Audiovisual Speech Recognition using Multi-View Lip Movement for Amharic Language using Deep Learning

Wondimagegn Tamene

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

May,2022

Adama ,Ethiopia

Audiovisual Speech Recognition using Multi-View Lip Movement for Amharic Language using Deep Learning

Wondimagegn Tamene

Advisor : Mesfin Abebe (Phd)

A Thesis Submitted to

The Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfilment of the Requirement for the Degree of Masters in computer science and Engineering

Office of Graduate Studies

Adama Scinece and Technology University

May, 2022

Adama, Ethiopia

Declaration

I hereby declare that this MSC thesis is my original work and has not been presented as a partial degree requirement for a degree in anu other university and that all sources of materials used for the thesis have been duly acknowleged.

Name: Wondimagegn Tamene

Signature:_____

This MSC thesis has been submitted for examination with my approval as a thesis advisor.

Name: Mesfin Abebe (Phd)

Signature:_____

Submission Date:_____

Recommendation

I have the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled “Audiovisual Speech Recognition using Multi-View Lip Movement For Amharic Language using Deep Learning” prepared under my guidance by Wondimagegn Tamene submitted in partial fulfillment of the requirements for the degree of Master’s of Science in Computer Science and Engineering. Therefore, I recommend the submission of revised version of the thesis to the department following the applicable procedures.

_____	_____	_____
Advisor	Signature	Date

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by Wondimagegn Tamene have read and evaluated his thesis entitled "**Audiovisual Speech Recognition using Multi-View Lip Movement For Amharic Language using Deep Learning"** and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Computer Science and Engineering.

_____	_____	_____
Supervisor/Advisor	Signature	Date
_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
_____	_____	_____
External Examiner	Signature	Date

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the God, the Almighty, for His shows of blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor Dr. Mesfin Abebe for providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance. I am extremely grateful for what he has offered me.

I am extremely grateful to my family for their love, prayers, caring and sacrifices for educating and preparing me for my future and complete this research work.

Dedication

This research paper is dedicated to my wife and children, who has been nicely my supporter until my research was fully finished, and encouraged me attentively with their fullest and truest attention to accomplish my work with truthful self-confidence.

List of Acronym and Abbreviation

AAVC	Amharic Audio-Visual data corpus
ANN	Artificial Neural Network
ASR	Automatic speech recognition
AVSR	Audio-visual speech recognition
BLSTM	Bidirectional Long Short-Term Memory
CHHM	Coupled Hidden Markov Model
CNN	Convolutional neural networks
CV	Consonant-Vowel
DFT	Discrete Fourier transform
DNN	Deep Neural Networks
DWT	Discrete wavelet transform
GMM	Gaussian mixture model
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
KNN	K-nearest neighbors
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
LVCSR	Large vocabulary continuous speech recognizer
MFCC	Mel-frequency cepstral coefficient

PCA	Principal Component Analysis
RNN	Recurrent neural network
ROI	Region of Interest
SSD	Shot Multibox Detector
SWT	Stationary Wavelet Transform
VSR	Visual speech recognition

Table of Contents

List of Tables

List of Figure

Abstract

Speech is the most common form of interpersonal communication. The use of computers to automatically transcribe natural speech to enable human-machine interaction is known as automatic speech recognition. In order to achieve effective human-machine interactions, noise-robust voice recognition becomes essential. The performance of speech recognition systems can be improved by recognizing lip movements and describing their correlations with speech sounds, especially when operating in noisy environments.

There are few researches done on Amharic speech recognition. Most of them are done on Automatic Speech Recognition and Visual Speech Recognition separately. Only Befikadu Belete conducted a research on Audion Visual Speech Recognition for Amharic Language(Belete, 2017). However, Befikadu Belete uses only frontal view to conduct his research. Which doesn't applicable in occasions where side view only is available.

The aim of this study is to develop audio-visual speech recognition for Amharic Language using multi-view lip movement. By combining possible view angles, this study focuses on using bidirectional LSTM RNN to learn representation from audio data and CNN plus bidirectional LSTM RNN to extract sequential features for visual part and combining them in fusion part using multi-modal RNN to learn automatically from training data.

In order to develop the proposed approach, The data is collected from YouTube and other Broad casting website including dramas and programs engaged in news presented using single and interviews with multiple peoples using Amharic language. The dataset contains 50 words with 5000 utterances which is divided into training and testing with the training stack receiving around 80% of the data, the testing stack receiving around 20% of the data.

Finally, the result demonstrate that using multi-view enhance accuracy of word recognition in audio visual speech recognition of Amharic Language. Combining Frontal view with left three quarter demonstrate higher accuracy than others with accuracy of 97.5%.

Keywords: Amharic; Lip-reading; Audio-Visual Speech Recognition; Multi-Modal LSTM

1 Introduction

1.1 Background of the Study

Speech is the most prevalent way for people to communicate with one another (Mehrabani, Bangalore, & Stern, 2015). For various reasons, including scientific curiosity about the mechanics underlying the mechanical realization of human speech skills and the desire to automate basic tasks that fundamentally need human-machine interactions, research in automatic speech recognition is being conducted. It is a field that has gotten a lot of attention during the last five decades(Patil et al., 2019; Vadwala, Suthar, Karmakar, Pandya, & Patel, 2017).

Speech recognition software that works automatically (ASR) is a key component of future human-computer interfaces that aim to enable natural pervasive and ubiquitous computing through, among other things, speech (Vadwala et al., 2017). It is the use of computers to transcribe natural speech automatically to enable human-machine interaction. Many devices today have a speech interface having ASR technology. This technology enables humans to communicate with a computer interface using their voices in a manner that is similar to normal human conversation. However, noise that occurs in a real environment is the crucial problem that degrades the recognition performance of ASR (Virtanen, Singh, & Raj, 2012).

Noise-robust speech recognition becomes crucial in attaining effective human machine interactions. Audio-visual Speech Recognition also known as bimodal or multi-modal speech recognition minimizes this degradation by incorporating visual features, such as speaker's lip movements and facial expressions, into automatic speech recognition (ASR) systems has shown improvements in performances especially under noisy conditions.

Lip reading is the ability to recognize what is being said from visual information alone (Chung, Senior, Vinyals, & Zisserman, 2017). Because the visual signal is unaffected by acoustic noise, lip reading has the potential to improve acoustic speech recognition in loud circumstances. It relies on visually analyzing the movement of the lips, face, and tongue. Lip reading can be done by using different methods. However, lip reading using neural networks

yields better accuracy than conventional methods (Wand, Koutník, & Schmidhuber, 2016). Lip reading plays an important role in a wide variety of speech recognition tasks such as dictating messages in noisy environment, transcribing archival silent films, assisting people with speech impairments.

This study focus on designing and developing Audio-visual speech recognition using multi-view lip movement for Amharic Language, by extending the work done on audio-visual speech recognition for Amharic language by using frontal view only, using bidirectional LSTM RNN to learn representation form audio data and CNN plus bidirectional LSTM RNN to extract sequential features for visual part and combining them in fusion part using multimodal RNN to learn automatically from training data.

1.2 Motivation of the Study

The motivation for the work in this study is inspired by previous work done by B. Belete on Audio-Visual Speech Recognition Using Lip Movement for Amharic Language(Belete, 2017). He used frontal face lip reading for his audio visual speech recognition system. The premise of this thesis is that instead of using single view for lip reading multiple views increases the accuracy of the AVSR.

Another motivation is the advancement of audio visual speech recognition systems. Deep learning has emerged as an effective method for audiovisual speech recognition systems (Bhaskar & Thasleema, 2019). In deep learning without pre-assumed rules, the feature extraction in the front-end and the back-end feature recognition is learned from the training data.

Thus, this study deal in the combination of frontal view and side view to develop audio visual speech recognition system for Amharic language by combining possible view angle and focuses on using bidirectional LSTM RNN to learn representation form audio data and CNN plus bidirectional LSTM RNN to extract sequential features for visual part and combining them in fusion part using multi-modal RNN to learn automatically from training data.

1.3 Statement of the Problem

Automatic speech recognition is a technology that translates voice into text transcriptions and is a key tool in man to machine communications (X. Lu, Li, & Fujimoto, 2020). In various gadgets, automatic speech recognition (ASR) is frequently employed as an effective interface for Computers, mobile phones, robotics, and car navigation systems. In the low noise level environments, audio-only ASR system is obtained high recognition accuracy and it has been put to practical use.

There are few researches done on Amharic speech recognition. For example, M. Melese et. al created a 7.43-hour read voice corpus in the tourist domain(Melese, Besacier, & Meshesha, 2016). M. Y. Tachbelie et.al investigated the use of syllables for Amharic speech recognition (Tachbelie, Abate, & Besacier, 2014) and presents the application of morpheme-based and factored language models in an Amharic speech recognition task.

B. Belete conducted research on audio visual speech recognition for Amharic language employing lip movement and frontal face for visual speech recognition (Belete, 2017). However, there are different occasion where frontal lip reading is not possible like in occasion in which side view only is possible. Zimmermann et.al explored the influence of multi-view on visual speech recognition by combining possible view angle both at future levels and decision level(Zimmermann, Ghazi, Ekenel, & Thiran, 2017). In consequence the result gained from the different view improves the correctness and performance significantly than frontal-view only.

Even though, researches have been done on Audio-visual speech recognition using lip movement extracted from multiple view images for English language, it has not been carried out for Amharic language or for any other Ethiopian local language.

Therefore, the aim of this study is to design and develop Audio-visual Amharic speech recognition using lip reading from multiple viewpoints.

1.3.1 Research Question

The following are the research questions that have to be addressed at the end of this study:

- To what extent profile view improve performance of Amharic speech recognition?
- What is the effect of using multi-view over profile and front view only AVSR?

1.4 Objective of the Study

1.4.1 General Objective

The general objective of this study is to develop audio-visual speech recognition model by using multi-view lip reading for Amharic language.

1.4.2 Specific Objectives

The following are the specific objectives to achieve the general objective of the study

- To perform a comprehensive literature review related to AVSR
- Collect audio visual data to train and test
- Extract audio and visual speech features.
- To build a model and integrate the developed ASR models.
- To test the performance of the developed model.
- To analyze the results and give a conclusion and forward recommendations.

1.5 Scope and Limitation of the Study

This study focus on audio-visual speech recognition using lip movement taken from multiple viewpoints by combining possible view angles because studies showed that result gained from different view improve the correctness and performance of AVSR system significantly than frontal-view only. The data of the study is limited to video data distilled from You Tube videos like TV broadcasts, films and others. This study focuses on words of Amharic language only. Due to time constraint, it does not consider continuous speech, facial expressions, hand and body gestures.

1.6 Application of Results

Lip reading plays an important role in a wide-ranging variety of speech recognition tasks (Butt & Lombardi, 2013). The major significance of this study is designing and developing Amharic Audio-Visual Speech Recognition and the findings of this work can be used

- In 'dictating' instructions or messages to a phone in a noisy environment
- For transcribing and re-dubbing archival silent films
- To assist people with speech impairments
- For resolving multi-talker simultaneous speech and
- To resolve incomplete or unavailable acoustic information.

1.7 Thesis Organization

This thesis is organized in six chapters and the contents of each chapter are described as follows. The first chapter introduces the whole paper and is consisted of the subtopics: background of the study, motivation of the study, statement of the problems and its justification, objectives of the research, scope and limitation of the study and finally application of the study.

The second chapter, which is literature review and related works, discusses the different underlying principles and theories in speech terminologies. It starts by discussing the speech and speech recognition and then discusses lip reading, audio-visual speech recognition and deep learning. Continuing, this chapter presents the speech recognition in Ethiopia and finally, previous related works in the area of this particular research work are presented.

In the third chapter, the methodology and the tools and techniques used in the development of the multi-view audio-visual speech recognition are discussed in detail.

The fourth chapter discusses about the proposed model starting from its architecture followed by discussing step by step procedures.

The fifth chapter is all about the experimentation and evaluation carried out on the developed models. And therefore, in this chapter, the researcher tried to show the results of the experimentations in line with a discussion on the results obtained from the developed model.

The last chapter, which is chapter six, presents the conclusions arrived in by the researcher and also the recommendations that need to be considered in future works in the area.

2 Literature Review and Related Works

2.1 Overview

This chapter details words and terminologies used in this paper. Section2.2 defines speech and speech recognition. Section 2.3 details the different types of lip reading and the challenges of lip reading in speech recognition. Section 2.4 describes audio visual speech recognition, its parts and challenges in its implementation. Section 2.5 describes the different deep learning methods used in audio visual speech recognition. Section 2.6 details the different related works done on speech recognition and the last section summarize and conclude on which gap this paper focus and how it tends to solve the problem indicated.

2.2 Speech and Speech Recognition

2.2.1 Speech

Speech is the most efficient and preferred communication means of exchanging information for human beings (Wuth, Correa, Núñez, Saavedra, & Yoma, 2021). It is the simplest and natural way of communicating with others and expressing ourselves. It can be represented as acoustic and visual information. In general humans understand speech by processing acoustic information. However, it is difficult using acoustic information for deaf persons and in noisy environments(Han, Kang, & Yoo, 2017; Sukale, Gornale, & Yannawar, 2016).

Researchers use speech to communicate with machines over the last five decades. Therefore, designing and production of speech recognition system is the target of research centers. Many researchers use acoustic information for speech recognition system (Saksamudre, Shrishrimal, & Deshmukh, 2015). The performance of this acoustic information reached 100% for closed environments.

However, in noisy environments it is difficult to use acoustic information for speech communication in areas like aviation, manufacturing and construction. Such noisy environments require visual information which is helpful to communicate instructions and commands.

2.2.2 Speech Recognition

Speech recognition is the process of converting speech signal to words using Algorithm in computer program. Speech recognition incorporate different fields like pattern recognition, artificial intelligence, machine learning and computer vision. Figure 2.1 shows the taxonomy required in speech recognition process(Vadwala et al., 2017).

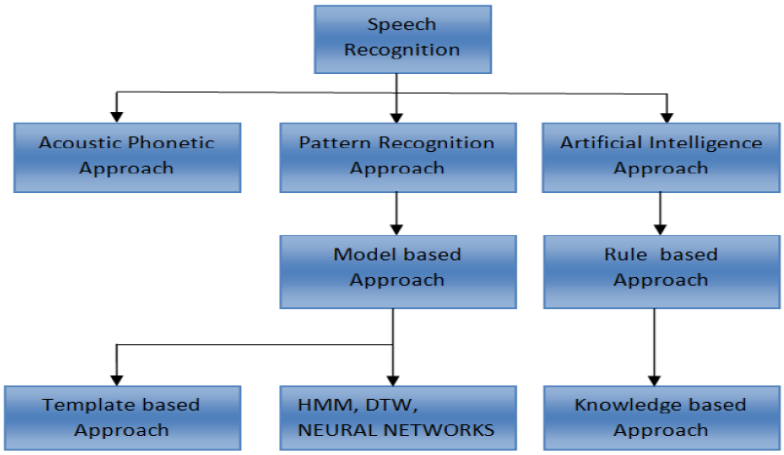

Figure 2.1Taxonomy of speech recognition

In the fields like telephony, human computer interface and artificial intelligence automatic speech recognition is an assistive tool as an alternative technique for individuals with disabilities(El Maghraby, Gody, & Farouk, 2016).Speech recognition process include different stages starting from analysis followed by feature extraction and then classification. Figure 2 shows the process speech recognition system(Vadwala et al., 2017) should follow.

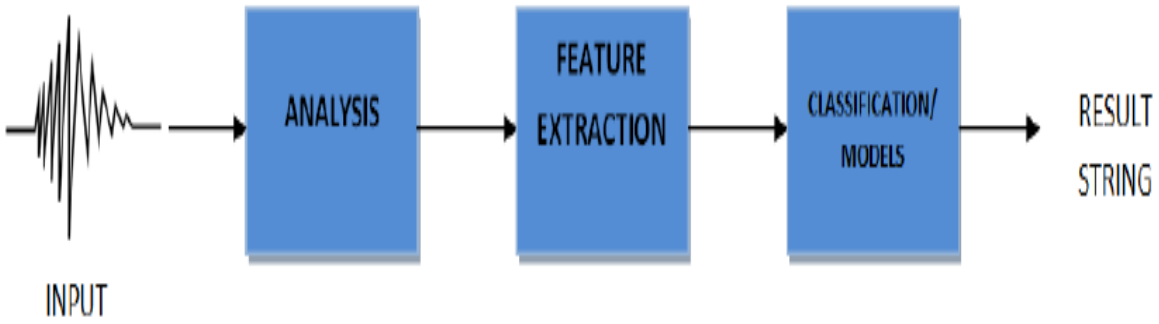

Figure 2.2Process of speech recognition

2.3 Lip Reading

Lip reading, also referred to as visual speech recognition, is the method of recognizing speech by observing only the lip movements without having access to the audio signal. It is the field where the computer vision and speech recognition communities meet each other and combine the advances of each field. It can be achieved with a processing pipeline primarily based on neural networks, yielding considerably better accuracy than conventional methods (Chung & Zisserman, 2017; Chung & Zisserman, 2018; Stafylakis & Tzimiropoulos, 2017).

Lip reading, or the ability to understand speech using only visual cues, is a very impressive skill that plays an important role in human communication and speech comprehension (Chung & Zisserman, 2016). It plays a vital part in human communication and speech understanding. It has obvious applications in speech transcription where audio is unavailable, such as for archival silent films or (less ethically) off-the-mike exchanges between politicians or celebrities (the visual equivalent of open-mike mistakes) (Chung & Zisserman, 2018). It is additionally complementary to the audio understanding of speech, and indeed can adversely affect perception if audio and lip motion are not consistent for such reasons, lip-reading has been the subject of a vast research effort over the previous few decades (Patil et al., 2019) , (Petridis et al., 2018).

Lip reading examples include word recognizing in continuous speech, phrase recognition, and sentence level transcribing of continuous speech(Barkhan, Alizadeh, & Maihami, 2019).

A. Word Recognition

An isolated word speech recognition system necessitates that the speaker provides a brief intermission between words. It doesn't denote that it recognizes single words, but does entail a single utterance at a time. This is acceptable in situations where the user is required to provide only one word responses or commands; however, it is extremely unusual for multiple word inputs (Vadwala et al., 2017).

B. Phrase Recognition

Connected word systems (or, more accurately, 'connected utterances') are analogous to isolated words; however, they allow separate utterances to be run-together with a nominal pause between them.

C. Sentence Recognition

Recognition of Sentences Users can use continuous speech recognizers to speak roughly naturally while the computer finishes the content. It includes an immense pact of "co-articulation", where adjoining words are spoken together without temporary halts or any other noticeable division between words. Continuous speech recognition systems are extremely difficult to develop because they must use extraordinary methods to determine utterance boundaries as the list of vocabulary grows and the confusability between different word sequences grows.

A VSR system includes image reception, lip identification, feature extraction, and speech recognition. Its precision is highly dependent on the identification of the lips and its limits(Patil et al., 2019).

Lip reading can be done from frontal view, side view and multi-view across different views.

1. Frontal View

In lip reading frontal or near-frontal faces are considered due to two reasons: first availability: most video material has mainly near-frontal faces; and second technological: until recently,

profile face detectors and profile landmark detectors were far inferior to their frontal counterparts(Barkhan et al., 2019).

2. Side View

The vast majority of previous works has focused on frontal view lip-reading. This is in contrast to evidence in the literature that human lip readers prefer non-frontal views, most likely due to more pronounced lip protrusion and lip rounding. It is thus reasonable to assume that non frontal lip views contain useful information that can be used to improve frontal view lip reading performance or in cases where a frontal view of the mouth roi is not available. This is especially true in realistic in-the-wild scenarios where the face is rarely frontal(Chung & Zisserman, 2018).

According to Chung et al it is possible to read lips in profile but the standard is inferior to frontal face lip reading. They proposed a multi-stage strategy to automatically collect multi-view lip reading dataset using CNN face detector based on single shot multi box detector and aligning the visual face sequence and the words in subtitle in two stages: aligning audio to text and video to audio(Vadwala et al., 2017). However, according to Anina et al. Recognition results show that the best VSR performance does not come from the frontal view or those close-to-frontal views(Huang & Kingsbury, 2013).

Zimmermann et al. have explored the influence of multi-view fusion on visual speech recognition using OuluVS2 dataset. In the results they have shown that exploiting multiple-view data can improve the recognition results significantly particularly for combinations involving the frontal view, the 45° and 90° view angles. The other 30°, 60° views do provide additional information, however, the improvements are not as noteworthy(Zimmermann et al., 2017).

Furthermore, Lan et al. showed that frontal lip reading is not the optimal view angle instead 30° performed best by focusing in multiple camera views using LiLIR dataset simultaneously recorded from 0°,30°,45°,60° and 90°. they tested the lip-reading system in a view-dependent and view-independent conditions (Paleček & Chaloupka, 2013).

- **Multi-View**

ElMaghraby et al. Proposed multi task based approach for multi-view training on the OuluVS2

dataset which contains l52 speakers saying three kinds of utterances 3 times each. Their method is composed of CNN and bidirectional LSTM. They concluded that their model achieved 95.0% accuracy(ElMaghraby, Gody, & Farouk, 2020).

According to Petridis et al., the presented an end-to-end model to jointly learn features directly from pixels and performs visual speech classification using BLSTM networks from multiple views by using OuluVS2 database with 5 lip views between 0 and 90. It has 52 speakers who say each utterance three times, for a total of 156 examples per utterance. The utterances are the following: "excuse me", "goodbye", "hello", "how are you", "nice to meet you", "see you", "i am sorry", "thank you", "have a good time", "you are welcome". They demonstrated that the combination of the frontal and profile views leads to significant improvement over the frontal view only(Chung & Zisserman, 2018).

Lip reading models can further be grouped as speaker dependent or speaker independent.

1. **Speaker Dependent**

Speaker-independent systems are being developed for a variety of speakers with varying articulations. It recognizes the speech patterns of a large group of people. In speaker dependent a user must provide samples of his or her speech before using them, a speaker dependent system is supposed for the use of a single speaker.

2. **Speaker Independent**

Speaker independent systems are planned for range of speakers having different articulations. It identifies the speech patterns of a huge collection of people. This system is most tricky to develop; most steep and provides a reduced amount of accuracy than speaker dependent systems.

In noisy environments, it is difficult for human to understand each other using audio-conversation. Hence, lip-reading fills the gap by helping humans to achieve better communication. However, in the absence of context, lip reading is difficult task for humans(Patil et al., 2019).

2.3.1 Limitations of lip reading

One of the fundamental limitation of lip-reading is homophones, words that sound different but involve similar lip movements, that cannot be identified using visual information alone degrade the performance of visual information for example in English the phonemes ‘B’, ‘P’ and ‘M’ are visually identical and words like Bark, Mark and Park are homophones and cannot be identified by using lip reading only (Chung & Zisserman, 2018).

The other limitation of lip reading is that it is a challenging problem in any case due to adversarial imaging conditions (such as poor lighting, strong shadows, motion, resolution, foreshortening, etc.) and intra-class variations (such as accents, speed of speaking, mumbling)(Chung & Zisserman, 2018).

2.4 Audiovisual Speech Recognition (AVSR)

Audiovisual speech recognition (AVSR) is the problem of recognizing speech by combining audio and video information where audio technique rely on reception of audio signals and visual method involve image processing, artificial intelligence, object detection, pattern recognition and so on (Petridis et al., 2018), (Barkhan et al., 2019).

Because visual data provides a stream of information that is separate from the audio and invariant to acoustic noise, AVSR has shown noticeable advances over audio-only speech recognition in both clean and noisy conditions(Bhaskar & Thasleema, 2019).The major goal of AVSR is to complement corrupted audio speech inputs using visual information derived from a speaker's lip motion(Salama, El-Khoribi, & Shoman, 2014).

2.4.1 Feature Fusion

The fusion of audio and visual options is employed to require advantage of the visual and audio option within the recognition process.

There are different types of fusion exists, but generally two levels of fusion are performed that is “Feature level Fusion‟ or “early‟ fusion and Decision level fusion‟ or ”late‟ fusion(Sukale et al., 2016).

2.4.2 Feature Level Fusion

Feature Level Fusion is also known as early fusion. Before entering a recognition network, early fusion is used. It converts raw data into a more manageable intermediate format. Feature level fusion is employed to obtain a combined feature vector that's passed to classifier to perform a recognition process(El Maghraby et al., 2016).

2.4.3 Decision Level Fusion

The decision level integration is also known as late integration in which two classifiers are used for classification of each modality. Late fusion is the fusion after recognition. i.e., In late integration each modality is first classified separately and final classification is based on the fusion of both modalities by estimating their joint occurrence (Thangthai, Harvey, Cox, & Theobald, 2015).

2.5 Deep learning

In the machine-learning community, deep learning approaches have recently attracted increasing attention as a result of deep neural networks that effectively extract robust latent features that modify numerous recognition algorithms to demonstrate revolutionary generalization capabilities under numerous application conditions(Heimbach, 2018). It is widely used due to its ability to tackle audio visual speech recognition problems in its achievement in both speech recognition and image recognition.

2.5.1 Artificial Neural Network (ANN)

An artificial neural network (ANN) is a tensile mathematical arrangement which is proficient of identifying complex nonlinear relationships among input and output data sets. ANN models have been found valuable and well-organized, mainly in troubles for which the characteristics of the processes are complicated to describe using physical equations.

2.5.2 Deep Neural Network

DNN are ANN models of hidden units between inputs and outputs with multiple layers. Applying DNN for a solution to the problem of image classification is made due to three factors:

popularization, availability of large dataset and development of powerful machine learning techniques (Heimbach, 2018).

Thangthai et al. explored the use of DNNs in visual and audiovisual speech recognition by building multimodal speech recognizer using DNN. In contrast to other researches done in conventional method they used large dataset instead of small and restricted dataset used in previous methods like HMM. For visual speech (lipreading), DNNs gave 85% word accuracy, a huge improvement on the baseline HMM performance of 33% (Feng, Guan, Li, Zhang, & Luo, 2017).

2.5.3 Convolutional Neural Network

CNN is a variant of DNN utilized for image classification and integrate three architectural ideas local receptive fields, shared weights and spatial sub sampling(Salama et al., 2014). It is one of the most utilized neural network architecture for image clustering problems and it is used as visual feature extraction mechanism(Heimbach, 2018). A CNN is a type of neural network of which the architecture is inspired by the architecture of the visual cortex. A neuron in the visual cortex responds to stimuli occurring in a restricted subsection in the field of vision. The specific region of space to which the individual neuron responds is called the receptive field. A convolutional layer applies several filters to the visual input in order to obtain feature maps. Usually, a CNN extracts an increasing amount of feature maps as its convolutional layers get deeper. However, the feature maps are reducing in size every layer it passes. By having multiple layers of convolution, increasingly complex features are extracted. Moreover, apart from convolutional layers, a CNN also contains pooling layers. The role of the pooling layers is to attain dimensionality reduction of the data by creating a summary of each subspace in the feature maps(Zhang et al., 2017).

2.5.4 Recurrent Neural Network

Recurrent Neural Networks (RNNs) are a powerful model for sequential data. An RNN is a type of neural network that is designed to process sequential data. In sequential data, the chronological order of events is one of the key characteristics. This is due to an event at an

arbitrary time step is dependent on the event that occurred before it. Examples of data for which this is true are time series data, human action recognition, speech recognition, and optical character recognition. In optical character recognition, the possible characters at position t depends on the characters at position tz until t − 1 where z is a non-negative integer. How large z is, and thus how far back the dependency goes, remains an empirical question. Moreover, RNNs take into account the chronological order of the input data by having an internal state that keeps track of the events it has seen before and the order in which it has seen said events.

Denoting the input layer, hidden layer and output layer at time t as x(t), h(t) and o(t), respectively, the output o(t) can be calculated by(Wand et al., 2016).

$$a(t)=b1+Wh(t-1)+Ux(t) \quad 2.1$$

$$h(t) = \sigma a(t)$$

$$o(t) = b2 + V\, h(t),$$

where b1 and b2 are bias vectors, U, V , and W are the weighting matrices of the input-to-hidden connection, hidden-to-output connection, and hidden-to-hidden connection, respectively, and _ is an activation function. RNNs are computationally expensive and sometimes difficult to train(Dribssa & Tachbelie, 2015).

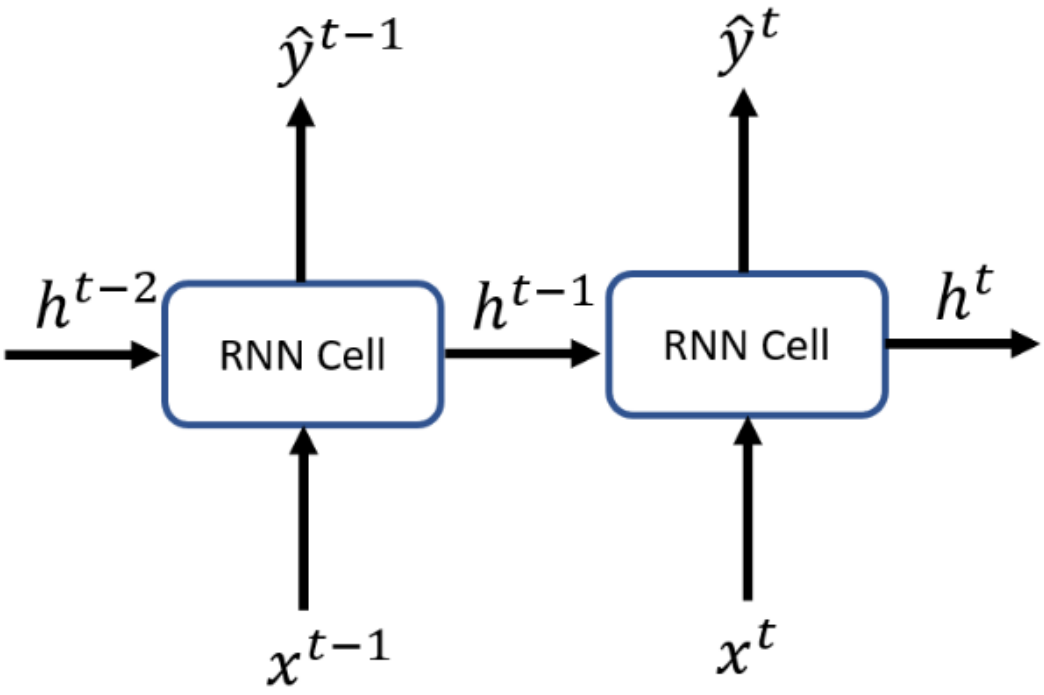

Figure 2.3 Unrolled RNN with hidden state

The figure above shows that how each recurrent neuron has an output as well as a hidden state that is passed on to the next neuron in the sequence.

2.5.5 Long Short Term Memory RNN

An LSTM layer consists of a group of recurrently connected blocks, referred to as memory blocks. In a digital computer these blocks are often thought of as a differentiable version of the memory chips. Each contains one or more recurrently connected memory cells and 3 increasing units: the input, output and forget gates that give continuous analogues of write, read and reset operations for the cells. More precisely, the input to the cells is increased by the activation of the input gate, the output to net is increased by that of the output gate, and also the previous cell values are increased by the forget gate. The net will solely act with the cells via the gates.

The LSTM lipreader with a single feed-forward network, which learns the features automatically together with training the LSTM sequence classifier, consistently achieved almost 80% word accuracy in speaker dependent lipreading(Stafylakis & Tzimiropoulos, 2017).

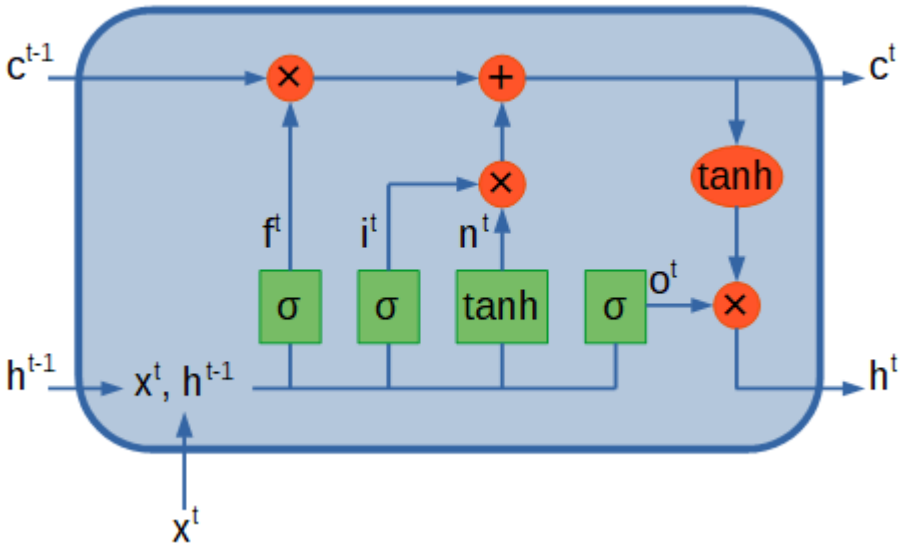

Figure 2.4 Representation of an LSTM cell

The forward propagation inside an LSTM cell can be seen in the image above. It is significantly more complex than a simple RNN. It contains four networks, each with its own set of parameters that are activated by the sigmoid function (σ) or the tanh function.

Finally, Chung and Zisserman shown that CNN and LSTM architectures can be used to classify temporal lip motion sequences of words with excellent results (Petridis et al., 2018).

2.6 Speech Recognition in Ethiopia

The first Amharic speech recognition is developed by Solomon Berhanu in 2001 [27]. The author developed isolated Consonant-Vowel syllable Amharic recognition system which recognizes a subset of isolated consonant-vowel (CV) syllable using HTK Hidden-Markov Modeling Toolkit (Hidden-Markov Modeling Toolkit). Subsequently, several speech recognition research attempts are made as discussed by(Tachbelie et al., 2014).

A. E. Dribssa and M. Y. Tachbelie studied the possibility of developing a large vocabulary continuous speech recognizer (LVCSR) large vocabulary continuous speech recognizer for Amharic using the different syllable types V, CV, VC, CVC, VCC and CVCC found in the language as acoustic units. They developed the recognizer using HMM as a modeling technique using Amharic read speech corpus which contains 20 hours of training speech

collected from 100 speakers. Their result shows that the use of all Amharic acoustic syllables are promising units for Amharic LVCSR provided that enough training data is available [27].

M. Melese et. al studied Amharic speech recognition for Amharic-English speech translation in tourism domain. They developed read-speech corpus of 7.43hr in tourist domain by recording the speech corpus after translating Basic Traveler Expression corpus under normal working environment. They used for acoustic models phoneme and syllable units and for language model morpheme and word. Their result indicated recognition accuracy of 89.1%, 80.9%, 80.6% and 49.3% at character, morph, word and sentence level respectively(Melese et al., 2016).

M. B. B. Frew design and develop the first automatic audio-visual Amharic speech recognition using lip reading. In his study he used Viola-Jones object recognizer called haar cascade for face and mouth detection after the mouth detection ROI is extracted and DWT for visual feature extraction and LDA to reduce visual feature vector. He used MFCC for audio feature extraction. Decision fusion is used to integrate audio and visual features. As a result he used three classifiers. The first one is the HMM for audio only speech recognition, the second one is HMM for visual only speech recognition and the third one is CHHM for audio- visual integration. He used his own data corpus called AAVC and evaluated using speaker dependent and speaker independent sets using both phone(vowels) and isolated word recognition using frontal face only. His result indicated accuracy for over all dataset 71.45% for visual only, 76.34% for audio only and 83.92% for audio-visual in word recognition whereas in vowels recognition achieved accuracy of 68.04% for visual, 71.96% for audio and 76.79% for audio-visual(Belete, 2017).

According to a Emiru et al., using a phoneme-based BPE system with a syllabification algorithm helped users achieve better accuracy or a lower WER (18.42 percent) when using the CTC attention end-to-end technique. In order to model subwords, Emiru et al. employed CTC-attention end-to-end speech recognition at the phoneme level and looked at the usage of subword-based sub-words in Amharic ASR (Emiru, Xiong, Li, Fesseha, & Diallo, 2021).

2.7 Related Works

In this section, we briefly reviewed the related Audio visual speech recognition models, including: the HMM classifier, multi-modal RNNs, long short-term memory (LSTM), and also the related deep learning models for AVSR.

Table 2-1 Related Works

S. N	***Year***	***Author***	***Title***	***Objective***	***methodology***	***Finding/conclusion***	***Gap***
1	2017	Weijiang Feng1, Naiyang Guan2, Yu Li1, Xiang Zhang1, Zhigang Luo2	Audio Visual Speech Recognition with Multimodal Recurrent Neural Networks	To consider sequential characteristics of both audio and visual modalities	CNN followed by LSTM RNN for visual modality, LSTM RNN to handle audio modality and multimodal layer to fuse the outputs of both modalities.	The proposed multi-modal RNN model and its variants succeed to incorporate visual speech recognition and out-perform the best known audio visual speech recognition on AVletters.	1. small dataset 2. only consider frontal faces
2	2019	Befkadu Belete	Audio-Visual Speech Recognition Using Lip Movement for Amharic	To design and develop automatic audio-visual Amharic speech	HMM classifier for audio only speech recognition, another HMM classifier for visual only speech	For speaker dependent they found overall word recognition 60.42% on visual only, 65.31% on audio only and 70.1% for audio-visual. They also found overall vowels (phone) recognition	1. used static background while capturing video

			Language	recognition using lip reading	recognition and CHHM for audio-visual integration.	71.45% on visual only, 76.34% on audio only and 83.92 % on audio-visual speech based on speakers' dependent evaluation criteria. -For speaker independent They got overall word recognition 61% for visual only, 63.544% for audio only and 67.08 % for audio-visual. They also found the overall vowels (phone) recognition 68.04% for visual only, 71.96% for audio only and 76.79 % for audio-visual speech.	2. Do not consider continous speech 3. Used frontal face for visual speech recognitio n
3	2017	Marina Zimmermann, Mostafa Mehdipour Ghazi, Hazım Kemal Ekenel and Jean-Philippe Thiran	Combining Multiple Views for Visual Speech Recognition	Explore this aspect and provide a comprehensive study on combining multiple views for visual speech recognition	Extracts features through a PCA-based convolutional neural network, followed by an LSTM network. Finally, these features are	Their result shown that exploiting multiple-view data can improve the recognition results significantly for combinations involving the frontal view, the 30 and 60 view angles.	1. used simple decision fusion 2. the dataset is limited to simple phrase recognitio

					processed in a tandem system, being fed into a GMM-HMM scheme.		n
4	2018	Joon Son Chung, Andrew Zisserman	Learning to Lip Read Words by Watching Videos	To recognise the words being spoken by a talking face, given only the video but not the audio	CNN and LSTM architectures	Show that CNN and LSTM architectures can be used to classify temporal lip motion sequences of words with excellent results.	1. Do not consider audio
5	2016	SadhanaSukale, Prashant Borde, ShivanandGornale and PravinYannawar	Recognition of Isolated Marathi words from Side Pose for multi-pose Audio Visual Speech Recognition	Presents a new multi pose audio visual speech recognition system based on fusion of side pose visual features and	The feature sets derived from visual features for Side pose are extracted using 2D Stationary Wavelet Transform (2D-SWT) and acoustic features extracted using (Linear Predictive	90% correct classification was achieved on "vVISWa" data set with consideration of only side pose AV streams. These results clearly indicates that multi-pose AVSR system based on robust visual features offers enhancement is recognition of words from even	1. Do not consider frontal view 2. Only classify the words without fusing audio and

				acoustic signals.	Coding) LPC are fused and classified using the KNN algorithm.	side pose where the acoustic recognition is poor	video modalities

2.8 Summary

In this chapter different types of speech recognition are discussed in detail and the objectives of those systems are elaborated. The objective of audio visual speech recognition is to use visual signal to support audio signal in noisy conditions. The different parts of audio visual speech recognition, features of audio visual speech recognition in audio visual speech recognition are discussed. Although there a lot of works done on audio visual speech recognition for different languages like English, hindi, Marathi and so on. The only audio visual speech recognition for Amharic language is done in 2017 by Befikadu Belete in the title Audio-Visual Speech recognition Using Lip Movement for Amharic Language. However, in his work he considered only full frontal face to recognize the lip movement. M. Zimmermann et al. explored the influence of multi-view on visual speech recognition by combining possible view angle both at future levels and decision level. In consequence the result gained from the different view improves the correctness and performance significantly than frontal-view only (Zimmermann et al., 2017).

This study focus on designing and developing Audiovisual Speech Recognition using multi-view Lip Movement for Amharic Language by extending the work of (Belete, 2017). Lip gestures such as lip protrusion and lip rounding are more pronounced because the system performs best in a non-frontal view when viewed from an angle.

3 Methodology of the Study

3.1 Overview

This part contains the thesis methodology. In this section, the author discusses in more detail the research strategy, method, approach, the data collection techniques, the research measures or metrics, the research procedure or technique, the method of testing and evaluation, the algorithm used.

3.2 Overall Methodology

The overall methodology used to conduct this thesis is described in Figure 3.1 below. The methodology starts by reviewing different articles and journal papers written in the area of audio visual speech recognition and identifying the gaps on related works done on the area. Next the problem to be addressed is formulated from the gaps identified. Next data collection is performed using different methods and pre-processing is done on the collected data. Next the development and implementation strategies followed using different algorithms are identified when reviewing literatures then the data is trained, tested and validated. Finally the findings are discussed in detail stating the conclusion and recommendation of the researcher.

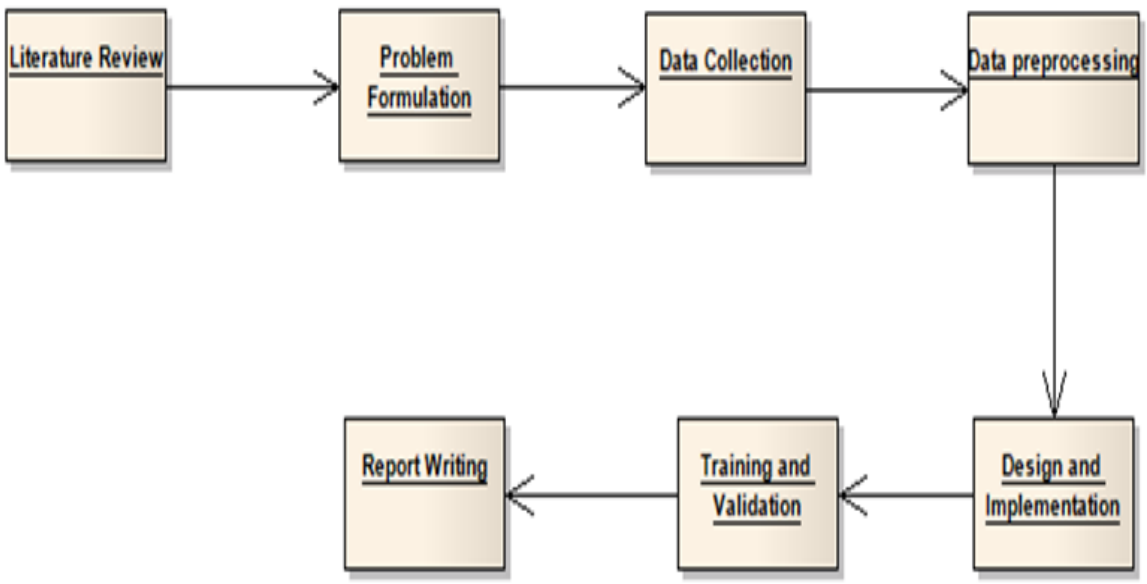

Figure 3.1Overall methodology

3.3 Literature Review

Comprehensive literature review of books, journal articles, and relevant documents from the Internet and other sources is performed in order to understand the underlying principles and theories of the various approaches, techniques and tools that can be employed in developing Audio-Visual Speech Recognition system. Deeply analyze which Deep Learning model is used to solve the identified problem and including how they are applied in previous researches. Furthermore, research works done both locally and internationally in the area of speech recognition are investigated as it helps the researcher to have a good understanding of the problem area and to support the results in a scientific manner and evaluate empirically. During the review, the gap of the previous solution will be identified to use as input for the proposed solution.

3.4 Problem Formulation

This part identifies limitations in current researches related to Audio-Visual Speech Recognition and finds out possible solution for the identified problem. By evaluating the gaps on different literatures the researcher focused on Audio-Visual Speech Detection and Recognition. The main aim of this study is to develop Multi-View Audio-Visual Speech Recognition for Amharic Language. To achieve this, deep learning methods such as CNN is trained on images from collected videos and annotated accordingly. Then LSTM RNN is used to train to learn from sequential data's of the spoken word in visual as well as audio. Then the extracted features are fused and recognized using Multimodal RNN. The deep learning model will be evaluated based on the evaluation metrics and classification accuracy.

3.5 Data Collection

The author follows a multi-stage strategy to automatically collect a large scale dataset for multi-view lip reading. It contains videos of talking faces covering all views, from frontal to profile as indicated in (Chung & Zisserman, 2017).

In order to develop the proposed approach, the data is collected from YouTube and other Broad casting website including dramas and programs engaged in conversation of single and multiple

peoples using Amharic language. The dataset contains videos of talking faces covering all views from frontal to profile.

3.6 Design and Implementation

The goal of the work is to implement an Amharic Language AVSR system. Following are the steps involved in

- As mentioned in the tools used sections, Keras is installed from those mentioned above. Those mentioned are the major ones.
- Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.
- So as to develop our own system, some shell scripts are written to configure our AVSR system, according to proposed architecture
- After modeling our ASR portion of the architecture, the model is trained with custom dataset created.
- The trained model is tested against different views.

3.7 Training and Evaluation

The aim is measuring the performance of the developed model that recognizes sentences/phrase spoken by different persons. The performance evaluation plays a vital role in accuracy measurement through different process of the developed model. In this study, an evaluation was conducted on the model based on the provided dataset.

3.8 Tools Used

Investigation of software tools with their related libraries is conducted to use the appropriate tool for the implementation of multi-modal RNN algorithm for Audio-Visual Speech Recognition. During the investigation, tools for deep learning are selected using different criteria such as appropriateness to the work, freely availability with corresponding libraries, enough learning materials. Software Tools to implement CNN, LSTM RNN and multimodal RNN algorithms are python programming language with Opencv on anaconda environment.

3.8.1 Software Tools

Anaconda

Anaconda is used as a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), with the aim of simplifying package management and deployment.

Anaconda Navigator is a desktop graphical user interface (GUI) included in the Anaconda distribution, this tool allows users to launch applications and manage Conda packages, environments, and channels without having to use command line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux. The following applications are available by default in Navigator: Jupyter Lab, Jupyter Notebook, QtConsole, Spyder, Glue, Orange, RStudio, Visual Studio Code.

OpenCV

OpenCV is a massive open source library for computer vision, machine learning, and image processing, and it now plays a significant role in real-time operation, which is critical in today's systems. It can process images and videos to identify objects, face, and even human keras handwriting.

Keras

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Keras is a neural network application programming interface API for Python that is tightly integrated with tensorflow, which is used to build machine learning models. Keras' models offer a simple, user-friendly way to define a neural network, which will then be built for you by TensorFlow. Long short-term memory (LSTM) networks are built in the Keras framework (Brownlee, 2017).

SyncNet

SyncNet learns a joint embedding between the sound and the mouth motions from unlabeled

data. This network can be used to perform audio visual synchronization tasks such as

- Removing temporal lags between audio and visual signals in a video and
- Determining who is speaking among multiple faces in a video subtitle edit.

Subtitle Edit 3.6.5

Subtitle edit is a free video subtitle editor that allows you to easily adjust a subtitle if it is out of sync with the video, among other things.

Microsoft Word

Microsoft Word is a word processor that was created by Microsoft.. It is the document editor that can be used for Report Writing.

Zotero

Zotero is a reference manager. It is intended to store, manage, and cite bibliographic references, such as books and articles. In Zotero, each of these references creates an item. More broadly, Zotero is a powerful tool for collecting and organizing research information and sources.

EdrawMax

EdrawMax can be used to build diagrams or charts with its built-in editable symbols and templates for a variety of categories.

3.8.2 Hardware Tools

To accomplish the tasks specified in this paper, hardware with the following specific components are used. Hp Computer with processor speed of 2.30GHZ , Operating System type Windows 10, RAM 8GB, GPU type NVIDIA , Hard Disk space of 500GB and System type of 64-bit processor.

4 Proposed Solution

4.1 Overview

In this chapter, the proposed framework and main steps are discussed in detail according to the following four parts. The pipeline of the proposed AVSR model is shown in The proposed method is included in this paper, as shown in **Error! Reference source not found.** The figure below displays all the steps included in the process from data collection until recognition, each of which is discussed in detail in the following sub-sections.

which explains the stages for the proposed system including data preparation, pre-processing, Feature extraction and model generation. Firstly, we demonstrate the flow chart of the proposed method. Secondly: we describe the data collection strategy. Then we need to preprocess the videos, including separating audio and video signals, extracting key frames and positioning the mouth (Y. Lu & Li, 2019). Then, features are extracted from the preprocessed image dataset by using OpenCV for visual data And MFCC for audio data. Then the extracted feature of image is preprocessed using CNN and transferred to BLSTM and the extracted feature of audio is directly transferred to BLSTM for model generation. Finally, Multimodal RNN is used to fuse the audio and visual features and dynamically perform the classification.

4.2 Proposed Method

The proposed method is included in this paper, as shown in **Error! Reference source not found.** The figure below displays all the steps included in the process from data collection until recognition, each of which is discussed in detail in the following sub-sections.

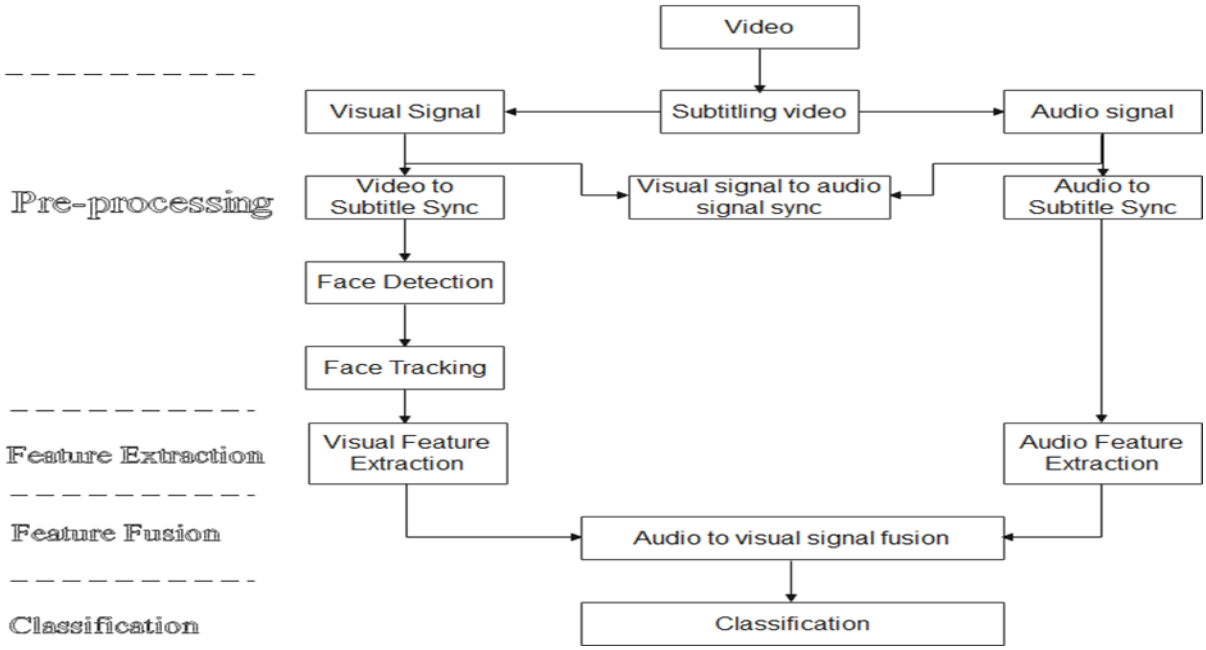

Figure 4.1proposed AVSR framework

4.3 Data collection

We gathered the 23 videos from a broadcasting news and YOUTUBE videos, including news and interviews. The sources and number of videos collected from each source are indicated in *Table 4-1* below.

Table 4-1 Data sources

Sources	Number of videos
Fana TV	7
Walta TV	5
ETV	5
OtheYouTube sources	6

After downloading the video its format is converted to Avi format then the video files are broken into 10 minute segments after downloading so that subtitles and other audio-to-visual synchronization can be conducted easily. This is done using ffmpeg command shown below in Equation __. Figure 4.2 shows the example image of an actual news broadcast video from

Fana TV.

Figure 4.2 Sample news image from Fana TV

4.4 Data Preprocessing

Data preprocessing in machine learning refers to the method of preparing (cleaning, organizing) raw data suitable for building and training machine learning models. It is an important step in helping to improve the quality of the data and facilitate the extraction of meaningful insights from the data. It aims to facilitate the training/testing process by appropriately transforming and scaling the entire dataset.

In our experiment, we preprocess the multimodal data to get a standardized input format of the audio and the image data. For the audio data, we keep using 16kHz mono-channel audio data and normalize the value with zero mean and standard deviation of one. For the image data, we extract 25 image frames per second from the video using FFMPEG(Li, 2018). To reduce the computational cost, we extract only the lip part of each image with the face landmark detection module (Kazemi & Sullivan, 2014)from the DLIB library by using a face

landmark detection model. The preprocessing of data used in this paper consists of the following stages and is closely related to (Chung & Zisserman, 2017) which include from adding subtitles to video to detecting faces and combining them into face tracks as listed below:

- Adding subtitles to the video
- Aligning the audio and subtitles in the video
- Correcting audio-to-video synchronization
- Detecting faces and combining them into face tracks

4.4.1 Subtitling video

Subtitles really convey the dialog happening in a video and it is used to split the video into word size and also used for testing and prediction purposes. However, actually many of the Broadcasting news, talks and dramas written in Amharic doesn't have subtitle, which is fairly significant. So, the subtitles mostly are manually created for every video in a basically major way. The subtitles definitely are created by the use of subtitle edit 3.6.5.

4.4.2 Splitting audio signal from video

A video to video converter splits the audio signal from the video. From the video file, the audio file is converted to wave format. Waveform Audio File Format is a standard for storing an audio bit stream on PCs that was developed by IBM and Microsoft. It is the most used uncompressed audio format on Microsoft Windows systems.

4.4.3 Aligning subtitle to video

After the subtitle is created the other major task to be done is on synchronization of subtitle to the visual signal, this is done using subtitle edit.

4.4.4 Aligning audio to video

In many news and dramas the audio to video is not properly synchronized. However, speech recognition and multimedia communication rely heavily on audio and video synchronization. In live video conferencing, audio-visual sync is a key issue. It's due to the

usage of a variety of hardware components, which introduces varying delays and software environments. The breakdown of synchronization between audio and video causes viewers to lose interest in the presentation and reduces its effectiveness. The purpose of synchronization is to maintain the temporal alignment of the audio and video streams (Thenmozhi & Kannan, 2018). Therefore, the other task is aligning this audio signal to the visual signal and this is done using the subtitle edit.

4.4.5 Aligning the audio and video to the subtitle

For classification purpose aligning the audio and video signal with the subtitle is required. The audio signal is aligned with the subtitle using subtitle edit. This is also necessary for training and testing purpose too.

4.4.6 Word Extraction

The videos are broken down into individual words using subtitles. This is accomplished by utilizing a bash script to break video files into separate pieces depending on the time codes from the subtitle file. By extracting the time codes from the subtitle file, this script creates a separate video file for each subtitle duration. All of the words are under one second long.

4.4.7 Face Detection and Tracking

In talks where many speakers are participated one of the major activity needed is audio visual speech recognition is determining which person is speaking. SyncNet is used to determine which face is speaking in talks or debates where many persons are participated. CNN with Single Shot Multibox Detector(SSD) is used to track face appearances in individual frame. SSD detects faces from all angles, and shows a more robust performance while being faster to run than HOG-based detector [29].

show video frames of faces tracked using CNN while speaking a single word.

Figure 4.3Sample face tracking images

4.4.7.1 Facial pose estimation

Head pose estimation (HPE) is defined in a computer vision context as the calculation of a person's head position and orientation in relation to the camera in front of which they are standing(Galilea, 2016). The major goal of this challenge is to determine the human's head's relative orientation (and location) in relation to the camera. An ideal head posture estimator would produce findings that were completely independent of any image-altering variables, such as lighting, distortion, noise, or facial emotions. It is usual to estimate relative orientation with Euler angles – yaw, pitch, and roll – in the head pose estimation task. Six degrees of freedom (DOF), three rotation angles, and three linear shifts are all available to the head. Pitch (X), yaw (Y), and roll (R) are the Euler angles that describe the three angles (Z).

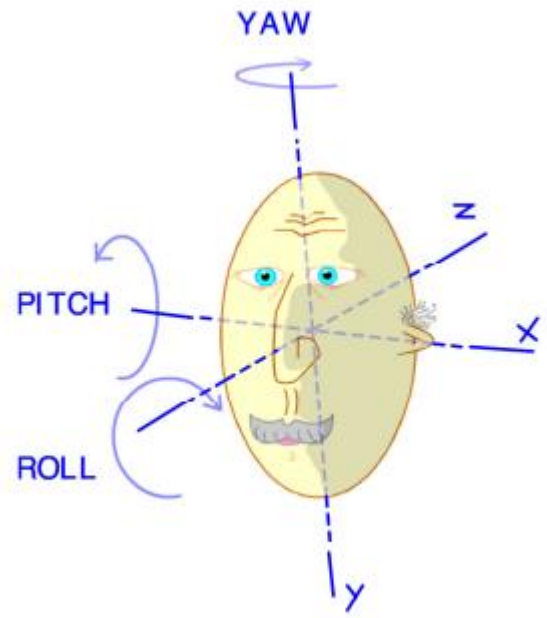

Figure 4.4Head six degrees of freedom(Guàrdia Santesmases, 2014)

In order to facilitate the testing of the multi-view model, we divide the data into 5 pose categories based on the yaw-rotation of the face such as Left profile, Left three quarter, Front, Right three quarter and Right profile.

4.4.7.2 Lip detection

Many applications, such as face detection and lip reading, rely on lip detection. When it comes to lip reading, the look or shape the picture takes when speaking is crucial to comprehending the letters or words being spoken. Every frame of the face track determines facial landmarks. According to (Dastidar, Basak, Hota, & Athar, 2018), we utilize a Support Vector Machine (SVM) based on feature distance based lip detection. SVM is a type of supervised learning that may be used as a lip recognition classifier. When developing the SVM technique, it is required to specify and parameters, which are the two aspects determining model quality. Figure 4.5 shows lip detected using SVM from each frame of video for a single word. Mouth areas were resized to have all the same size so they could be used in a single neural network, as shown in the following figure.

Figure 4.5 Sample extracted lip

The screen capture is restricted to the mouth area. In this case, the significant portion of the face is. The mouth corners (left, right, top, and bottom) are used to automatically locate this location. In a single 3D model, the top and bottom) were manually marked because all meshes have the same size, they just need to be tagged once. The topology is the same the appropriate 3D model is generated in each cycle.

4.5 Feature Extraction

After preprocessing our data, there are different methods used to extract various features from the audio and visual signal so that we can train our model to predict recognition accuracy. The features are extracted using the following mechanisms.

1. Audio feature is extracted by using MFCC.

Mel Frequency Ceptral Coefficient is a very common and efficient technique for signal processing. They are a small group of characteristics that precisely describe the entire shape of a spectral envelope (Bahuleyan, 2018). The MFCC uses the MEL scale to divide the frequency band into sub-bands. Then, the Cepstral Coefficients are extracted by computing the Discrete Cosine Transform (DCT). It is a standard method for feature extraction in speech recognition. Figure 4.6 shows sample audio file.

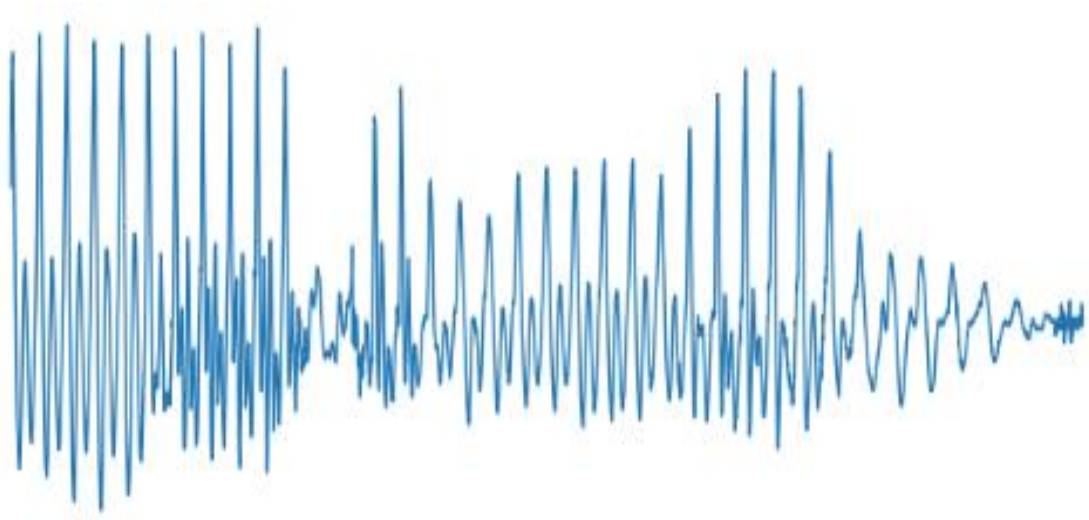

Figure 4.6 Sample audio file

```
Out[11]: array([[-1.56690262e+02, -1.98373322e+02, -2.36911713e+02,
                 -2.18462753e+02, -2.08714325e+02, -2.23578064e+02,
                 -2.73393616e+02, -3.09872437e+02, -2.87865265e+02,
                 -2.49795731e+02, -2.37506821e+02, -2.66506622e+02,
                 -3.32645264e+02, -3.22087708e+02, -3.24081787e+02],
                [ 1.35796967e+02,  1.49132385e+02,  1.80547089e+02,
                  1.85651825e+02,  1.92163986e+02,  2.00321808e+02,
                  1.90261719e+02,  1.58207413e+02,  1.61632156e+02,
                  1.60926056e+02,  1.57111633e+02,  1.53721924e+02,
                  1.39830627e+02,  1.12320312e+02,  8.51186218e+01],
                [ 2.23621635e+01,  2.89749870e+01,  4.42613907e+01,
                  3.36856651e+01,  1.55171604e+01,  1.03643360e+01,
                  4.49606400e+01,  5.57627716e+01,  5.77108688e+01,
                  5.11904869e+01,  4.04268837e+01,  4.51377258e+01,
                  5.93734894e+01,  2.51565170e+01, -4.30821848e+00],
                [ 3.47907181e+01,  3.09793396e+01, -4.66927814e+00,
                 -8.24095726e+00,  3.24215603e+00,  5.06811810e+00,
                  5.11743164e+00,  2.58995705e+01,  2.47440720e+01,
                  1.88667984e+01,  3.28517380e+01,  5.12528648e+01,
```

Figure 4.7 Audio file extracted using MFCC

Figure 4.7 above shows the audio files after extracted by MFCCs

2. Visual feature is extracted using CNN.

CNNs are a common deep learning architecture that has been widely employed in computer vision and audio recognition. CNN is used to create features automatically and mix them with the classifier. The list of layers that translate input volume to output volume is the simplest of all the classifiers in this method (Jogin, Madhulika, Divya, Meghana, & Apoorva,

2018), which is one of the advantages of CNN classifier.

4.6 Zero Padding

The changing length of frame sequences in the input was another more technical issue that arose during the creation of both lip reading networks. This problem was addressed in separate areas of both networks using diverse ways. Only the time steps matching to actual frames from the video were evaluated during back propagation in time using zero padding in the input level and dynamic LSTM cells in the BiLTSM network (Doukas, 2017).

Because each sample has a variable number of frames, zero padding is required post-processing (Sayed, ElDeeb, & Taie, 2021). By employing zero padding, the audio signal is padded to accommodate a maximum of 1s. The term "zero padding" refers to the practice of padding a signal (or spectrum) with zero. The rest of the grid is Zero-padded in visual signals with utterances of fewer than 25 frames. There is no padding used, therefore all frame sequences are the same length.

4.7 Audio Visual Fusion and Classification

The late fusion process is what we employ for feature fusion. This method employs a different modeling procedure for each modality, which takes the properties of one modality as input and generates an output choice. The decision integration unit then combines these to generate the final outcome. The outputs of the modeling processes share the same representation in late integration, making merging them easier than mixing feature vectors, as in early integration(Katsaggelos, Bahaadini, & Molina, 2015). After extracting the audio and visual part separately using bidirectional LSTM and CNN plus bidirectional LSTM respectively the feature is fused using multi-modal RNN. The Audio-Visual fusion was done using multi-modal RNN because multi-modal learning improves the classification performance through exploiting the sharable knowledge of the data across modalities.

5 Implementation of the Proposed Solution

5.1 Introduction

This chapter presents the experiments and the procedures followed while conducting the experiments in this research. As the objective of this research is to develop audio-visual speech recognition by using multi-view lip reading for the Amharic language, the dataset of words is taken from online videos released through YouTube and other broadcasting sites as discussed on proposed model above. For this, the dataset is divided into training and testing with the training stack receiving around 80% of the data, the testing stack receiving around 20% of the data. All the training and testing conducted are presented as follows.

5.2 Dataset Description

The dataset contains views of faces from 5 different angles such as left profile, left three quarter, right three quarter, right profile and front views. This views are separated by using ranges from -18 to +18 as front , -54 to -18 as left three quarter, -90 to -54 as left profile, +18 to +54 as right three quarter and +54 to +90 as right profile. Each individual word images and audio are kept in different folders using similar folder structure stating the speaker and word in number. Table 5.1 contains the dataset used for this paper

Table 5-1 Dataset by their views

Angles	***Number of words***	***Number of utterances***
Right profile	25	990
Right three quarter	30	895
Front	50	1285
Left three quarter	25	910
Left Profile	25	920

The dataset used in this work includes 5000 utterances from 50 different words that are frequently used in TV programs, as seen in the table above. Frontal views are the most common statements since the majority of TV programs, such as news and shows, are shot from this perspective.

Before the audio and video can be used for recognition, they must be properly subtitled to match the spoken words. Each word and speaker is assigned an ID. Furthermore, we assign sequence numbers to videos of a word. As a result, we assigned a speaker id, a word id, and a video id sequence number to the captured films. Table 5.2 lists the words we've collected along with their meanings.

Table 5-2 Sample collected words

ID	*Word*	*Number of utterances*
1	ተመልካቾች	100
2	ሰዎች	100
3	ሰዓት	100
4	መንገድ	100
5	እንችላለን	100
6	በዋናነት	100

5.3 Evaluation Protocol

In this paper we focus on two types of datasets: training datasets and test datasets. Training datasets are used by machine learning algorithms to learn how to perform tasks such as image recognition, speech recognition, natural language processing (NLP) etc. The video segments in the dataset partitioned into training, validation, and test set. After training, validation, and testing the dataset, the model is evaluated by its recognition accuracy of different view angles.

Table 5-3 Training, Validation and Test set sizes

Sets	*Number of utterances*
Training	4000
Testing	1000

5.4 Feature Extraction

The feature extraction, which extracts features from speech signals, is the most significant aspect of the speaker recognition. Feature extraction is a method that converts input data into a set of features. It decreases the dimension of the input vector while preserving the crucial discriminating feature of a speaker in feature extraction.

The audio features are extracted using MFCC (Mel Frequency Cepstral Coefficients). It is the widely used technique for extracting the features from the audio signal. To acquire the MFCC coefficient, the input speech signal is windowed and converted into frequency domain using the Discreet Fourier transform (Singh, Khan, & Shree, 2012). We're covering the Mel frequency with a bank filter here. The log magnitude of each Mel frequency is then obtained. The generated signal is then converted into the cepstral domain using an inverse discrete Fourier transform (DFT).

CNN is used to extract the video features (Ravanbakhsh, Mousavi, Rastegari, Murino, & Davis, 2015). After the video has been transformed into a series of photos using OpenCV, the feature detection is performed. Because of its simplicity, quick processing time, and strong demand in computer vision applications, CNN is one of the most popular used visual feature extraction mechanisms.

5.5 Model Implementations

Identifying the data you need and preparing the data for machine learning models is critical to the success of machine learning implementation. We implement the model

using python. Python is a popular and general-purpose programming language. The reason why Python is so popular among data scientists is that Python has a diverse variety of modules and libraries already implemented that make our life more comfortable.

5.5.1 Model Training

Model training is essential for determining the accuracy of your model. While training we use 50 epochs for the model to train with batch size of 32. The audio and visual signals are trained individually and model for each signal is generated and trained with the data. The visual signal is trained using mobile net framework with width of 128 and height of 128. The epoch and batch size for each signal is similar. After the models are generated for each signal we concatenate the model together.

5.5.1.1 Audio model training

Figure 5.1 depicts the epoch training for an audio-only model, with a training accuracy of 100% and a testing accuracy of 80%. Figure 5.2 shows the accuracy curve of the audio-only model against the epoch.

```
Epoch 36/50
1/1 [==============================] - ETA: 0s - loss: 0.0033 - acc: 1.0000
Epoch 36: val_loss did not improve from 0.52025
1/1 [==============================] - 0s 42ms/step - loss: 0.0033 - acc: 1.0000 - val_loss: 1.3152 - val_acc: 0.8000
Epoch 37/50
1/1 [==============================] - ETA: 0s - loss: 0.0019 - acc: 1.0000
Epoch 37: val_loss did not improve from 0.52025
1/1 [==============================] - 0s 47ms/step - loss: 0.0019 - acc: 1.0000 - val_loss: 1.3536 - val_acc: 0.8000
Epoch 38/50
1/1 [==============================] - ETA: 0s - loss: 0.0052 - acc: 1.0000
Epoch 38: val_loss did not improve from 0.52025
1/1 [==============================] - 0s 44ms/step - loss: 0.0052 - acc: 1.0000 - val_loss: 1.3967 - val_acc: 0.8000
Epoch 39/50
1/1 [==============================] - ETA: 0s - loss: 0.0023 - acc: 1.0000
Epoch 39: val_loss did not improve from 0.52025
1/1 [==============================] - 0s 49ms/step - loss: 0.0023 - acc: 1.0000 - val_loss: 1.4373 - val_acc: 0.8000
Epoch 40/50
1/1 [==============================] - ETA: 0s - loss: 0.0017 - acc: 1.0000
Epoch 40: val_loss did not improve from 0.52025
1/1 [==============================] - 0s 46ms/step - loss: 0.0017 - acc: 1.0000 - val_loss: 1.4468 - val_acc: 0.8000
Epoch 41/50
1/1 [==============================] - ETA: 0s - loss: 0.0027 - acc: 1.0000
Epoch 41: val_loss did not improve from 0.52025
1/1 [==============================] - 0s 54ms/step - loss: 0.0027 - acc: 1.0000 - val_loss: 1.3881 - val_acc: 0.8000
Epoch 42/50
```

Figure 5.1Training of Epochs for Audio

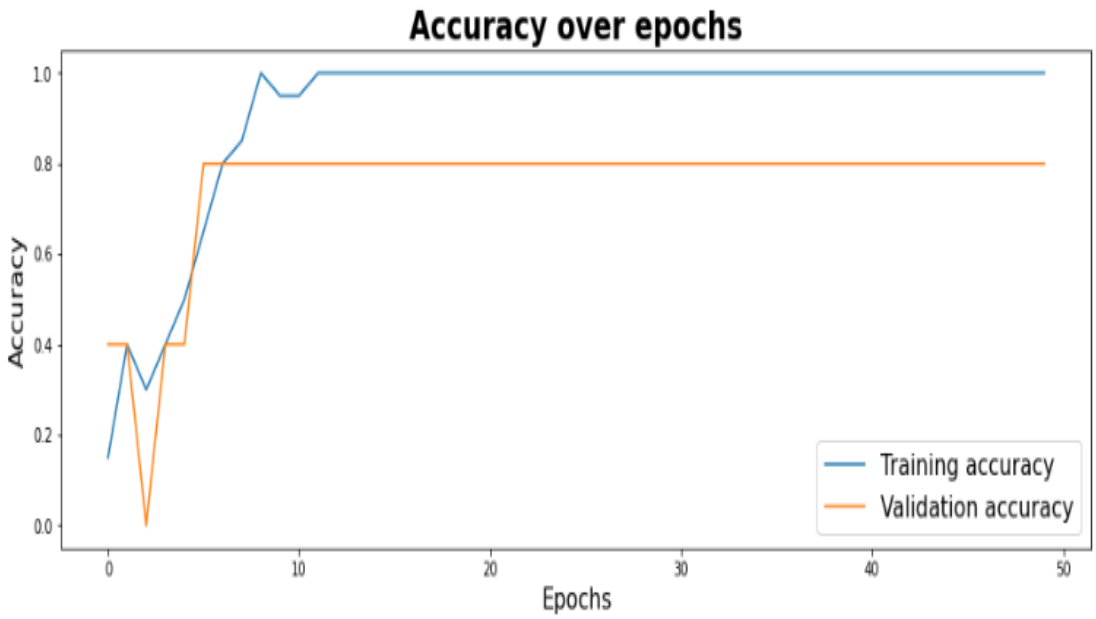

Figure 5.2Model training accuracy curve for audio, epoch v accuracy

We get the graph above for the current audio model training. It demonstrates that training accuracy is improving: the model is learning something and has reached 100% accuracy. However, the validation accuracy does not improve significantly, indicating that the model will require some fine tuning before it becomes usable.

5.5.1.2 Visual model training

The number of epoch employed for the visual model training model is shown in Figure 5.3 with training accuracy of 99 percent and testing accuracy of 96.08 percent. As the number of training epoch increases, so does the accuracy result. Figure 5.4 shows the model accuracy model for the visual model and the graph is epoch versus accuracy. This graph shows the training and validation graph. depicts the visual model's model accuracy model, with an epoch against accuracy graph. The training and validation graphs are shown in this graph. Figure 5.5 depicts the visual model's loss curve, which includes both training and validation losses.

```
Epoch 21/50
7/7 [==============================] - 71s 10s/step - loss: 0.9596 - accuracy: 0.9104 - val_loss: 1.2837 - val_accuracy: 0.7843
Epoch 22/50
7/7 [==============================] - 71s 10s/step - loss: 0.8541 - accuracy: 0.9353 - val_loss: 1.1145 - val_accuracy: 0.8039
Epoch 23/50
7/7 [==============================] - 70s 10s/step - loss: 0.7387 - accuracy: 0.9353 - val_loss: 1.0158 - val_accuracy: 0.8431
Epoch 24/50
7/7 [==============================] - 73s 11s/step - loss: 0.6571 - accuracy: 0.9502 - val_loss: 0.9515 - val_accuracy: 0.8039
Epoch 25/50
7/7 [==============================] - 71s 10s/step - loss: 0.5894 - accuracy: 0.9602 - val_loss: 0.8043 - val_accuracy: 0.8627
Epoch 26/50
7/7 [==============================] - 70s 10s/step - loss: 0.5075 - accuracy: 0.9602 - val_loss: 0.7342 - val_accuracy: 0.8627
Epoch 27/50
7/7 [==============================] - 71s 10s/step - loss: 0.4584 - accuracy: 0.9701 - val_loss: 0.6462 - val_accuracy: 0.9216
Epoch 28/50
7/7 [==============================] - 72s 10s/step - loss: 0.4006 - accuracy: 0.9801 - val_loss: 0.5732 - val_accuracy: 0.9608
Epoch 29/50
7/7 [==============================] - 71s 10s/step - loss: 0.3507 - accuracy: 0.9851 - val_loss: 0.5142 - val_accuracy: 0.9608
Epoch 30/50
7/7 [==============================] - 71s 10s/step - loss: 0.3055 - accuracy: 0.9900 - val_loss: 0.4474 - val_accuracy: 0.9608
Epoch 31/50
7/7 [==============================] - 78s 11s/step - loss: 0.2687 - accuracy: 0.9950 - val_loss: 0.4009 - val_accuracy: 0.9608
Epoch 32/50
7/7 [==============================] - 76s 11s/step - loss: 0.2339 - accuracy: 0.9900 - val_loss: 0.3615 - val_accuracy: 0.9608
```

Figure 5.3 Training of Epochs for visual speech

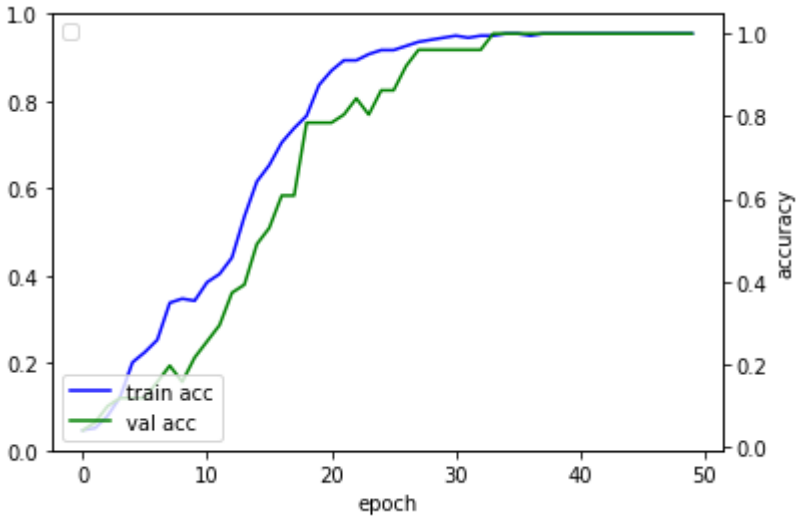

Figure 5.4Model training accuracy curve for visual speech, epoch v accuracy

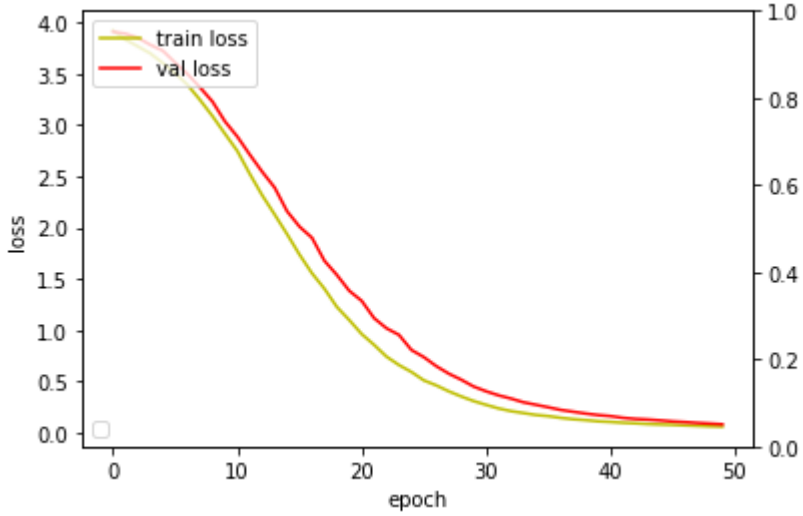

Figure 5.5Model training loss curve for visual speech, epoch v loss

5.5.2 Model Testing

We should test our model after training it to make sure it functions properly. This may be accomplished by running various data sets through the model. Based on the concatenated model created using our test data in our dataset, we have tested our model.

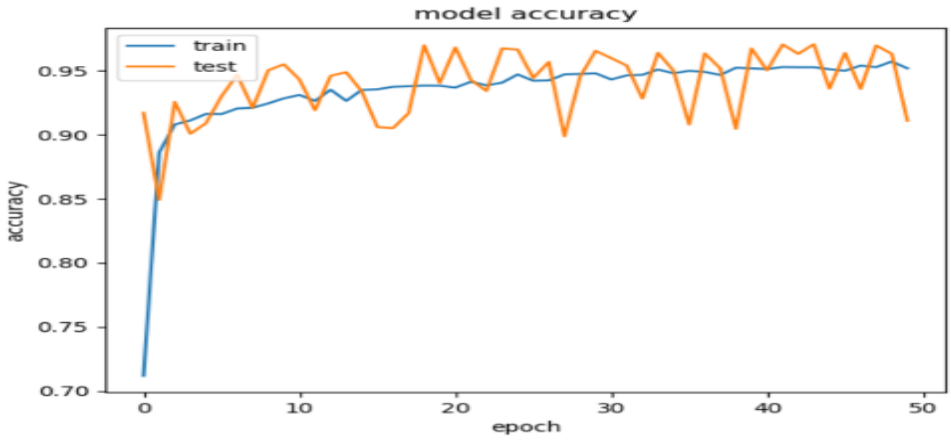

Figure 5.6 Testing of epoch on concatenated model

Figure 5.6 above shows the test accuracy of our concatenated model. It has a test accuracy of around 95%.

6 Result and Discussion

6.1 Overview

The author performed different experiments engaging the multimodal corpus presented in this paper. Results realized with the developed audio-visual speech recognition (AVSR) model trained and tested with the material confined in the corpus are presented and discussed with comparative test results. This chapter starts by discussing the results obtained depending on the side employed that is single side view result and multiple side view result. Next, there is a discussion sub section in which we discuss and analyze the results obtained stating the improvements gained using side view , effect of using combined views and effect of multi-view in audio visual speech recognition.

6.2 Results

We tested the isolated word audio visual recognition system on the database we developed. Each word in the database is repeated one hundred times by different speakers in the database. For each word, eighty examples of each word were used for training and the remaining example was used for testing. In our experiments we compared the accuracy of the single side view result and multiple side view results using the multimodal audio visual speech recognition (AVSR) model described in Chapter 4. For each of the audio-only and video only recognition tasks, we model the observation sequences using Bidirectional LSTM and CNN followed by BLSTM.

This section summarizes the results of the model's testing using the accuracy attained in each perspective. Model accuracy is a metric used to determine which model is the best at identifying relationships and patterns between variables in a dataset based on input or training data.

We compare the audio visual speech recognition results with the results attained for each view using multimodal-RNN model in this section. The results are separated in

two the single view results as well as multiple view results.

6.2.1 Single view results

We conducted different experiment using the same setup in the each single view experiment. Each side's findings are taken, including front view, left three quarter, right three quarter, left profile, and right profile. As shown in the table below, the left three quarters and the right three quarters have better results. This viewpoint is more powerful than others in terms of pronouncing words spoken. Table 6-1 represents the single-view experiment result on test data. Our model produces the reasonable result of the prediction.

Table 6-1 Single-view word test accuracy results for each side view.

Side	***Accuracy***
Front	92.5%
Left three quarter	95%
Right three quarter	93%
Left profile	92%
Right profile	90.5%

6.2.2 Multiple view results

We conducted different experiment using the same setup by combining different views together in each experiment. This result demonstrates the accuracy attained by combining multiple views from various perspectives. Table 6.2 shows that combining frontal and left three quarter views improves accuracy over other methods.

Table 6-2 Muti-view results

Side	*Accuracy*
Front with left three quarter	97.5%
Left three quarter with left profile	94%
Front with Right three quarter	96.35%
Right three quarter with Left profile	92.5%
All views	96.5%

6.3 Discussion

The researcher discusses the findings from the results section above in this section stating how each outcome should be interpreted. From the results stated in the above sub section we can make several observations. Here the discussion part is classified into three, the first part states the improvements attained using side view, the second states improvements attained by combining side view with frontal view and the last part states the multi-view recognition accuracy and its effects.

6.3.1 Improvements by using side view

Frontal view test accuracy is 92.5% as seen in table 6.1, the left three quarter 95% and right three quarter view results 93% which is are more accurate than the frontal view. As a result, employing those views for lip reading is more potent than using the frontal view.

6.3.2 Combining side view with frontal view

Looking at the results in

Table 6-2 we can see that there are various improvements over the baseline results for the Front view while combined with other views like left three quarter view and right

three quarter view.

As seen

Table 6-2, combining the left three quarter view with the frontal view improves accuracy over other viewpoints. Frontal view alone produces worse accuracy than frontal view with left three quarters. As a result, combining those views is more accurate than using just one.

6.3.3 Multi-view recognition accuracy

As long as there are more than two view points in the aggregate, all views are called multi-viewpoints. As a result, while employing all views yields a lower accuracy of 96.5 percent than frontal view with left three quarter views, however, the accuracy is still higher than using a single frontal view. As a result, multi-view recognition improves speech recognition accuracy in the lip reading process. Comparing the combinations of more than just two views, we can see that they improve the results further.

7 Conclusion, Recommendation and Future Work

7.1 Conclusion

In this paper we have presented our newly recorded multi-view audiovisual database for Amharic language. It contains 50 videos of more than 17 speakers collected from YouTube and other broadcasting websites. It is preprocessed using CNN and MFCC. The features are extracted using LSTM RNN for audio and for the video separately and fused using multi-modal RNN.

As a result, the accuracy of combining frontal view with left three quarter view is 97.5 percent, which is higher than single view. When it comes to Amharic audio visual speech recognition, multi view is more accurate than single view.

Therefore, we can give a qualified answer to the question raised in the research question section which says “To what extent side view improve performance of Amharic speech recognition?” There is an improvement of at least 0.5% in right three quarter view and 2.5% in left three quarter view. Even though, there is a decrease in accuracy while using both profile views due to the stated improvement side view improve the performance of Amharic Speech Recognition at least in 0.5%.

The next question raised is “How to combine frontal view with side view to form multi-view lip reading?” there are different ways of combining frontal with side view the first is using moving faces that rely on different side view for single word. The other is combining the dataset of different view for single word to train and test. We use the second one in this paper which is more robust and simple to execute.

The last raise question is “What is the advantage of using multi-view over side and front view only AVSR?” using multi-view improves the accuracy of speech recognition than the front view on audio visual speech recognition(AVSR).

7.2 Contributions

The contributions and main outcomes of this paper can be listed as below:

• A thorough analysis on combination of all possible view angle combinations in a multiple view setup is provided and their influence on the performance is evaluated. It has been found that combining multiple views results in significant performance enhancement.

• Different view angles contributions to the performance are analyzed. Front and Left three quarter are found to be more useful for audio-visual speech recognition.

7.3 Recommendation and Future Work

7.3.1 Recommendation

One of the contributions of this paper is that it shows how multi-view lip reading can improve audio visual speech recognition. As the results show, it is one of the more accurate aspects of lip reading. As a result, the author suggests that each lip reading process be done in frontal view with a left three quarter view.

7.3.2 Future Work

This research demonstrates that audio-visual speech recognition is a broad field that has been studied by a number of researchers. We acquired a result in detecting single Amharic word in this thesis effort from different views. However, there are several holes that future work should fill. Because this research cannot be employed as a full-fledged Amharic language speech recognition system, we urge that future research incorporate the following components.

- ✓ This study uses Multi-modal LSTM machine learning languages however there are still a more powerful languages exist. Therefore, Future Amharic lip reading researches should consider extending the work done in this study by

using different machine learning languages.

- ✓ We plan now to increase the size of the dataset further to see if the availability of more training data will be of benefit. Also, it will be interesting to investigate how deep learning has learnt to select relevant information for each view, and whether different architectures, e.g. increasing capacity, will improve performance.
- ✓ This study uses a small number of dataset due to processing capacity constraints. Therefore, Future researches can check the accuracy employing more dataset.
- ✓ Researches should be expanded to the recognition of continuous speech by using the output of the isolated word recognition.
- ✓ We expect that the best score would come out in the multi-view result before experiments are performed; the results also reported the same. To view the results in a more dataset is necessary as the performance of DNN increases as the samples used in the database increase. Experiments with various data augmentation from multi-view data are future work.

Reference

Bahuleyan, H. (2018). Music genre classification using machine learning techniques. *arXiv preprint arXiv:1804.01149*.

Barkhan, M., Alizadeh, F., & Maihami, V. (2019). Designing and implementing a system for Automatic recognition of Persian letters by Lip-reading using image processing methods. *Journal of Advances in Computer Engineering and Technology*, *5*(2), 71-80.

Belete, B. (2017). *College of Natural Sciences.* Addis Ababa University.

Bhaskar, S., & Thasleema, T. (2019). *Scope for deep learning: A study in audio-visual speech recognition.* Paper presented at the 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE).

Brownlee, J. (2017). A Gentle Introduction to Long Short-Term Memory Networks by the Experts. *Machine Learning Mastery*, *19*.

Butt, W., & Lombardi, L. (2013). *Comparisons of visual features extraction towards automatic lip reading.* Paper presented at the EDULEARN13 Proceedings.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). *Lip reading sentences in the wild.* Paper presented at the 2017 IEEE conference on computer vision and pattern recognition (CVPR).

Chung, J. S., & Zisserman, A. (2016). *Lip reading in the wild.* Paper presented at the Asian conference on computer vision.

Chung, J. S., & Zisserman, A. (2017). Lip reading in profile.

Chung, J. S., & Zisserman, A. (2018). Learning to lip read words by watching videos. *Computer Vision and Image Understanding*, *173*, 76-85.

Dastidar, J. G., Basak, P., Hota, S., & Athar, A. (2018). SVM based method for identification and recognition of faces by using feature distances *Intelligent Engineering Informatics* (pp. 29-37): Springer.

Doukas, M.-C. (2017). *Deep Lipreading.* Imperial College London.

Dribssa, A. E., & Tachbelie, M. Y. (2015). *Investigating the use of syllable acoustic units for amharic speech recognition.* Paper presented at the AFRICON 2015.

El Maghraby, E. E., Gody, A. M., & Farouk, M. H. (2016). *Enhancing quality and*

accuracy of speech recognition system by using multimodal audio-visual speech signal. Paper presented at the 2016 12th International Computer Engineering Conference (ICENCO).

ElMaghraby, E. E., Gody, A. M., & Farouk, M. H. (2020). Noise-robust speech recognition system based on multimodal audio-visual approach using different deep learning classification techniques. *The Egyptian Journal of Language Engineering, 7*(1), 27-42.

Emiru, E. D., Xiong, S., Li, Y., Fesseha, A., & Diallo, M. (2021). Improving Amharic speech recognition system using connectionist temporal classification with attention model and phoneme-based byte-pair-encodings. *Information, 12*(2), 62.

Feng, W., Guan, N., Li, Y., Zhang, X., & Luo, Z. (2017). *Audio visual speech recognition with multimodal recurrent neural networks.* Paper presented at the 2017 International Joint Conference on Neural Networks (IJCNN).

Galilea, M. A. (2016). *Contributions to head pose estimation methods.* Universidad Pública de Navarra.

Guàrdia Santesmases, R. (2014). Head pose estimation from 2D/3D images.

Han, H., Kang, S., & Yoo, C. D. (2017). *Multi-view visual speech recognition based on multi task learning.* Paper presented at the 2017 IEEE international conference on image processing (ICIP).

Heimbach, K. (2018). *An Empirical Evaluation of Convolutional and Recurrent Neural Networks for Lip Reading.*

Huang, J., & Kingsbury, B. (2013). *Audio-visual deep learning for noise robust speech recognition.* Paper presented at the 2013 IEEE international conference on acoustics, speech and signal processing.

Jogin, M., Madhulika, M., Divya, G., Meghana, R., & Apoorva, S. (2018). *Feature extraction using convolution neural networks (CNN) and deep learning.* Paper presented at the 2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT).

Katsaggelos, A. K., Bahaadini, S., & Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE, 103*(9), 1635-1653.

Kazemi, V., & Sullivan, J. (2014). *One millisecond face alignment with an ensemble*

of regression trees. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Li, G. (2018). Special treatment of video image based on FFmpeg. *vol, 137*, 266-271.

Lu, X., Li, S., & Fujimoto, M. (2020). Automatic speech recognition *Speech-to-Speech Translation* (pp. 21-38): Springer.

Lu, Y., & Li, H. (2019). Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. *Applied Sciences, 9*(8), 1599.

Mehrabani, M., Bangalore, S., & Stern, B. (2015). *Personalized speech recognition for Internet of Things.* Paper presented at the 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT).

Melese, M., Besacier, L., & Meshesha, M. (2016). *Amharic speech recognition for speech translation.* Paper presented at the Atelier Traitement Automatique des Langues Africaines (TALAF). JEP-TALN 2016.

Paleček, K., & Chaloupka, J. (2013). *Audio-visual speech recognition in noisy audio environments.* Paper presented at the 2013 36th International Conference on Telecommunications and Signal Processing (TSP).

Patil, M. S., Chickerur, S., Meti, A., Nabapure, P. M., Mahindrakar, S., Naik, S., & Kanyal, S. (2019). LSTM Based Lip Reading Approach for Devanagiri Script.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). *End-to-end audiovisual speech recognition.* Paper presented at the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP).

Ravanbakhsh, M., Mousavi, H., Rastegari, M., Murino, V., & Davis, L. S. (2015). Action recognition with image based CNN features. *arXiv preprint arXiv:1512.03980*.

Saksamudre, S. K., Shrishrimal, P., & Deshmukh, R. (2015). A review on different approaches for speech recognition system. *International Journal of Computer Applications, 115*(22).

Salama, E. S., El-Khoribi, R. A., & Shoman, M. E. (2014). Audio-visual speech recognition for people with speech disorders. *International Journal of Computer Applications, 96*(2).

Sayed, H. M., ElDeeb, H. E., & Taie, S. A. (2021). Bimodal variational autoencoder

for audiovisual speech recognition. *Machine Learning*, 1-26.

Singh, N., Khan, R., & Shree, R. (2012). MFCC and prosodic feature extraction techniques: a comparative study. *International Journal of Computer Applications, 54*(1).

Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. *arXiv preprint arXiv:1703.04105*.

Sukale, S., Gornale, P. B. S., & Yannawar, P. (2016). Recognition of isolated marathi words from side pose for multi-pose audio visual speech recognition. *ADBU Journal of Engineering Technology, 5*(1).

Tachbelie, M. Y., Abate, S. T., & Besacier, L. (2014). Using different acoustic, lexical and language modeling units for ASR of an under-resourced language–Amharic. *Speech Communication, 56*, 181-194.

Thangthai, K., Harvey, R. W., Cox, S. J., & Theobald, B.-J. (2015). *Improving lip-reading performance for robust audiovisual speech recognition using DNNs.* Paper presented at the AVSP.

Thenmozhi, A., & Kannan, P. (2018). PERFORMANCE ANALYSIS OF AUDIO AND VIDEO SYNCHRONIZATION USING SPREADED CODE DELAY MEASUREMENT TECHNIQUE. *ICTACT Journal on Image & Video Processing, 9*(1).

Vadwala, A. Y., Suthar, K. A., Karmakar, Y. A., Pandya, N., & Patel, B. (2017). Survey paper on different speech recognition algorithm: challenges and techniques. *Int J Comput Appl, 175*(1), 31-36.

Virtanen, T., Singh, R., & Raj, B. (2012). *Techniques for noise robustness in automatic speech recognition*: John Wiley & Sons.

Wand, M., Koutník, J., & Schmidhuber, J. (2016). *Lipreading with long short-term memory.* Paper presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Wuth, J., Correa, P., Núñez, T., Saavedra, M., & Yoma, N. B. (2021). The Role of Speech Technology in User Perception and Context Acquisition in HRI. *International Journal of Social Robotics, 13*(5), 949-968.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., & Courville, A. (2017). Towards end-to-end speech recognition with deep convolutional

neural networks. *arXiv preprint arXiv:1701.02720*.

Zimmermann, M., Ghazi, M. M., Ekenel, H. K., & Thiran, J.-P. (2017). Combining multiple views for visual speech recognition. *arXiv preprint arXiv:1710.07168*.

Appendix A. Audio feature extraction

```
import pandas as pd
import os
import librosa

audio_dataset_path='data/audio/'
metadata=pd.read_csv('data/audio_ds.csv')

def features_extractor(file):

    audio, sample_rate = librosa.load(file_name, res_type='kaiser_fast')

    mfccs_features = librosa.feature.mfcc(y=audio, sr=sample_rate, n_mfcc=40)

    mfccs_scaled_features = np.mean(mfccs_features.T,axis=0)



    return mfccs_scaled_features

import numpy as np

from tqdm import tqdm

### Now we iterate through every audio file and extract features

### using Mel-Frequency Cepstral Coefficients

extracted_features=[]

for index_num,row in tqdm(metadata.iterrows()):

    file_name =
os.path.join(os.path.abspath(audio_dataset_path),str(row["wordid"])+'/',str(row["spea
ker"])+'/',str(row["f_name"]))

    final_class_labels=row["word"]
```

data=features_extractor(file_name)

extracted_features.append([data,final_class_labels])

Appendix B Audio Padding

```
for root,dirs,file in os.walk(speaker):
    
    for files in file:
        in_wav=files
        wavpath=os.path.join(speaker,in_wav)
        audio = AudioSegment.from_wav(wavpath)
        print(speaker)
        Fs,x=wavfile.read(wavpath)
        length=len(x)/Fs
        print("Dist  ",len(x))
        print('Duration: ' , length)
        max_len=1
        if(length>max_len):
            Fs=Fs[:,:max_len]
        elif (length<max_len):
            silence=AudioSegment.silent(duration=max_len-length+1)
            padded=audio+silence
            print(dest_speaker)
            newfilename=os.path.join(dest_speaker,in_wav)
            padded.export(newfilename,format="wav")
```

Appendix C Video padding

```
path=os.getcwd()
data1="dataset/video/"
newpath=os.path.join(path,data1)
itr=iter(os.walk(newpath))
root,dirs,files=next(itr)

maxlength=25

for d in dirs:
    dir=d
    labelpath=os.path.join(newpath,dir)
    for root,dirs1,files in os.walk(labelpath):
        for d1 in dirs1:
            dir1=d1
            personpath=os.path.join(labelpath,dir1)
            for root,dirs2,files in os.walk(personpath):
                list = os.listdir(personpath)
                number_files = len(list)
                print (number_files)
                i=number_files+1
                for file in files:

                    for i in range(number_files,maxlength+1):
                        imagecopy= np.copy(my_img_3)
                        m=str(i)
                        name=os.path.join(personpath,m)
                        cv2.imwrite(name+".jpg",imagecopy)
                        i=i+1
```

Appendix D Lip reading Model

```
# 2. Buliding a Model
# declare input layer for CNN+LSTM architecture
video = Input(shape=(timesteps,img_col,img_row,img_channel))
# Load transfer learning model that you want
model = applications.mobilenet.MobileNet(input_shape=(img_col,img_row,img_channel), weights="imagenet", include_top=False)
model.trainable = False
# FC Dense Layer
x = model.output
x = Flatten()(x)
x = Dense(1024, activation="relu")(x)
x = Dropout(0.3)(x)
cnn_out = Dense(64, activation="relu")(x)
# Construct CNN model
Lstm_inp = Model(model.input, cnn_out)
# Distribute CNN output by timesteps
encoded_frames = TimeDistributed(Lstm_inp)(video)
# Contruct LSTM model
encoded_sequence = LSTM(256)(encoded_frames)
hidden_Drop = Dropout(0.3)(encoded_sequence)
hidden_layer = Dense(128, activation="relu")(encoded_sequence)
outputs = Dense(n_labels, activation="softmax")(hidden_layer)
# Contruct CNN+LSTM model
model = Model([video], outputs)
# 3. Setting up the Model Learning Process
# Model Compile
adam = tf.keras.optimizers.Adam(lr=Learning_rate, beta_1=0.9, beta_2=0.999, epsilon=None, decay=0.0, amsgrad=False)
model.compile(loss="categorical_crossentropy", optimizer=adam, metrics=["accuracy"])

# 4. Training the Model
hist = model.fit(X_train, Y_train, batch_size=batch_size, validation_split=validation_ratio, shuffle=True, epochs=num_epochs)
```

Appendix E Audio Model

```
from tensorflow.keras.layers import LSTM,Bidirectional,TimeDistributed

from tensorflow.keras.layers import Dense,Attention

# Neural network model


input_shape=(X_train.shape[1],X_train.shape[2])

optimizer = Adam(0.005, beta_1=0.1, beta_2=0.001, amsgrad=True)

n_classes = 80


model = Sequential()

model.add(Bidirectional(LSTM(256, return_sequences=True), input_shape=input_shape))

#model.add(Attention(636))

model.add(Dropout(0.2))

model.add(Dense(400))

#model.add(ELU())

model.add(Dropout(0.2))

model.add(Dense(n_classes, activation='softmax'))
```

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Appendix E Concatinating Audio Model with Video Model

```
import tensorflow

from tensorflow.keras import Sequential, Model

from tensorflow.keras.layers import Embedding, GlobalAveragePooling1D, Dense, concatenate

import numpy as np

from tensorflow import keras

model1=keras.models.load_model("C:/Users/leamlak/Desktop/close/saved_models/lipreading.hdf5")

model2=keras.models.load_model("C:/Users/leamlak/Desktop/close/saved_models/audio_recognition.hdf5")

print(model1)

print(model2)

model_concat = concatenate([model1.output, model2.output], axis=-1)

model_concat = Dense(1, activation='softmax')(model_concat)

model = Model(inputs=[model1.input, model2.input], outputs=model_concat)

model.compile(loss='binary_crossentropy', optimizer='adam')

model.summary()

model1.save("concatinate.hdf5")
```