually the source of its issues.

At time 1, the units are colored based on their sensitivity to the input. (With the black at the top and the white at the bottom.) As can be seen, the first input's impact diminishes substantially over time [57].

These gates can figure out whether a sequence of data should be preserved or destroyed. This enables it to convey relevant data down a long chain of sequences to create predictions. Almost all state-of-the-art outcomes based on recurrent neural networks are achieved using this network. The structure of the Long Short Term Memory (LSTM) network is shown in below, Long Short Term

Memory (LSTMs) rely on the cell state and its multiple gates. The cell state acts as a thoroughfare, transporting data all the way down the sequence chain. It's possible to think of it as the network's "memory." In theory, the cell state can carry useful information throughout the sequence's processing. As a result, data from past time steps may flow into succeeding time steps, reducing the impact of short-term memory. During the process, information is added or removed from the cell state via gates. The gates are neural network components that determine whether or not data is allowed to enter the cell state. The gates are in charge of learning and selecting what information is important to keep or forget during the course of the game.

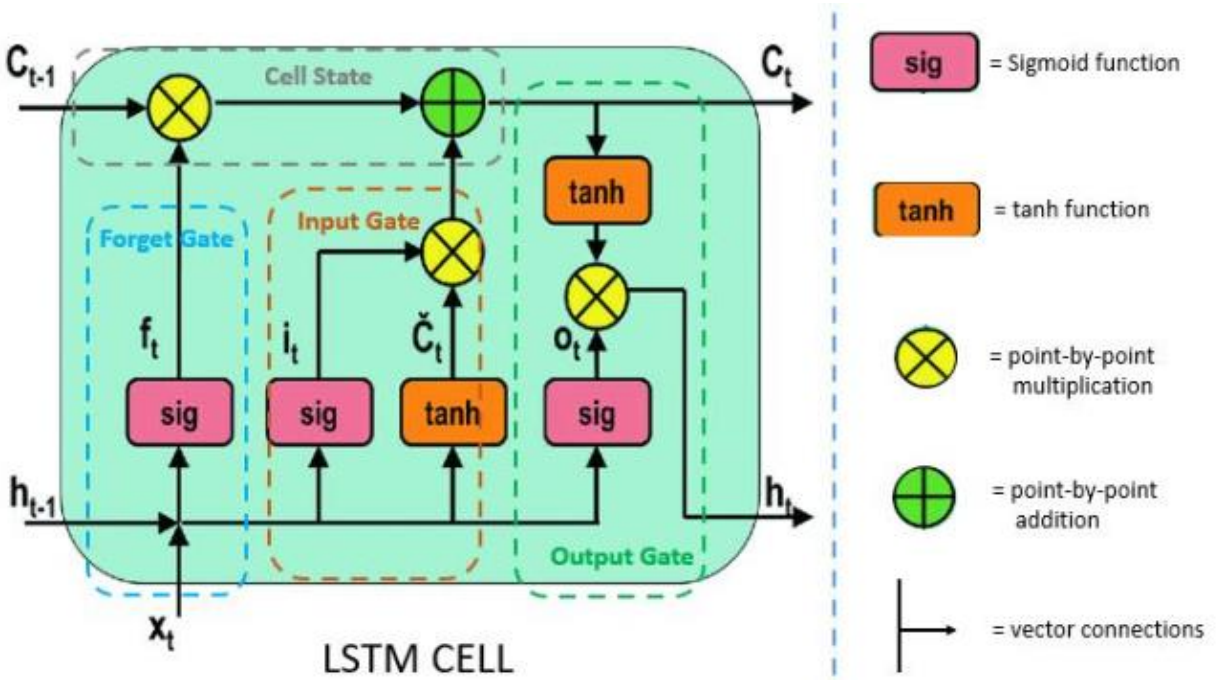

Figure 10: Long Short Term Memory (LSTM) network's structure

In Gates, there is sigmoid activation. The *tanh* activation function and the sigmoid activation function are comparable. The sigmoid function squishes data between 0 and 1, whereas *tanh* squishes values between -1 and 1. Because every integer multiplied by 0 equals 0, values are ignored or "forgotten." This is vital to consider for subsequent processing or forgetting data. When we multiple a number by one, the result is the same. As a result, the value remains the same or is "preserved." As a result, the network may learn which data to discard and which data to keep (to be kept).

Forget Gate: this gate makes the decision about which data should be discarded and which should be saved. The sigmoid function is used to combine the data from the previous hidden layer and the data from the current input. The numbers 0 and 1 are used to indicate the values. A value close to 0 indicates that the data should be ignored, while a value close to 1 indicates that the data should be maintained.

The Input Gate is responsible for updating the cell's state. In a sigmoid function, the previous hidden layer state and the current input are merged. By expressing the values in a range of 0 to 1, this is utilized to determine which values will be changed. If the value is 0, it means it is inconsequential, and if it is 1, it means it is extremely important. Use the *tanh* function with the hidden layer state and current to compress values between -1 and 1, which can aid with network control. The output of the sigmoid is then multiplied by the output of the *tanh*. The sigmoid output will determine which of the *tanh* output data should be kept.

Cell State: The cell state is multiplied by the forget vector at each position. It's possible that the value will be dropped in the cell state if it's multiplied by a value close to zero. The output of the input gate is then used to perform point-wise addition, which is then used to update the cell state to new values determined by the neural network. New cells are generated as a result.

The Output Gate specifies what the hidden state should be for the following hidden state. It's worth mentioning that the hidden layer state takes into account data from previous inputs. Predictions are also made on the hidden layer. The previous hidden layer state and the current input are supplied to the sigmoid function first. The newly altered cell state is then submitted to the *tanh* function. The *tanh* output is multiplied by the sigmoid output to determine what data should be conveyed by the hidden layer state. The concealed states are set as the output. A new cell state as well as a new concealed are introduced in the next time step. Throughout the feature extraction and classification process, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), notably Long Short Term Memory (LSTM), are used.

3.5 Summary

The system architecture employed comprises a video stream that was used as a collection of datasets, and these streams were used to extract critical information via the design. After the face

part was eliminated, the region of interest for the frames was extracted. We extracted a series of frames from the film using the OpenCV library. Based on the original input data, features from the images that contained all of the information required from the input data for the collection of datasets were obtained. The convolution layers and max pooling layer, which is placed in between the convolution layers, the dense layer, loss functions, and activation function, were used to extract features from datasets.

CHAPTER FOUR: EXPERIMENT

4.1 Introduction

The implementation of a phone-based and isolated word-based Visual Speech Recognition (VSR) system for the Amharic language is discussed in this chapter. The prototype's implementation tools, database, training, and testing methods will all be detailed.

The tests and experimental setup of our proposed Amharic visual speech recognition system are discussed in this chapter. The technique we utilized, the materials we used, the methods and procedures we took to conduct our experiment, and other experiment-related themes and difficulties are all presented here. In this chapter, we'll also talk about our dataset, which is a major topic. Datasets are well acknowledged to be the most important prerequisite in the era of Deep Neural Networks (DNN).

4.2 Data Collection

The type of dataset we're gathering is a video dataset, as previously stated. The video is a compilation of different speakers saying the sentence in Amharic. We collected the video dataset from the Ethiopian Satellite Television (ESAT) YouTube channel. The speaker sat about 2 (two) to 3 (three) meter away from the camera when the video was taken. As it is a news station they attempted to maintain a consistent lighting system across the space. The lighting system coordination would have an impact on the speaker's facial region and the consistency of our data.

We collected from 6 (six) speakers for our dataset. The videos are not of the same duration. The videos range in length from 2 (two) to 3 (three) seconds. We tried to keep the sampling consistent in terms of gender and mustache style. In the data collection process, 4 (four) female and 2 (two) male participants are involved as speakers. Our data mostly affect our model as it is more female speakers. After capturing and gathering the video data, we use a free video cutter and jointer to delete the undesirable portions of the video.

The most commonly spoken terms in the community are taken into account in our sample data gathering. As previously said, we have gathered 4 (four) phrase samples. Table 1 discusses the sentences that we used as an example.

Table 1: Sample phrases that the dataset has been collected

ID	Phrases	Written In English Letters	Their Meaning In English
1.	አብራችሁን ቆዩ	ABRACHEHUN QOYU	Stay with us
2	ከዜናዎቹ ጋር	KEZENAWOCHU GAR	With the recent news
3.	ጤና ይስጥልኝ	TENA YSTELGN	Wishing you great Health (Salutations)
4.	የሰዓቱን ዜናዎች	YSEATU ZENAWOCH	The news at this time

4.3 Experimental Setup

Different parameters, such as hardware features, dataset nature and size, single thread or simultaneous task processing, all affect execution time in experimental circumstances. Our tests were conducted on HP 15-BS E7440 laptop computer with an Intel Core i5 processor and 8 GB of RAM, as well as a Windows 10 operating system and the Google Collaborator cloud service. Data preprocessing, various data cleansing operations, and other duties such as document preparation have all been completed on the computer. Feature extraction and classification operations have been completed on the cloud service. We like to utilize Google Collaborator for feature extraction and classification activities since we can easily use any of the latest Python packages, and because we are working with video data, if our data exceeds the CPU's capacity, Google Collaborator can provide GPU service. All of Google Collaborator's services are available for free.

The most recent version of Python is employed in terms of software. The OpenCV library is an open-source image processing and computer vision toolkit that can be readily integrated with Python and used to conduct image processing and computer vision operations. We utilized different python packages in both situations (locally on the machine and in the cloud with Google collaborators). To name a few of the packages we've used:

Numpy is a Python programming package that provides array functionality. Pixels represent data points in an image, which is just a typical Numpy array. As a result, we can use fundamental Numpy operations like slicing, masking, and clever indexing to change the pixel values of an image. The image can be loaded with Skimage, and it can be displayed with Matplotlib.

The **OpenCV** library (Open Source Computer Vision Library) is a well-known computer vision library. The Python API for OpenCV is known as OpenCV-Python. OpenCV-Python is not only rapid, but also simple to create and deploy because the background code is written in C/C++ (due to the Python wrapper in the foreground). As a result, it's a fantastic choice for computer vision algorithms that require a lot of work.

Matplotlib: is primarily used for 2D visualizations, although it may also be used to handle images. Matplotlib is effective in altering images to extract information from them, despite not supporting all file types.

The Hyper Parameters We Used: We put in a lot of effort to find the best hyper-parameters for our experiment. It's impossible to remember all of the hyper-parameters we attempted. As a result, on the table form, we list the hyper parameters that we found to be optimal. We tried varying the number of hidden layers from one to five, but there was no noticeable difference after the third one for both CNN and RNN.

Table 2: Hyper-parameters used

The hyper-parameter	**Amount**	**Some of our trials**	**Selected as Optimum**
1st (Time distributed Conv2D)	Number of layers	32, 64, 128, 256	32
	kernel	(3,3), (5,5), (7,7)	(5,5)
	Activation function	tanh, Relu, Sigmoid	Relu
	Dropout	0.1, 0.2, 0.25, 0.3, 0.5	0.3
2nd (Time distributed Conv2D)	Number of layers	32, 64, 128, 256,	64
	kernel	(3,3), (5,5), (7,7)	(3,3)
	Activation function	tanh, Relu, Sigmoid	Relu
	Dropout	0.1, 0.2, 0.25, 0.3, 0.5	0.3
3rd (Time distributed Conv2D)	Number of layers	32, 64, 128, 256	128
	kernel	(3,3), (5,5), (7,7)	(7,7)
	Activation function	tanh, Relu, Sigmoid	Relu
	Dropout	0.1, 0.2, 0.25, 0.3, 0.5	0.3
1st (RNN (GRU, LSTM, BiLSTM))	Number of layers	64, 128, 256, 1024	128
	Activation function	tanh, Relu, Sigmoid	Relu

	Dropout	0.1, 0.2, 0.25, 0.3, 0.5	0.3
2nd(RNN(GRU, LSTM, BiLSTM))	Number of layers	64, 128, 256, 1024	256
	Activation function	tanh, Relu, Sigmoid	Relu
	Dropout	0.1, 0.2, 0.25, 0.3, 0.5	0.3
3rd (RNN (GRU, LSTM, BiLSTM))	Number of layers	64, 128, 256, 1024	1024
	Activation function	tanh, Relu, Sigmoid	Relu
	Dropout	0.1, 0.2, 0.25, 0.3, 0.5	0.3
Dense layer	Number of class	----------	4
	Activation function	Softmax, sigmoid	Softmax
Optimizer		sgd, adam, adadelta, rmsprop, adagrad	adam

4.4 Experiment Result

Our entire dataset was divided into three sections: training, validation, and testing. We used 70% of the complete dataset for training, and the remaining data was used for validation and testing the model. Three distinct categorization algorithms were used in the experiment. CNN-LSTM, CNN-GRU, and CNN-BiLSTM are the three. Each of these models comes with its own set of drawbacks and benefits. We'll talk about their outcomes and performance with each of the models in this section. Because many academics agree that Convolutional Neural Network (CNN) is preferable for feature extraction, all of the algorithms use Convolutional Neural Network (CNN) for feature extraction. We used Time-distributed layers instead of the standard Convolutional Neural Network (CNN) layer for the Convolutional Neural Network (CNN) layer.

4.4.1 CNN-LSTM

For feature extraction, we utilized Convolutional Neural Network (CNN), and for classification, we used conventional Long Short Term Memory (LSTM). The CNN-LSTM model has a better performance than CNN-GRU, but not as excellent as CNN-BiLSTM. The same is true when it comes to resource use. It uses more resources (both time and memory) than the Gated Recurrent Units (GRU) but less than the Bidirectional Long Short Term Memory (BiLSTM). Because Gated Recurrent Units (GRU) has two update and reset gates, it performs better than CNN-GRU. It can't tackle intricate problems like lip-reading because it only has two gates. Because it has three layers and is more complicated than the Gated Recurrent Units (GRU), it consumes more resources.

Another explanation for Long Short Term Memory (LSTM's) success is that it works better with long sequence data. We attained a training accuracy of 97.2% and a validation accuracy of 94.2%. One epoch takes 128 seconds to complete in terms of resource utilization. Figures 11 show a graphical representation of our CNN-LSTM model findings.

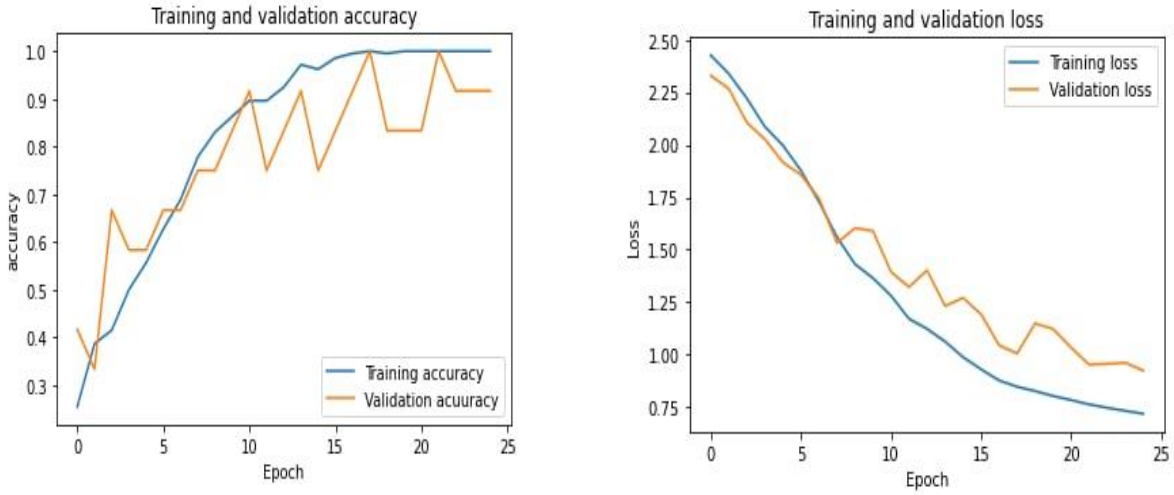

Figure 11: CNN-LSTM model training, validation accuracy, and validation lose

4.4.2 CNN-GRU

For feature extraction, we utilize Convolutional Neural Network (CNN), and for classification, we use Gated Recurrent Units (GRU). Of the three models, the CNN-GRU model performed the worst. It does, however, use the least amount of resources of the three models. Gated Recurrent Units (GRU) is the easiest to learn because it only has two gates (reset and update). Because of the nature of Gated Recurrent Units (GRU), the CNN-GRU model uses fewer resources. However, the Gated Recurrent Units (GRU's) nature precludes it from tackling sophisticated issues like lip-reading. Furthermore, the Gated Recurrent Units (GRU) is incapable of dealing with extended sequences. As a result, it received the lowest score of the three models. Figures 13 show a graphical representation of the CNN-GRU model's experimental results.

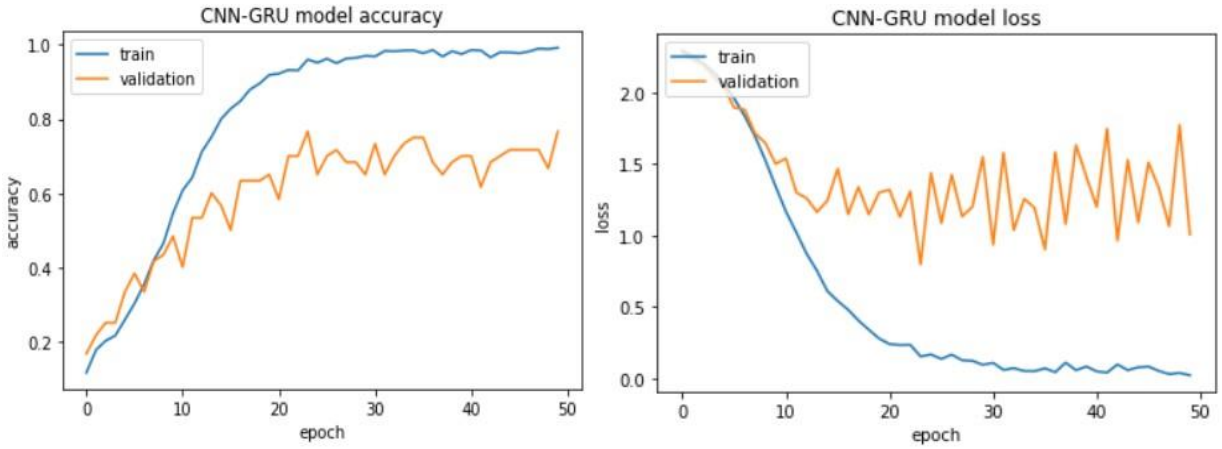

Figure 12: CNN-GRU training, validation accuracy, and Validation loss

4.4.3 CNN-BiLSTM

For feature extraction, we used Convolutional Neural Network (CNN), and for classification, we used Bidirectional Long Short Term Memory (BiLSTM). The CNN-BiLSTM model has the greatest performance of the three models we tested in the experiment. The main disadvantage of this model is that training takes a long time and requires a lot of resources. With this model, we were able to achieve 98.8% training accuracy and 97.6% validation accuracy. A single epoch takes 190 seconds to complete in terms of resource use. We got 92% when we test the model with our testing data. Back propagation is the major reason for this model's greater performance and higher resource consumption. In a given time frame, Bidirectional Long Short Term Memory (BiLSTM) analyzes the future as well as the prior value. It doubles the Long Short Term Memory (LSTM's) resources because it propagates in both ways. Bidirectional Long Short Term Memory (BiLSTM) will be the optimum option if the nature of the classification problem necessitates considering future input, such as sequence classification. To circumvent the constraints of a regular Recurrent Neural Network (RNN), (Mike Schuster and Kuldip K in 1977) proposed a Bidirectional Recurrent Neural Network (BRNN) that can be trained using all available input information in the past and future of a certain time frame. Figures 13 show a graphical representation of the CNN-BiLSTM model's experimental results.

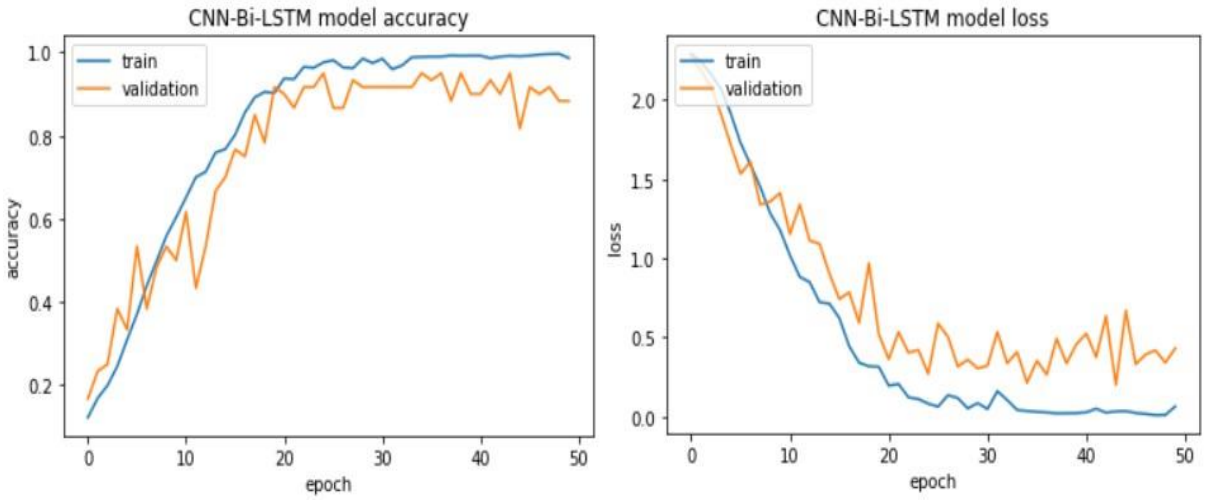

Figure 13: CNN-BiLSTM training, validation accuracy and and validation loss

4.4.4 Confusion Matrix and Classification Report for our Selected Model

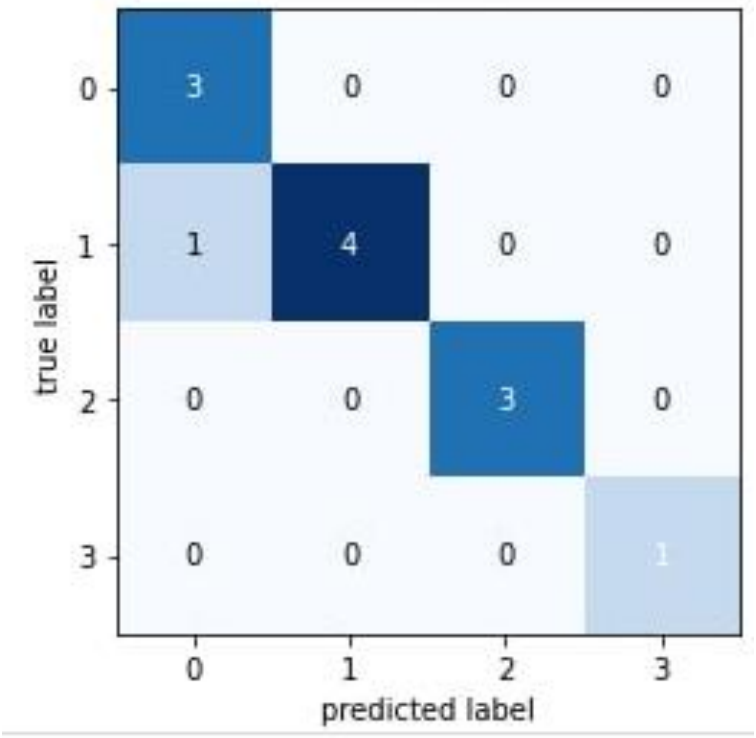

Figure 14: Confusion matrix for our Best model (CNN-BiLSTM)

```
                    precision    recall  f1-score   support

አብራችሁን ቆዩ.          0.75      1.00      0.86         3
ከዜናዎቹ ጋር           1.00      0.80      0.89         5
ጤና ይስጥልኝ            1.00      1.00      1.00         3
የሰዓቱን ዜናዎች         1.00      1.00      1.00         1

    accuracy                           0.92        12
   macro avg       0.94      0.95      0.94        12
weighted avg       0.94      0.92      0.92        12
```

```
print(predictions.argmax(axis=1))
print(y_valid.argmax(axis=1))
```

```
[0 3 1 2 2 0 1 2 0 0 1 1]
[0 3 1 2 2 0 1 2 1 0 1 1]
```

Figure 15: The classification report for our best model

4.4.5 Performance Comparison of other Related Works

Table 3: Performance comparison of our work with related works

Author	Year (GC)	Feature Extraction	Classification	Natural Language	Accuracy (%)
Amjad et al.	2001	Optical flow	Optical flow	Arabic	70
Alin G	2009	AMM	HMM	Dutch	46.6
Yannis M. et al.	2017	CNN	LSTM+CTC	English	95.2
Joon S. et al.	2018	CNN	LSTM	English	70
Befkadu B.	2017	DWT, LDA, DF	HMM, CHMM	Amharic	76.79
Muluken K.		Shape info, YIQ	ANN & SVM	Amharic	66.43
Current work	2022	CNN	BiLSTM	Amharic	92

4.5 Summary

We explained our methodology, dataset, material, and results in this chapter. We gathered and processed the dataset. The dataset was acquired with a high definition camera which is sated up in Ethiopian Satellite Television (ESAT) to take use of the camera's ability to capture high-quality

films. We collected data from 4 (four) female and 2 (two) male speakers. The camera is positioned 2 (two) to 3 (three) meter away from the speakers. Video cutter software Adobe Premium Pro removes unneeded parts of the video data after it has been collected. The remaining (relevant) section is then used to extract a frame sequence. The sequence of frames is subjected to various image preparation techniques.

We extracted a sequence of frames from each video. Padding has been used because the movies are of varying lengths. The maximum number of frames in this case is 20. Our videos, however, do not all have the same number of frames, If the number of frames in the video is less than the maximum number of frames, the last frames from the available frames are added to the sequence at the end. We erase frames in a constant interval if the number of frames in the movie exceeds the maximum number of frames. The retrieved frames are subsequently subjected to noise reduction. The extracted frames are subjected to face localization and Region of Interest (ROI) extraction procedures. This series of images is supplied to our deep learning model after all image preparation actions have been applied to the frames.

Our deep learning model uses Convolutional Neural Network (CNN) for feature extraction and a variety of Recurrent Neural Network (RNNs) for classification (Gated Recurrent Units (GRU), Long Short Term Memory (LSTM), and Bidirectional Long Short Term Memory (BiLSTM)). Convolutional Neural Network (CNN) is an excellent method for feature extraction, according to many studies. However, Convolutional Neural Network (CNN) isn't very excellent at classifying sequential data. Because Convolutional Neural Network (CNN) lacks memory, it is unable to handle the temporal aspect of sequential data. We used Recurrent Neural Network (RNN) for classification in order to solve this challenge.

We tested three different models: CNN-LSTM (Convolutional Neural Network (CNN) as feature extraction and Long Short Term Memory (LSTM) as classification), CNN-GRU (Convolutional Neural Network (CNN) as feature extraction and Gated Recurrent Units (GRU) as classification), and CNN-BiLSTM (CNN as feature extraction and Gated Recurrent Units (GRU) as classification) Convolutional Neural Network (CNN) as feature extraction and Bidirectional Long Short Term Memory (BiLSTM) as classification. CNN-BiLSTM outperforms the others out of all of these models. However, CNN-GRU outperforms the other in terms of speed and memory utilization. In both circumstances, CNN-LSTM serves as an intermediary.

CHAPTER FIVE: CONCLUSION AND FUTURE WORK

5.1 Conclusion

Understanding the movement of the speaker's lip to predict the speaker's words is known as Visual Speech Recognition (VSR). For humans, visual speech recognition is a difficult task. A trained person can recognize roughly 20% of the speech on average. When we are unable to listen to the speaker's voice or merge it with an audio file in order to better speech recognition tasks, we must recognize visual speech. The amount of video files available on the internet is constantly increasing. Some of the video files lacked synchronized audio files, or the audio files were distorted. We can't hear the sounds from some of the video clips because they were shot from afar. Furthermore, hearing-impaired people are unable to communicate with hearing people or with one another. To address these issues, visual speech recognition is used. Because Visual Speech Recognition (VSR) is such a challenging process for humans, it must be automated.

Computer vision is essential for automating the task of Visual Speech Recognition (VSR). Computer vision is an interdisciplinary field of science that investigates and automates tasks that the human visual system is capable of. Computer vision tasks include capturing, processing, analyzing, and comprehending visual data (images or video) as well as extracting high-dimensional data from real-world data. An image or a series of images can be used as a visual piece of information (video). We are discussing a video in this paper (sequence of images). The video data is first collected by the camera, and then the video file is broken down into a series of images. To train the machine, this sequence of photos is tagged and supplied to machine learning algorithms (in our case, Deep Neural Network (DNN)).

Various investigations on visual speech recognition have been conducted by a number of researchers. Machine learning algorithms had previously done a lot of work. Hidden Markov Model (HMM) was utilized for feature extraction while Support Vector Machine (SVM) was used for classification in the majority of earlier research. Deep Neural Networks (DNN) have lately showed outstanding image processing capability. We proposed this study utilizing a deep neural network to take advantage of the deep neural network's advantages.

We collect and arrange our datasets in this project. For 4 (four) selected Amharic phrases, we gathered the lip reading dataset. We use a high-definition to acquire our data. Our data is in the

form of a video. The video lasts between 1 (one) and 2 (two) seconds. We used video cutting software to delete unwanted scene of the video. After that, each video is turned into a frame sequence. Each frame was subjected to image preprocessing before being fed into the neural network. Some of the preprocessing that has been applied to the frames done includes denoising the images, extracting the facial part from each frame, and extracting the mouth region.

We employed a time-distributed Convolutional Neural Network (CNN) for feature extraction and three deep learning algorithms for classification after preprocessing our sequence of frames. We employed the classification algorithms Long Short Term Memory (LSTM), Bidirectional Long Short Term Memory (BiLSTM), and Gated Recurrent Unit (GRU) in the experiment. Bidirectional Long Short Term Memory (BiLSTM) outperforms the rest of these algorithms. Gated Recurrent Unit (GRU) outperforms the other methods in terms of calculation time. The fact that the Bidirectional Long Short Term Memory (BiLSTM) algorithm has two directions is the key reason for its high accuracy (forward and backward). Although the Long Short Term Memory (LSTM) method exceeds the Gated Recurrent Unit (GRU) algorithm in terms of performance, it takes longer to finish. We can see that the key advantage of bidirectional Long Short Term Memory (LSTM) is that it can manage a little more variance in video duration. We fill the shortest videos by copping the last frames of the video because the length of the videos varies. As a result, videos with fewer than the maximum number of frames are obliged to include certain frames that aren't relevant. Bidirectional Long Short Term Memory (BiLSTM) is more efficient than the other algorithms at handling this process.

In general, this work proposes a method for visual speech detection in Amharic. Front side perspectives are included in our dataset. Before training our model, we conducted different video and image preprocessing tasks to it, which resulted in good accuracy. We used three different models to conduct our investigation.

5.2 Contribution of the Study

We prepare the phrase level Amharic lip reading dataset for the Amharic language when we propose this study. This dataset may be valuable to other researchers that are conducting visual speech recognition research. The speaker's frontal views were captured and prepared for this dataset.

- This effort adds to scientific inquiry by providing a phrase-level Amharic dataset.
- This paradigm is used to communicate with hearing individuals by hearing impaired people.
- Apart from sign language.
- For videos captured by surveillance cameras.
- To integrate with a video recognition system in order to increase speech recognition performance.

To understand speech from distorted video files. Although the technology will not be able to recover the damaged audio, it can be utilized to comprehend it.

5.3 Future Works

We proposed a visual speech recognition model for the Amharic language in this research. We gathered the data, preprocessed it, and used it to train the system. We've been through a lot of ups and downs in order to complete these primary objectives. Despite the fact that we completed these duties, future work will fill in some gaps. Based on the restricted time and resources, we tried to achieve the goals that are proposed here in this paper. Our future investigations, or any other researchers who are interested in it, will finish the following work.

- All of our speakers had normal dental conditions when we collected our data. Different dental situations can arise in the actual world. In the actual world, some people have one, two, or all of their front teeth broken, or have their teeth replaced with alternative materials such as gold. We didn't think about teeth at all in our research. As a result, other than usual dental problems are expected to be included in future studies.
- Even though our accuracy is good, there is always room for improvement. Improvements can be achieved in future research by employing alternative strategies, such as using more data than we did previously, better picture preparation techniques, or other methods.
- In this work, we look at video files from two perspectives (front and 90^{o}). We propose considering more than two perspectives in future projects (like 30^{o}, 60^{o}, 45^{o}).
- This work is intended for phrase-level visual speeches, which are one to two-second videos, but it could be advanced to the sentence-level in future works.

Reference

[1] B. Rahim, S. and Naz, “Audio-Visual Speech Recognition Development Era; From Snakes to Neural Network:,” in *A Survey Based Study.*, 2011.

[2] H. A. M. Amjad Al-Ghanim, Nora Al-Oboud, Shatha Al-Tammami, “I See What You,” *Say Arab. Lip Read. Syst. Conf. Pap. Res.*, 2013.

[3] L. J. . R. Alin G. chitu, “Visual Speech Recognition: automatic system for lip reading of Dutch.,” 2009. [Online]. Available: Preliminary paper presented at an international conference.

[4] A. Z. Joon Son Chung, “Learning to Lip Read Words by Watching Videos. Visual Geometry Group,.”

[5] M. F. Font, “Multi-microphone signal processing for automatic speech recognition in meeting rooms.,” 2005.

[6] I. B. PALEN, “METHODS USED IN TEACHING LIP-READING TO SPEAKING CHILDREN. American Annals of the Deaf,” p. pp.190-197.

[7] J. Shaikh, A.A., Kumar, D.K., Yau, W.C., Azemin, M.C. and Gubbi, “2010, October. Lip reading using optical flow and support vector machines.,” *2010 3Rd Int. Congr. image signal Process.*, vol. (Vol. 1, p, 2010.

[8] Guenther, F.H., Hampson, M. and Johnson, D., “A theoretical investigation of reference frames for the planning of speech movements. Psychological review, 105(4),” p. p.611, 1998.

[9] Kelso, J.A., Saltzman, E.L. and Tuller, B., “The dynamical perspective on speech production: Data and theory.,” *J. Phonetics, 14(1),* p. pp.29-59, 1986.

[10] M. Tarigan, K.E., Ginting, F.Y.A. and Stevani, *Pyscholinguistic: Models of Speech Production and Lexical. Yayasan Kita Menulis*. 2020.

[11] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J. and Mashari, A., “Audio visual speech recognition,” in *(No. REP_WORK). IDIAP*, 2000.

[12] Q. Z. Jing Hong, Daniel Avery Nisbet, Alex Vlissidis, “Deep Learning Methods for Lipreading. The University of California, Berkeley Department of Electrical Engineering & Computer Sciences,” 2017.

[13] Napier, J. and Leeson, L., “Sign language in action. In Sign language in action Palgrave Macmillan, London,” 2016, p. (pp. 50-84).

[14] Befkadu B., “Audio-Visual Speech Recognition Using Lip Movement for the Amharic Language.,” Addis Ababa University., 2017.

[15] K. M. Thein, T. and San, “Features Point Extraction Based on Lip Movement for Lip Reading System.,” *Int. J. Inf. Technol. (IJIT)*, vol. 4(4).), 2018.

[16] F. M. S. Adriana Fernandez-Lopez, *Survey on automatic lip-reading in the era of deep learning. Department of Information and Communication Technologies, University*

Pompeu Fabra, Barcelona, Spain. Image and Vision Computing. 2018.

[17] R. B. Milan S., Vaclav H., "Image processing, analysis, and machine vision. Thomson.," 2008, p. ISBN 978-0-495-08252-1.

[18] T. J. B. Soltani A.A., Huang H., Wu J., Kulkarni T.D., "synthesizing 3D shapes via modeling Multi-view Depth maps and solhouttes with deep generative networks.," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, p. (pp. 1511-1519).

[19] DeLand, F. and Montague, H.A., "The Story of Lip-Reading; Its Genesis and Development.," 1968.

[20] Ahad, M.A.R., "Computer vision and action recognition:," in *A guide for image processing and computer vision community for action understanding (Vol. 5)*, 2011.

[21] M. S. Mayank Chauhan, "Study and analysis of different face detection techniques.," *Int. J. Comput. Sci. Inf. Technol.*, vol. vol.5 (2), pp. 1615-1618..

[22] M. R. R. Angali, Avinash Kumar, "(). An algorithm for face detection and feature extraction.," *Int. J. Sci. Eng. Technol. Res. (IJSETR)*, vol. volume3 is, 2014.

[23] D. Yu, "The Application of Manifold based Visual Speech Units for Visual Speech Recognition"," Thesis, Dublin City University, Dublin, Ireland.

[24] R. A. Kaucic Jr, "Lip Tracking for Audio-Visual Speech Recognition.," in *AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH.*, 1997.

[25] S. Khalid, S., Khalil, T. and Nasreen, "A survey of feature selection and feature extraction techniques in machine learning.," in *science and information conference*, 2014, p. (pp. 372-378). IEEE.

[26] R. H. Iain Matthews, Tim Cootes, J. Andrew Bangham, Stephen Cox, "'Extraction of Visual Features for Lipreading'. School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK Department of Medical Biophysics, University of Manchester, Manchester M13 9P. Regular paper.," 1999.

[27] I. Potamianos, G., Neti, C., Luettin, J. and Matthews, "Audio-visual automatic speech recognition: 22," in *An overview. Issues in visual and audio-visual speech processing*, 2004, p. p.23.

[28] A. W. Bregler, H. Hild, S. Manke, "" Improving connected letter recognition by lip-reading". In Proc. , volume 1," in *International Conference on Acoustics, Speech and Signal Processing*, 1993, p. pages 557–560,.

[29] Ahmad Basheer Hassanat, "Visual Words for Automatic LipReading. University of Buckingham United Kingdom.," 2009.

[30] M. S. and T. K. Tessema Mindaye, "The need for Amharic WordNet.," International conference of Ethiopian studies. Jorge Mason University.

[31] A. W. and R. Wynn, "Amharic Language and Culture Manual: National Language of Ethiopia.," T exas State University., 2011.

[32] B. G. and G. Eriksson, “Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL, pages,” in *Natural Language Processing at the School of Information Studies for Africa.,* 2005, pp. 49–56, Ann Arbor.

[33] B. T. Solomon Tefera, Wolfgang Menzel, “An Amharic speech corpus for large vocabulary continuous speech recognition. Fachberech Informatic,” Universitat Hamburg.

[34] S. H. S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, “‘Audio-Visual Speech Recognition Using Deep Bottleneck Features and High Performance Lipreading’ i, .,” in *n Proceedings of APSIPA Annual Summit and Conference, Asia-Pacific,* 2015.

[35] T. Mitchell, . “Machine Learning. McGraw Hill, New York.,” p. ISBN 0-07-042807 OCLC.

[36] C. Bennett, K.P. and Campbell, “Support vector machines: hype or hallelujah?. ACM SIGKDD explorations newsletter,” 2000, p. pp.1-13.

[37] J. H. . Friedman, “Data Mining and Statistics: What’s the connection?.,” in *Computing Science and Statistics.,* pp. 29 (1): 3–9.

[38] P. Bengio, Y. Courville, A. Vincent, *Representation Learning: A Review and New Perspectives.* 2013.

[39] B. B. and D. B. . Eric Petajan, “An improved automatic lip-reading system to enhance speech recognition,” 1988.

[40] L. Assael, Y.M., Shillingford, B., Whiteson, S. and De Freitas, N., “End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.,” 2016.

[41] Muluken B., “Automatic word recognition based on lip motion for Amharic speech: a computer vision approach. T,” hese work Bahir Dar University, Bahir Dar, Ethiopia.

[42] J. Dupont, S. and Luettin, “Audio-visual speech modeling for continuous speech recognition.,” in *IEEE transactions on multimedia, 2(3),* 2000, p. pp.141-151.

[43] Y. L. Jarrett, K. Kavukcuoglu, M. A. Ranzato, “‘What is the best multi-stage architecture for object recognition?’,” in *In International Conference on Computer Vision,* pp. 2146–2153. IEEE.

[44] Asratu Aemiro, “Audio-Visual Speech Recognition using LIP Movement for Amharic Language” 2015

[45] Solomon Berhanu, “Isolated Amharic Consonant-Vowel (CV) Syllable Recognition. An experiment using the Hidden Markov Model’, Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University,” 2001.

[46] Kinfe Tadesse, “Sub-word based Amharic speech recognizer: An experiment using Hidden Markov Model (HMM), MSc Thesis School of Information Studies for Africa, Addis Ababa University Ethiopia,.” 2002

[47] Zelalem Tamrie, “Amharic Language Visual Speech Recognition using Hybrid Features.,” *Abyssinia J. Sci. Technol.,* p. 6(2), pp.42-50., 2021.

[48] Befkadu Belete, “Audio-visual speech recognition using LIP movement for amharic language.,” *Int. J. Eng. Tech. Res.*, 2017.

[49] Muluken Birara, “Automatic word recognition based on lip motion for amharic speech: a computer vision aproach. Bahir Dar university.,” 2017.

[50] M. J. Paul Viola, “‘Rapid Object Detection using a Boosted Cascade of Simple Features,’” 2001.

[51] H. F. & C. Z. Linwei Fan, Fan Zhang, “Visual Computing for Industry, Biomedicine, and Art,” *Artic. number 7.*, vol. volume 2, 2019.

[52] D. K. L. K. + and Y. R. K. S. Suryanarayana, Dr. B.L. Deekshatulu, “().”Estimation and Removal of Gaussian Noise in Digital Images” ISSN 0974-2166 ,” *Int. J. Electron. Commun. Eng.*, vol. Volume 5, no. Number 1 (201), pp. 23–33, 2012.

[53] P. P. and S. Srivastava, “‘Image De-noising by Various Filters for Different Noise’.,” *Int. J. Comput. Appl.*, vol. Volume 9–, pp. 0975 – 8887, 2010.

[54] M. Kunaver, “‘Image feature extraction – an overview’.,” *EUROCON 2005.The Int. Conf.*, vol. Volume: 1, no. Computer as a Tool, 2005.

[55] G. Sarangi, Susanta; Sahidullah, Md; Saha, “‘Optimization of data-driven filterbank for automatic speaker verification’.,” in *Digital Signal Processing.*, p. 104.

[56] S. C. and N. M. F. (2016). “Feature E. of V. U. D. N. N. Iicc. Yoshihiro Hayakawa, Takanori Oonuma, Hideyuki Kobayashi, Akiko Takahashi, “CC16.”

[57] G. E. H. Alex Krizhevsky, Ilya Sutskever, “‘ImageNet Classification with Deep Convolutional Neural Networks’. Advances in Neural Information Processing Systems No Title,” 2012, p. 25, pp.l097, 1 105, 2.

[58] A. S. Manjunath Jogin, Mohana, Madhulika M S, Divya G D, Meghana R K, “‘Feature Extraction using Convolution eural Networks (CNN) and Deep Learning,’” in . *3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2018),*

[59] Vishali Aggarwa and Gagandeep, “‘A review:deep learning technique for image classification’.,” in *ACCENTS Transactions on Image Processing and Computer Vision, Vol 4(11),* p. ISSN (Online): 2455-4707.

[60] Y.-W. C. Weibin Wang, Dong Liang, Qingqing Chen, Yutaro Iwamoto, Xian-Hua Han, Qiaowei Zhang, Hongjie Hu, Lanfen Lin, “‘Medical Image Classification Using Deep Learning’.,” in *Deep Learning in Healthcare.*, p. pp 33-51.

[61] Y.-G. J. Zuxuan Wu, Ting Yawu, Yanwei Fu, “‘ Deep learning for video classification and captioning’.,”

[62] C. N. V. and E. C. Rafael Rego Drumond, Bruno A. Dorta Marques, “In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) - Volume 1: GRAPP,” in *“An LSTM Recurrent Network for Motion Classification from Sparse Data,”* 2018, p. pages 215-222.

[63] J. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., ´ Bunke, H., and Schmidhuber, "'A novel connectionist system for unconstrained handwriting recognition'. ," in *IEEE transactions on pattern analysis and machine intelligence*, 2009, pp. 31(5):855–868.

[64] Hochreiter, S. and Schmidhuber, J., "997. Long short-term memory.," in *Neural computation, 9(8),* 1997, p. pp.1735-1780.

Appendix A: OpenCV, the Face and Lip Identification.

```
import cv2
import numpy as np
import os

faceDetect = cv2.CascadeClassifier('D:/opencv-
4.x/data/haarcascades_cuda/haarcascade_frontalface_default.xml')
mercy = 1

for i in range (100):
    try:
        eyu = 'y' + str(mercy) + '.mp4'
        print(eyu)
        Newfolder = 'C:/Users/hp/OneDrive/Desktop/Dagmawi/dataset/Dagmawi-' + str
(i)
        os.makedirs(Newfolder)

        cam = cv2.VideoCapture("C:/Users/hp/OneDrive/Desktop/ Dagmawi/የስኡቱን
ዜናዎች/" + str(eyu))
        id = 0;
        mercy = mercy + 1

        while(True):
            ret, img = cam.read()
            #gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
            faces = faceDetect.detectMultiScale(img, 1.05, 5)
            for(x,y,w,h) in faces:
                #cv2.rectangle(img, (x,y), (x+w,y+h), (0,255,255),3)
                cv2.imwrite(Newfolder+"/_"+str(id)+".jpg",img[((y//2)+((y+h)//2))
:y+h, x:x+w])
                id = id+1;
                cv2.rectangle(img, (x,((y//2)+((y+h)//2))), (x+w,y+h),
(255,255,255),3)
            cv2.imshow('face',img)
            if(cv2.waitKey(1) == ord('s')):
                break
        cam.release()
        cv2.destroyAllWindows()
```

Appendix B: Sample Code

```
from google.colab import drive
drive.mount('/content/drive')

import cv2
import tensorboard
import tensorflow as tf
import keras
from tqdm import tqdm
import numpy as np
from random import shuffle
import time
from tensorflow.keras.callbacks import TensorBoard
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

def create_dataset():
  dataset = []
  images = []
  limit = 0
  count = 0

  for frames in tqdm(os.listdir(r"/content/drive/MyDrive/New Dataset/")):
    path = os.path.join(r"/content/drive/MyDrive/New Dataset/", frames)
    img = cv2.resize(cv2.imread(path), (IMG_SIZE, IMG_SIZE))

    images.append(np.array(img))
    limit += 1
    count += 1
    if limit == NUM_FRAMES:
      limit = 0
      if (count < 1430):
        dataset.append(np.array([images, np.array([1, 0, 0, 0])]))
      elif ((count >= 1430) and (count < 2860)) :
        dataset.append(np.array([images, np.array([0, 1, 0, 0])]))
      elif ((count >= 2860) and (count < 4290)) :
        dataset.append(np.array([images, np.array([0, 0, 1, 0])]))
      elif ((count >= 4290) and (count < 5720)) :
        dataset.append(np.array([images, np.array([0, 0, 0, 1])]))
      images = []

  shuffle(dataset)
  np.save("ሊፕ_ዳታሴት.npy", dataset)
  return dataset
```

```
from tensorflow.keras.layers import LSTM, GRU
from tensorflow.keras.layers import Bidirectional
from tensorflow.keras.layers import Conv2D
from tensorflow.keras.layers import Dense
from tensorflow.keras.layers import Flatten
from tensorflow.keras.layers import MaxPooling2D
from tensorflow.keras.layers import TimeDistributed
from tensorflow.keras.layers import Reshape
from tensorflow.keras.layers import MaxPooling3D
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import BatchNormalization

model = Sequential()
model.add(Conv2D(64, (3, 3), activation = "relu", input_shape = (22, IMG_SIZE,
IMG_SIZE, 3), padding="same"))
model.add(Conv2D(64, (3, 3), activation = "relu"))
model.add(MaxPooling3D((1,2,2)))
model.add(Conv2D(64, (3, 3), activation = "relu"))
model.add(MaxPooling3D((1,2,2)))
model.add(Conv2D(64, (3, 3), activation = "relu"))
model.add(MaxPooling3D((1,2,2)))

# 43264
#model.summary()
model.add(Reshape((22, 30976)))

#BiLSTM layers
lstm_fw = LSTM(units = 32)
lstm_bw = LSTM(units = 32, go_backwards = True)
model.add(Bidirectional(lstm_fw, backward_layer = lstm_bw))

# Dense layers
model.add(Dense(64, activation = "relu", kernel_regularizer =
keras.regularizers.l2(0.01)))
model.add(Dense(32, activation = "relu", kernel_regularizer =
keras.regularizers.l2(0.01)))
model.add(Dense(4, activation = "softmax"))
model.summary()
labelNames= ["አብራችሁን ቆዩ", " ከዘናዎቹ ጋር", "ጤና ይስጥልኝ", "የሰዓቱን ዜናዎች"]
from mlxtend.plotting import plot_confusion_matrix
from sklearn.metrics import classification_report
predictions = model.predict(x_valid, batch_size=BATCH_SIZE)
print(classification_report(y_valid.argmax(axis=1), predictions.argmax(axis=1),
target_names=labelNames))
```

Appendix C: Dataset Sample