Bilingual Speaker Recognition for Afaan Oromo and Amharic Language Speakers

Million Sime Demissie

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 Masters in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

Friday, September 1, 2022

Adama, Ethiopia

Bilingual Speaker Recognition for Afaan Oromo and Amharic Language Speakers.

Million Sime Demissie

Advisor-Teklu Urgessa (Ph.D.)

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 Masters in Computer Science and Engineering

Office of Graduate Studies
Adama Science and Technology University

Friday, September 1, 2022

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled "**Bilingual Speaker Recognition for Afaan Oromo and Amharic Language Speakers"** is my original work. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation.

Million Sime Demissie	_____	_____
Name of Student	Signature	Date

RECOMMENDATION

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled "**Bilingual Speaker Recognition for Afaan Oromo and Amharic Language Speakers**" prepared under my/our guidance by Million Sime Demissie and submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering. Therefore, I/we recommend the submission of a revised version of the thesis to the department following the applicable procedures.

Dr. Teklu Urgessa	_____	_____
Major Advisor	Signature	Date

APPROVAL SHEET OF M.SC. THESIS

I, the advisors of the thesis entitled "**Bilingual Speaker Recognition for Afaan Oromo and Amharic Language Speakers**" and developed by **Million Sime Demissie**; hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Dr. Teklu Urgessa	_____	_____
Major Advisor	Signature	Date

APPROVAL OF BOARD OF REVIEWERS

We, the undersigned, members of the Board of Examiners of the thesis by Million Sime Demissie have read and evaluated the thesis entitled "**Bilingual Speaker Recognition for Afaan Oromo and Amharic Language Speakers**" and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Reviewer 1	Signature	Date
_____	_____	_____
Reviewer 2	Signature	Date

Final approval and acceptance of the thesis are contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

_____	_____	_____
Department Head	Signature	Date
_____	_____	_____
School Dean	Signature	Date
_____	_____	_____
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGMENT

It is not easy to just acknowledge those who helped throughout all my ups and downs. God has been my ladder, my teacher my keeper in all of my path, he is there for me when I even forgot to call his name. I am so grateful, nothing would have been possible with your help almighty! I needed you in person so many times so you created mom to fill the gap in a reliable way. Mom I am nothing without you, you are my why to live. Next, I am glad to express my deepest gratitude to my advisor **Dr. Teklu Urgessa** for his supportive assistance, productive comments, criticism, and qualified advising till the completion of this thesis work. I should strongly appreciate his patience and democratic handling. I am very proud of working with you. I could get valuable life time lesson regarding research methods and life too.

My path was field with roses, those roses are not flowers that dies after a while but roses of life time. The special ones, my friends, beloved family, gossipers and my dear enemy's you all are the real reason to be who I am know. I am very happy to have you by my side. This research worth more than the time it takes, it was great way to feel scientist, it was a great reason to read the world of the scientific art therefore those who planned and participated in this assessment, supervision and advising deserve much respect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AM	Amharic
ASR	Automatic Speaker Recognition
CNN	Convolutional Neural Network
CSV	Comma Separated Values
EER	Equal Error Rate
HCI	Human Computer Interaction
LSTM	Long Short Term Memory
MFCC	Mel Frequency Cestrum Coefficient
MLP	Multilayer Perceptron
NN	Neural Network
OR	Afaan Oromo
PC	Personal Computer
PIN	Personal Identification
RNN	Recurrent Neural Network
ROC	Receive Operating Characteristics
SI	Speaker Identification
SV	Speaker Verification
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate True
TPR	True Positive Rate
VQ	Vector Quantization

ABSTRACT

Speaker recognition is identifying "who is talking?" it helps the machines to identify who is speaking. Identifying a speaker is useful to provide easy interaction between humans and computers. It also provides secure interaction since speech is a biometric identity verification tool. However, this is an easy task for humans but it is not for computer machines unless they are aided with deep learning models. Various studies have been done in this area and achieved promising results. Yet, speaker recognition is highly dependent on the language spoken and other factors. Therefore preparing a model to handle different languages is relevant, few studies have been done on Amharic speaker recognition these researches are done only in one language. This study prepared a deep learning bilingual speaker recognition model for two Ethiopian languages that have a vast number of speakers. The study was inspired by the current growth of voice assistance and call systems where many security issues are raised but remained unresolved, the scarcity of research in the field, and the interest of the researcher. In this study, an open set text independent bilingual Afaan Oromo and Amharic speaker recognition model was developed using MFCC by Librosa to extract speaker feature vectors with CNN, LSTM, and MLP. Keras model API is used to build the neural network. To train and test the model we prepared a new bilingual dataset that has a total of 960 speech samples by taking a 30-sec utterance of 16 different Afaan Oromo and Amharic speakers. The speech is taken carefully from YouTube by taking into account every similar condition of a speaker for both languages to reduce the effect of recording device variation, speaker mood, and speaker health status. In the end after training the model in the two languages, 95.93% average accuracy is achieved with MLP. The proposed model achieved 96.59% best accuracy when training and testing is done in Amharic and 95.45% trained and tested with only Afaan Oromo. The model resulted in 61.86% performance drop for cross lingual test. Therefore, this study provided a novel approach to model bilingual speaker recognition and presented a clear insight into the effect of language mismatch in the speaker recognition model and the necessity of bilingual speaker recognition.

Keyword: *Text-independent Speaker Recognition, CNN, MFCC, Feature Extraction, Keras Model API.*

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the Study

Speech analysis is the process of analyzing an audio signal to obtain relevant information from the signal in a form that is more compact than the audio signal itself. Speech signal levels have more complex variability in terms of amplitude, duration, pitch, timbre, and speaker variability Speech analysis has various related applications in speech activity detection, speech enhancement, speech recognition, text-to-speech, source separation, speech modification, emotional speech classification, keyword recognition, speaker diarization, and speaker recognition.

Speaker recognition is telling who is talking. It is one of the biometric tools that are unique to all humans. It is the task of identifying the owner of the given speech with the help of the unique behavioral and physiological characteristics of the speaker(Kabir et al., 2021). Every person has their way of speech, the way of uttering the words, the length it takes to say a single word, and the width and shape of the linguistic organs make one voice unique(Anukriti et al., 2013)(Furui, 1981).

It is helpful in various sectors including forensics to identify a criminal and personalization tools like accessing bank accounts and cellphones with voice. Mainly it is used when there is voice data. It is true that it is an easy task for the human mind but not for computers unless they are designed with the help of sophisticated algorithms like neural networks. Having the speaker recognized will help to use the identity of the owner for many purposes. Speaker recognition Applicable services include voice calls, using bank service via call dispatch, trading, database monitoring, information and reservation services, VOIP, security control for confidential information, and remote access to computers, as well as it is mostly used as a forensic tool.(Kabir et al., 2021) (Furui, 2009).

Speaker recognition has been studied for centuries using a different mechanism. So much research has been done to come up with the best solution since the first automatic speaker recognition (ASR) system came into existence in 1962 through an article by Lawrence G. Kersta, a Bell Laboratories physicist designated, "Voiceprint Identification"(Kabir et al., 2021).

Performance of speaker recognition is affected by many factors such as the health condition of the speaker, and language. Speaker recognition research that is done in a monolingual context using a single language for both training and testing have shown reliable result. Therefore, the first deficiency of the models that are prepared in other languages is they can't provide the same performance for another language. In a condition where bilingual service is expected those models cannot bring an efficient service. On the other hand, most of the research those have been done on the bilingual aspect of speaker recognition is done in tech-rich countries and highly resourceful languages.

1.2. Motivation of the Study

The faster growth of technology makes security the most required entity all over digital systems. Plenty of applications are coming to the computing world holding voice-dependent services since voice is a more convenient way than other HCI methods in terms of accessibility.(Clark et al., n.d.)(Hofmann, n.d.)(Zhang et al., 1999). Voice has a biometric quality of uniqueness in addition to its contactless accessibility therefore this makes it more preferable HCI method. In 2021, the global speech and voice recognition market was worth approximately $8.3 billion as per research by Markets and Markets[1] (published in August 2021). This will reach $22 billion by 2026 due to major advancements in AI systems(BasuMallick, n.d.). In many countries throughout the world, including Ethiopia, voice accessibility tools like speaker and speech recognition are helpful. To date research made on speaker recognition for Amharic is less than 6 as per our knowledge, while there is none yet in Afaan Oromo. There is none to mention for either bilingual or cross-lingual speaker recognition for Ethiopian languages. Thus, it tells us there is a lack of research on the local languages.

Speaker recognition is a process of identifying the person through their voice. Among plenty of factors that yield poor performance language is considered to be a major factor. To address the language variability issue a speaker recognition model designed for bilingual speakers is important. The same speaker might have a different type of speaking style for different languages. This leads to research on the identity of bilingual speakers and the real impact on the

[1]Revenue Impact & Advisory Company

recognition process. Therefore, this research aspires to come up with a best-performing deep learning model to recognize Afaan Oromo and Amharic speakers.

1.3. Statement of the Problem

Ethiopia is a multilingual country with over 80 distinct languages and with an estimated population size of more than 115 million[2]. Afaan Oromo and Amharic the two most spoken languages in the country cover the majority of the population. Where Afaan Oromo covers 33.80% Amharic speaker 29.10%. [3] Afaan Oromo is the mother tongue of the Oromo people who are the largest ethnic group in Ethiopia. It is the working language of the regional state of Oromia. It is also spoken in Kenya and Somalia.

Currently, the official language of Ethiopia is Amharic. Amharic is spoken across many areas of Ethiopia. Since it is the working language for the FDRE, it is used in all of the regions and city administrations. Amharic is also given as side subject in all primary and secondary schools across the country, so Amharic and Afaan Oromo bilingual speakers have considerable number.

Speaker recognition is important for a variety of applications, and it has a biometric value that makes it one's unique identity verification tool. In today's digital age, people can interact with their cell phones with their voices. They can verify their identity with the banker, they can call and make order on various services. All these interactions require the help of speaker recognition.

Currently speaker recognition field of study have showed a reliable performance with the adaptation of deep neural network(Kabir et al., 2021) Various studies have been carried out using ANN on a resourceful languages like English(Jones et al., 2022; Tadele et al., 2022). When it comes to a low resourced languages like Amharic and Afaan Oromo, this field of study has a huge gap of resource. Only 5 studies have been carried out in Amharic language speaker(Et. al., 2021)(A. D. Mengistu, 2017) and none for Afaan Oromo. Those studies have been carried out only on Amharic language.

[2]https://datacommons.org/place/country/ETH?utm_medium=explore&mprop=count&popt=Person&hl=en

[3]https://translatorswithoutborders.org/language-data-for-ethiopia

Automatic speaker recognition is affected by language variation(Matějka et al., 2017)(Tadele et al., 2022). A model prepared for a single language was unable to identify speakers for different language. Since people speaking two language can use them exchangeable, in any service area having a tool to operate in both language is more advantageous.

For humans it is easy to learn and identify who is speaking, to give this ability to computers there must be a reliable speaker recognition method. The computers must be capable of characterizing and learning the speech signals to make a reliable prediction. This can be possible with the help of a model like deep learning. The model must be fed with appropriate data to learn the speaker-dependent features and to make a reliable prediction.

One of the speaker-dependent features that impact the recognition performance of a speaker recognition model is language. If the monolingual trained model is used for bilingual users the model will not perform as expected. Many tech-rich languages have been occupied with plenty of research and models that arrived at a convenient level of accuracy, yet the adaptation of those models and sculpting a convenient model for Amharic and Afaan Oromo is necessary.

Generally, the problems are highlighted as

- The scarcity of research in bilingual ASR Afaan Oromo and Amharic languages.
- The dependency of ASR model performance on a language
- The need of ASR for HCI and security.

1.4. Research Questions

1) What is the impact of bilingual speakers compared on monolingual ASR models?
2) Which model is best to handle the bilingual Afaan Oromo and Amharic languages, with better performance?
3) What is the effect of cross-lingual tests on ASR models?
4) What is the effect of monolingual, and bilingual tests on the bilingual ASR model?

1.5. Objectives of the Study

1.5.1. General Objective

The General Objective of this study is to develop Afaan Oromo and Amharic bilingual speaker recognition using deep learning.

1.5.2. Specific Objectives

To meet the general objective of the study; the following specific objectives have been established:

- Review literature on the concepts of deep learning algorithms regarding speaker recognition
- Identify the source of data and collect voice data of bilingual speakers.
- Identify suitable deep learning algorithms for designing a speaker recognition model
- Build a text-independent bilingual speaker recognition model.
- To test and analyze the performance of the model in different experimental scenarios.

1.6. Delimitation of the Study

1.6.1. Scope of the Study

In Ethiopia there are more than 80 language spoken by different group of people. Those peoples are not only monolingual speakers but they are bilingual and multilingual. There are people who speak Tigrinya and Amharic, Wolaita and Amharic etc. However, the scope of this bilingual speaker recognition study is bounded to Afaan Oromo and Amharic speakers.

The speaker recognition state of the art has a wide range of family, like speaker diarization[4], robust speaker identification, segmentation, clustering, and detection and also it can be classified as text-dependent and text-independent. In this research only text-independent speaker verification is modeled.

Few researchers have tried to attach the speaker recognition task with the help of computer vision to design a pattern that identifies the speaker based on his facial look and voice for bilingual speaker recognition purposes, in this research we only considered the speech data.

The dataset for a speaker recognition study results huge impact on the result of the model. Having a large dataset and short dataset; as well as increasing the number of speakers in the dataset have their own distinct impact in speaker recognition model. Using the collection of speech's collected in a protected studio, speeches recorded in the wild, cell phone speech's[5]

[4] Speaker diarization is the process of dividing audio into segments belonging to each individual speaker.
[5] A survey on speaker recognition

results a realistic result. However this study gives less concern this issue, instead the speeches are collected from YouTube with certain criteria set purposively.

1.6.2. Limitation of the Study

Most of the limitations are raised from the time constraint to fill the resource and skill gap of the researchers additionally, since speaker recognition is a wide area of study addressing such a broad concept with a less man power is certainly difficult.

1.7. Significance of the Study

This study will help efforts to make technology accessible to the entire community in its unique way. For speaker recognition, numerous applications have been taken into account. To avoid identity theft and improve regional and national security against foreign assault. Today's constantly expanding technical security devices necessitate the adoption of trustworthy user authentication methods. Governmental, non-governmental, as well as individual sectors might benefit from speaker identification for it is good biometric system. Making this type of study on speaker recognition is sensible.

Speaker identification is a useful biometric system, therefore it may be used by a lot of businesses and people. The study can be used to help identify people in criminal justice and terrorist actions by forensic investigation professionals, courts, and police commissions. As an ultimate or ideal tool paired with speech recognition and enhancing the HCI mechanism. It is used to address the issues faced by physically challenged people. Therefore, this study is an excellent place to start to improve the likelihood that such a practical system would be created.

The study will introduce a new feature modeling and matching technique for Amharic language and Afaan Oromo bilingual speaker recognition development. As it attempts to use an ANN approach. As it mainly impacts the existing Ethiopian language speaker recognition model. An existing few types of research have been done with a single language model. In the meantime, it is an icebreaker for Afaan Oromo speaker recognition.

The study is significant hence it contributes to the advancement of research on bilingual speaker recognition. It will be a valuable contribution to the country having more than 80 languages. Thus it will allow upcoming researchers to compare and contrast the performance of this technique with others and work on the rest of the languages.

1.8. Thesis Organization

Five chapters make up the full thesis because each one is required to describe the main idea of the study. Therefore, the following is a description of the contents covered.

The first chapter is all about the introduction of the state of the speaker recognition including the significant scope and research questions. In this chapter, the actual problems are stated, and the interest and skill of the researchers are described. In the second chapter, we discussed the related works and existing literature in the problem domain and solution domain based on their direct relativity to bilingual speaker recognition. The literature reviewed based on their specific relativity on the given title as well as their year of publication for recent contribution has to be valued more and need more to be studied to come up with approval or disproval justification. In the third chapter, the methodology and tools for the proposed solution are discussed. It describes the methodologies and the procedures to be followed in doing the research. The definition of the tools consists in this chapter. In the fourth chapter the experiments, results, and discussion are briefly stated. The last and the fifth chapter's conclusion and recommendation conclude by briefing out the achievements, drawbacks, and future works to recognize and improve the work done by this specific research.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Overview of the Chapter

In this chapter, major ideas related to this specific research are discussed. The speech analysis and production, the state of art speaker recognition, and the biometrics essence of speaker recognition are discussed in brief. Nonetheless, it describes speaker recognition methods, speaker feature extraction techniques, and speaker modeling techniques. It will be a clear insight into the intended research work.

2.1.1. Biometrics

Biometrics is the science of uniquely identifying or verifying an individual among a set of people by exploring the user's physiological or behavioral characteristics(Guodong Guo, West Virginia University, et al., n.d.). In-depth biometrics is derived from the Greek terms bio, which means "life", and metrics, which means "to measure". Biometrics is the process of identifying or verifying a person based on physiological and behavioral traits. Several verifications or identity biometrics have evolved depending on several unique characteristics of the human body, simplicity of acquisition, public acceptance, and the level of security necessary. (Ashok et al., 2010)

2.1.1.1. Application of Biometrics

In today's world, when the majority of banking and other transactions are automated, security has become a major concern. Possessions (such as ID cards and keys) or secret knowledge are frequently used to provide security (like passwords, and PINs). This form of protection isn't foolproof because ID cards can be misplaced, and passwords can be forgotten or hacked. As a result, there was a considerable demand for more reliable authentication methods, and extensive research was conducted in this field. This gave rise to the idea of using human body parts or human mannerisms as security and authentication measures, and eventually to biometrics as a separate field. Biometric identification is now universally acknowledged as a requirement for any identification of a person.

2.1.1.2. Prerequisites for Good Biometrics

Any aspect of human physiology or behavior that can be accepted as a biometric should satisfy five properties described by Clarke(Ashok et al., 2010) which are as follows:

- Universality: Every person should have biometric characteristics.
- Uniqueness: No two persons should be the same in terms of the biometric characteristic
- Permanence: The biometric characteristic should be invariant over time.
- Collectability: The biometric characteristic should be measurable with some practical sensing device.
- Acceptability: The public should have no strong objection to the measuring or collection of biometrics. (Ashok et al., 2010)

2.1.1.3. Comparison of Biometric Methods

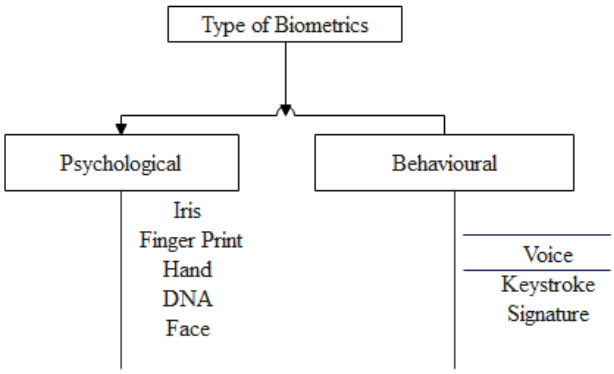


Figure 2-1 Types of Biometrics: Physiological and Behavioral

According to (Markowitz, 2000) speech, face, and signing up biometrics have high acceptability[6] and circumvention[7] ability compared to the rest. Although all of the biometric tools have their significant role and are unreplaceable comparably they have their limitation where one is accessible somewhere the others cannot. Speech has high accessibility since every human being communicates using voice. The use of voice is not limited to some specific location and condition. Services are delivered via speech communication including the voice assistance expert system that uses speech. Therefore having the comparison made by[15] as the benchmark

[6] The quality of being able to be reached or entered (translated by https://www.transate.google.com)
[7] The action of overcoming a problem or difficulty, typically cleverly and surreptitiously (translated by https://www.transate.google.com)

the biometric tools performed well in terms of their accessibility and can be outperformed by speech because of three major reasons. Table 2-1 depicts a comparison of speech biometrics with face and signing in regarding accessibility shows the comparison of speech biometrics with face and signing in terms of actuality.

Table 2-1 Comparison of Speech Biometrics with Face and Signing-In

Accessibility Conditions	**Speech**	**Face**	**Signing In**
Less Contact with the device	High	Medium	Low
Over cellular call	High	Low	Low
People with Visual impairment	High	Medium	Low

The three indicators taken for comparison are selected with common sense. Hence biometrics is derived from the biological characteristics of human beings we took into account the way people interact with those systems naturally. People using voice assistance systems make less contact with their device to provide themselves with the services that can be given by face and signing in. over a cellular call authentication with face and signing in to service is impossible, also people with visual impairment can get access to their digital tools with help of voice rather than face and signing in. Furthermore, speech is a preferable HCI tool because it enables end-to-end communication and interaction with digital systems.

2.1.1.4. Voice Biometrics

Voice-biometrics is a speech processing technology or procedure that uses voice to validate a person's identity in an unobtrusive and undetectable manner. Speech processing and biometric security are the two businesses that voice-biometrics systems belong to. To do their job, voice biometrics, like other speech-processing systems, take data from the stream of speech. They can be set up to work on many of the same acoustic parameters as voice recognition, their closest speech-processing relative. Voice biometrics systems rely on speech recognition to determine what a person is saying. The most common applications of biometrics-based technologies are security monitoring and fraud prevention. There are various types of voice biometrics, such as speaker verification, speaker identification, and speaker recognition. (Markowitz, 2000)

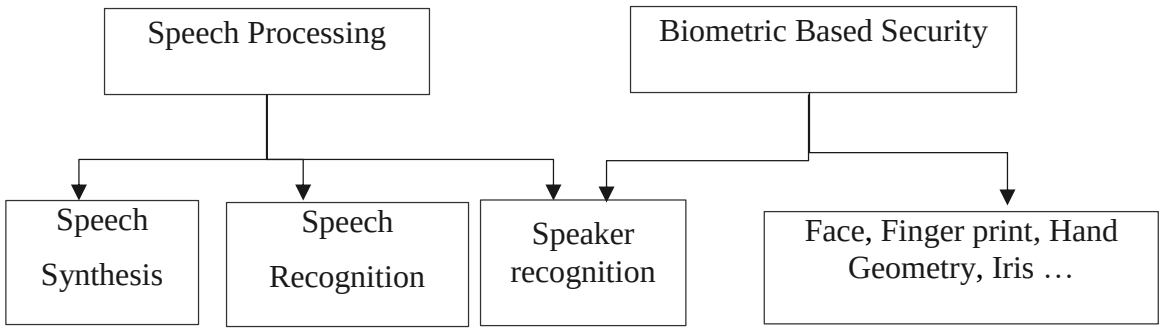


Figure 2-2 Voice Biometrics Overview

2.1.2. Language

Language is a means of communication. It is a means of conveying our thoughts, ideas, feelings, and emotions to other people. Jack C. Richards and Richard Schmidt define language: as "the system of human communication which consists of the structured arrangement of sounds (or their written representation) into larger units, e.g. morphemes, words, sentences, utterances. In common usage, it can also refer to non-human systems of communication such as the “language” of bees, the “language” of dolphins”.(Russo et al., 2018)

Writing letters or notes to the delivery man, chattering with friends, giving speeches, and even talking to ourselves in the mirror are just a few of the many ways we utilize language. If you give it some thought, there are a few repeating purposes that language generally fulfills despite the numerous varied purposes we give it. Others are loftier and nearly abstract, while some appear to be so routine as to virtually go overlooked as functions. But it's crucial to understand that they are all equally significant linguistically. No matter what societal value we may assign to different functions, language itself does not make any distinctions. Lyons refers to the study of meaning that is systematically encoded in the lexicon and grammar of the natural language as "linguistic semantics." Thus, linguistic semantics is a subfield of linguistics; according to Lyons, philosophical semantics is the more appropriate subfield for semantic problems that are more philosophical in nature. For example, languages can be used to release nervous/physical energy which is necessary for physiological. To scale, to create a record (recoding function), to identify and categorize objects (identifying function), as a tool for cognition (reasoning function), and as a way of exchanging thoughts and emotions giving pleasures is the ideational function.(Russo et al., 2018)

2.1.2.1. Amharic

A subset of the Semitic family of Afro-Asiatic languages, Amharic is one of the Ethiopian Semitic languages. The language is used as the official working language of Ethiopia and a number of the states that make up the federal structure of Ethiopia. Amharic is spoken by 21.6 million native speakers and 4 million secondary speakers in the country, according to the 1999 census. After Arabic, Amharic is the Semitic language that is most widely spoken worldwide. The language is also spoken by 3 million emigrants from outside of Ethiopia. Amharic is one of the six non-English languages recognized in Washington, DC, and is now used for government services and educational purposes. Amharic is a subject that is taught in schools all around the nation. To train Amharic teachers, journalists, and public relations professionals, many universities also offer Amharic departments. Amharic is being used by the country's press, radio, and television broadcasts to disseminate information to the general people(*A Typology of Verbal Derivation in Ethiopian Afro-Asiatic Languages*, n.d.).

2.1.2.2. Afaan Oromo

The Afaan Oromo is a language spoken by Oromo. Their language is known as Afaan Oromo, which translates to "Oromiffa". In Oromia region, the language has been adopted since 1991 as the official language, the language of instruction, the language of the courts, and the language of business(*A Typology of Verbal Derivation in Ethiopian Afro-Asiatic Languages*, n.d.). The most frequently used of the Cushitic languages is Afaan Oromo or Oromiffa, which is an Afro-Asiatic tongue. More than 40 million Oromo and neighbors in Ethiopia and Kenya speak it as their mother tongue(Jalata, 2010). 95% of Afaan Oromo speakers live in Ethiopia, mostly in the Oromia region (*Afaan Oromo*, n.d.). Additionally, it is significant to remember that both Oromo and non-Oromo individuals speak Afaan Oromo as a second language. There are a small number of speakers of the language in Somalia.

Currently Afaan Oromo is thought in various higher institutions including Addis Ababa University, to the doctoral and masters level. It is also a medium of learning in elementary schools available in Oromia

2.1.2.1. Impact of Language on Speaker Recognition

Many studies have shown that language is one of the factors to reduce the performance of speaker recognition models. The nature of the languages as well as the difference in utterance

and energy level when they create a sound have a great impact on the state of art. A study performed using OGI bilingual speaker recognition corpus, showed that the performance of the speaker recognition model is affected highly greatly by language variance, with equal error rates differing by more than a factor of three between the worst and best-performing languages(Kleynhans & Barnard, 2005) the effect of language on speaker recognition performance is also known as language familiarity effect.

Language is a barrier because Findings revealed that bilingual speakers produced notably different voice patterns in their two spoken languages across different speech tasks when vocal measures were explored. The difference implies that language alone is an acquired factor that contributes to the manifestation of within- and between-speaker variability of vocal attributes. In addition, results indicate that selection of speech samples can be a crucial factor in differentiating acoustic profiles, supporting previous literature that task effects emerge in acoustic measures within individuals. Language can be classified as monolingual, bilingual and cross lingual.

Monolingual: - a person who speaks only one language[8]

Bilingualism: - Bilingualism is fluency in use of two languages. Bilingualism, Ability to speak two languages. It may be acquired early by children in regions where most adults speak two languages example like one person speaking Afaan Oromo and Amharic[9].

Cross-lingual: - is of or relating to languages of different families and types especially, relating to the comparison of different languages[10].

2.1.3. Speech

Speech is the most common and natural form of communication. It is a toll to transmit information from one to another. Speech analysis is concerned with methods of processing that are intended to draw data from the speech waveform. It is the main responsibility of a speaker verification procedure. A key component of speech analysis is an understanding of the

[8] https://www.merriam-webster.com/dictionary/crosslinguistic
[9] https://www.britannica.com/topic/bilingualism?utm_source=pjaffiliate&utm_medium=pj&utm_campaign=kids-pj&clickId=4089542175
[10] https://www.merriam-webster.com/dictionary/crosslinguistic

phonetic[11] properties of human sound production. The act of generating articulated sounds or words is known as speech production. Speech is created by transporting air from the lungs to the larynx (respiration[12]), where the vocal folds may be held open to allow the air to flow through or may vibrate to produce a sound (phonation)

Speech creation does not start I lugs, it is created I our brain. After creation of the message and the lexico-gramatical structure in our mind we need a representation of the sound sequence and number of commands which will be executed by our speech organs to produce the utterance(Trujillo, 1994) The majority of speech is created by an air stream that rises from the lungs through the trachea (the windpipe) through the oral and nasal cavities. The many organs of speech modify the air stream as it passes through. Each of these modifications has unique acoustic effects that are utilized to differentiate sounds. The production of a speech sound can be divided into 4 distinct but interconnected processes: the ***initiation*** of the air stream, normally in the lungs; its ***phonation*** in the larynx via the operation of the vocal folds; its ***direction*** by the velum into either the oral cavity or the nasal cavity (the oro-nasal process); and finally its ***articulation*** in the oral cavity, primarily by the tongue.

The ***initiating*** process occurs when air is evacuated from the lungs. Speech sounds in English are produced by "a pulmonic aggressive air stream"(Trujillo, 1994)(*Rosistem Barcode - Biometric Education*, n.d.) albeit this is not true in other languages (ingressive sounds).

The larynx is where the ***phonation*** process takes place. The vocal folds are two horizontal folds of tissue in the flow of air through the larynx. The glottis is the space between these folds. The glottis can be shut(*Rosistem Barcode - Biometric Education*, n.d.). Then no air can pass through. It might also have a tiny aperture that causes the vocal folds to vibrate, resulting in "voiced noises." Finally, it can be wide open, as in normal breathing, which reduces the vibration of the vocal folds, resulting in "voiceless noises." The air might enter the nasal or oral cavities after passing via the larynx and throat. The velum is the component in charge of this selection. The oro-nasal process allows us to distinguish between nasal consonants (/m/, /n/, //) and other sounds.(Trujillo, 1994)(*Rosistem Barcode - Biometric Education*, n.d.)

[11] Phonetics is the study of human sounds and phonology is the classification of the sounds within the system of a particular language

[12] respiration is the movement of oxygen from the outside environment to the cells within tissues, and the removal of carbon dioxide in the opposite direction that's to the environment

2.1.3.1. Speech Sound Types

Speech sounds can be classified as voiced and unvoiced Voiced speech signal is characterized by deterministic acoustic waveforms, while unvoiced speech signal corresponds to stochastic waveforms. Voiced sounds are produced by forcing air through the glottis or an opening between the vocal folds. The vocal cord tension is altered so that they vibrate in an oscillatory pattern. The periodic stoppage of subglottal airflow causes quasi-periodic puffs of air to stimulate the vocal tract. The larynx produces sound, which is known as voice or phonation. The larynx stimulates voiced spoken sounds on a regular basis (Trujillo, 1994)

Unvoiced noises are produced by creating a constriction anywhere in the vocal tract and forcing air through it to create turbulence-like noise. A speaking sound can be uttered and constricted at the same time (mixed). Voiced speech sounds have garnered greater scientific interest than unvoiced speech sounds. This is owing to research on speech perception indicating that precise modeling of spoken speech is critical for natural-sounding speech.(Trujillo, 1994) (Rashid, n.d.)

The output of a linear, time-varying system triggered by either quasi-periodic pulses (during voiced speech) or random noise (during unvoiced speech) is used to represent speech sound waves(Furui, 2009).

A voiced speech signal y(n) is made up of a rapidly variable portion x(n) (excitation sequence) and a slowly varying part h(n) (vocal system impulse response). As in Equation 1

$$y(n)=x(n)* h(n)$$

Equation 1

where * represents the linear convolution operator This is a time-domain representation of a spoken speech frame that ends at time n. Using the pattern of the vocal system impulse response characteristics h(n), the claimed speaker is validated(Proakis, 2000; *Rosistem Barcode - Biometric Education*, n.d.; Trujillo, 1994). As a result, the speaker verification procedure is centered on h(n). However, the convolution makes it difficult to distinguish between the two portions x(n) and h(n)(Rashid, n.d.)(Trujillo, 1994) As a result, isolating them is a critical stage in the speaker certification process.

2.1.4. Automatic Speaker Recognition

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by

people accessing systems; that is, it enables access control of various services by voice. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers.

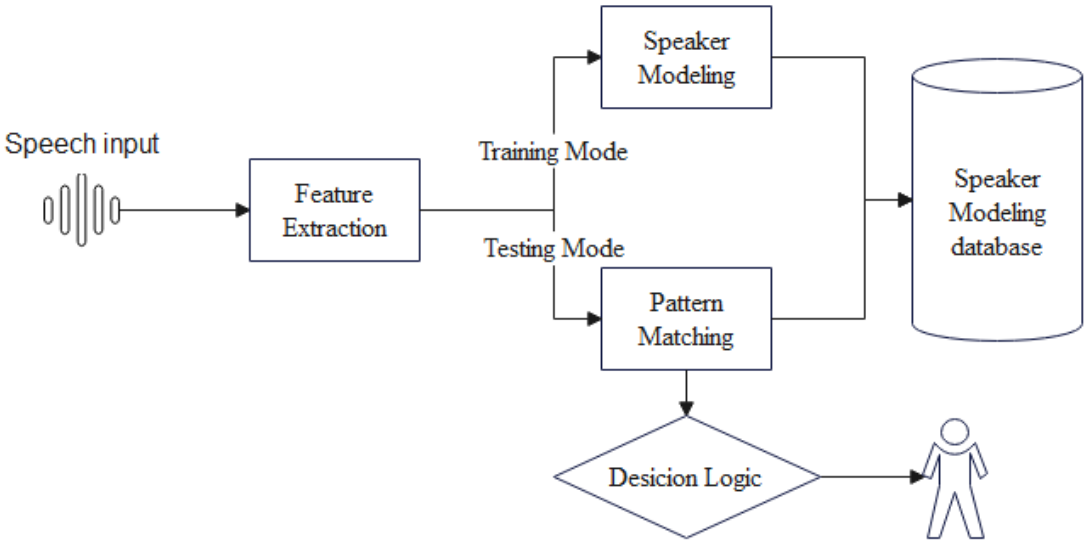


Figure 2-3Overview of ASR

2.1.4.1. Variants of Automatic Speaker Recognition

Based on the implementation technique and the depth of its functionality speaker recognition can be classified into various aspects. Text-dependent and text-independent categorization is given based on if the proposed solution is dependent on a specific word utterance. Like saying the same word for training and testing makes a speaker recognition model text-dependent or otherwise independent. On the other hand, if a speaker recognition is considering languages the categorization falls under cross-lingual, monolingual, and bilingual. Generally speaking, speaker recognition is classified as follows

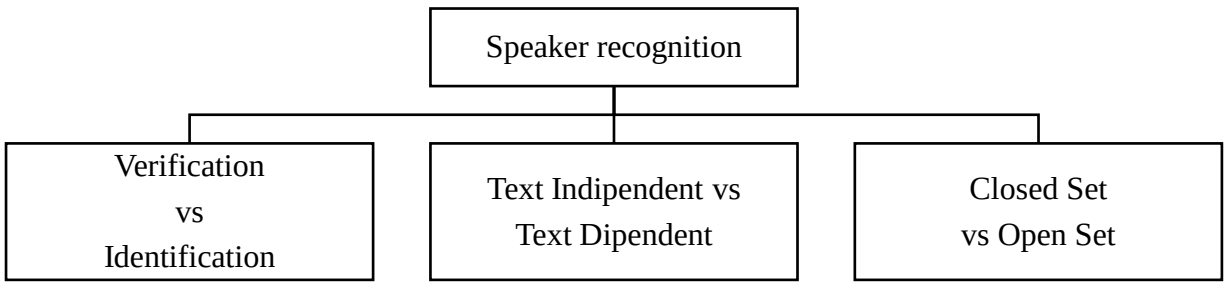


Figure 2-4 Variants of Speaker Recognition

Closed Set and Open Set

In the closed set problem, the speaker with the highest degree of similarity to the reference template among N known speakers is chosen. The input speech sample template of an unknown speaker is acquired. It's assumed that this unidentified speaker is one of the groups of speakers As a result, in a closed set problem, the system makes a compulsion decision by selecting the best match speaker from the database of speakers in the open set for an unknown object, identification, and a matching reference template are required(Kekre & Kulkarni, 2013).

Text-dependent vs Text-independent

The text-dependent speaker recognition uses some specific word utterances both for training and testing purposes. In that case, any system built with this scenario is dependent on similar word utterances. On the contrary, text-independent speaker recognition broadens its scope to unlimited word utterances while it try to depend on the person's physical and behavioral characteristics.

Identification vs Verification

Speaker Identification is a task of identifying a given speaker among set of speaker listed out in the database. The system has to check if there is a correct much of a given voice sample and has to identify who exactly is that person. While in speaker verification the system has to check only between the given voice sample and the suspected sample voice in the database. In speaker verification, the check works on a one-to-one basis while in speaker identification it works on many bases. Speaker verification uses voice samples to determine whether a person is whom she or he claims to be(Markowitz, 2000). The huge difference between speaker identification and verification is speaker identification has to tell who the speaker is, It is among the list of many known speakers in the database speaker identification is asked to figure out "To who does this voice belong?" and the SI is expected to reply "it is Mr. X" while the task of speaker verification is to say yes or no. It is speaker verification is asked if the person is Mr. X and it repays "Yes it is" or "No it is not"

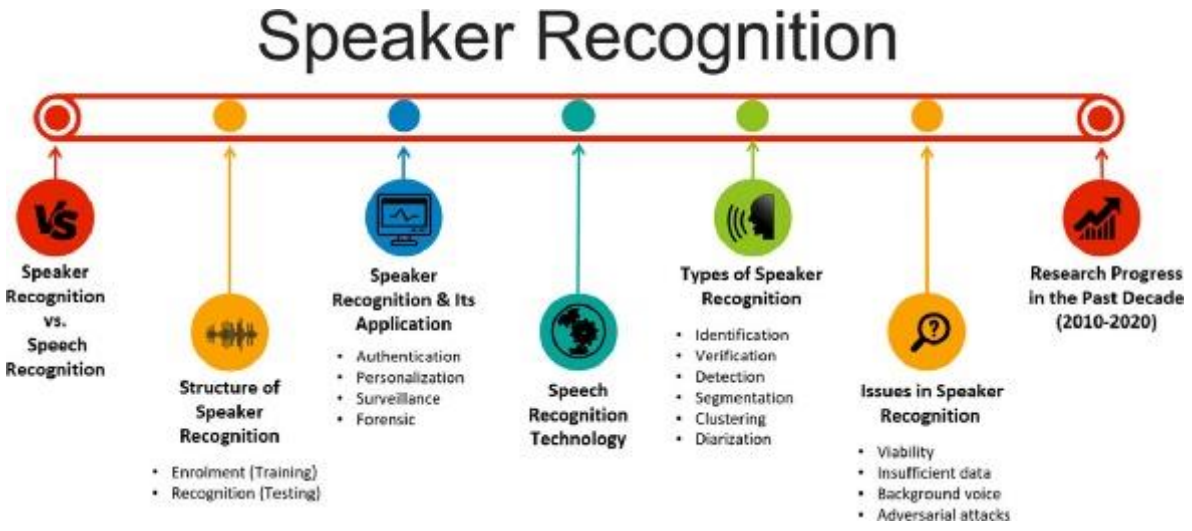


Figure 2-5 Overview of Speaker Recognition (Mohd et al., 2021)

2.1.4.2. Applications of Speaker Recognition

Speaker recognition is hardly known for its biometric identification purpose, as such, it is mostly used as a forensic tool. However, it has several applications including authentication, surveillance, forensic speaker recognition, security, speech recognition, multi-speaker tracking, and a personalized user interface.

- Forensic: speaker recognition plays a vital role in forensics to identify a target criminal with the help of voice samples as evidence. This is helpful when crimes a recommitted and there is voice evidence available.
- Mobile banking: Using a mobile telephone network, mobile banking enables customers to perform an automatic banking transaction or act in addition to the regular instructions of conducting banking operations remotely. It serves as the fundamental financial system's delivery channel. Through a mobile phone and voice authentication, the consumer is given secure access to his or her bank account.
- System for legal and judicial cases: For a user of the system to remotely access his or her court case through telephone validated by speech and learn about other court decisions and developments, speaker verification technology can be added to the Legal and Court Cases system.
- Password-free OS/Application Login: The speaker verification method can be utilized as authentication for login to an operating system or an application instead of remembering long, complex passwords or carrying tokens.

- On-site Entrance Control: A method for confirming speakers can be used to show identification before entering any physical institutions, such as those used for home detention, by storing an embedded speaker model chip.
- Shopping: To verify an online purchase, a speaker verification system might be utilized. Mobile phone shopping by customers
- Personalized user interface: recent AI applications learn user preferences by adopting their voice. Thus technologies are growing fast with voice assistance systems like (Siri, Alexa, Cortana… etc.)

2.1.5. Feature Extraction

Speech is a sophisticated motor skill that humans naturally develop. Adults who have it may produce 14 distinct sounds per second using the coordinated movements of around 100 muscles(Rashid, n.d.). By converting the speech waveform to a parametric representation at a comparatively low data rate for further processing and analysis, features are extracted.(Rashid, n.d.) Speech signal has various determining features the first feature is short-time energy, short-time zero crossing rate, and short-time autocorrelation in time-domain speech signals. Features of the frequency domain including the spectral centroid and spectral flux Pitch, stress, power spectral density, vowel length, rhythm, and intonation patterns are further aspects of speech signals. These qualities play a key role in meaningful and recognizable speech segmentation. Autocorrelation, short time zero crossing energy, and speech signals are three of the most crucial qualities(Rahman, 2012)(Poornima, 2016). Therefore, exceptional and quality attributes lead to the designation of anything as acceptable. The computation of a series of feature vectors in automatic speech recognition (ASR) results in a compact representation of the input speech signal. There are typically three basic stages to it. The first stage is known as the speech analysis or the acoustic front-end, which analyzes the speech signal's spectra and temporal dynamics and produces raw characteristics that describe the power spectrum's envelope for brief speech intervals. An extended feature vector made up of both static and dynamic features is assembled in the second stage. These extensive feature vectors are then converted in the last stage into more manageable and reliable vectors, which are subsequently provided to the recognizer. The speech feature extraction methods include Discrete Wavelet Transform (DWT), Line Spectral Frequencies (LSF), Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction

Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), and Perceptual Linear Prediction (PLP).

MFCC: Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed short time signal derived from the Fast Fourier Transform (FFT) of that signal (Dave, 2013)

LPC: the LPC calculates a power spectrum of the signal. It is used for formant analysis. LPC is one of the most powerful speech analysis techniques and it has gained popularity as a formant estimation technique(Dave, 2013)

PLP: the Perceptual Linear Prediction PLP model developed by Hermansky. PLP models the human speech based on the concept of psychophysics of hearing [2, 9]. PLP discards irrelevant information of the speech and thus improves speech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system(Dave, 2013)

LPCC: Linear predictive cepstral coefficients are linear predictive coding coefficients presented in cepstral domain. For evaluating the fundamental parameters of a speech signal, LPCC has become one of the predominant techniques. A cepstrum is the inverse Fourier transform of the estimated spectrum of the signal. Linear Predictive Coding is used to obtain the LPC coefficients from the speech tokens(Kaur et al., 2017)

Spectral Centroid: it is measure of the brightness of sound. The measure is obtained by evaluating the center of gravity.

2.1.6. Deep Learning

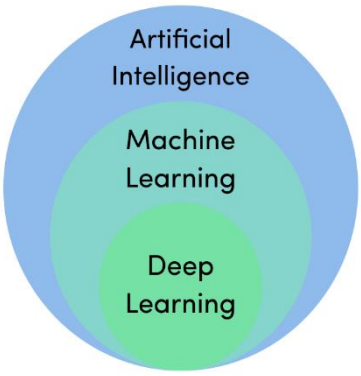


Figure 2-6 Hierarchy of Deep Learning in Artificial Intelligence

A subset of machine learning called deep learning recently outperformed machine learning in processing unstructured data. Current machine learning methods are outperformed by deep learning methods. Deep learning makes it possible for computer models to gradually learn features from data at various levels. It offers solutions for practically all types of data, including text, audio, and photos. To produce outcomes in a way that is similar to what the human mind does, neural networks attempt to imitate the brain.

Computational models with numerous processing layers can learn representations of data at various levels of abstraction thanks to deep learning. These techniques have significantly advanced the state-of-the-art in many other fields, including drug discovery and genomics, speaker recognition, visual object recognition, object detection, and many others. By employing the backpropagation technique to suggest changes to a machine's internal parameters that are used to compute the representation in each layer from the representation in the previous layer, deep learning can uncover detailed structure in massive data sets. While recurrent nets have shed light on sequential data like text and voice, deep convolutional nets have made advancements in the processing of pictures, video, speech, and audio. (Lecun et al., 2015)

2.1.6.1. Neural Network Architecture

Neural Networks are complex structures made of artificial neurons that can take in multiple inputs to produce a single output. This is the primary job of a Neural Network – to transform input into a meaningful output. Usually, a Neural Network consists of an input and output layer with one or multiple hidden layers within.

In a Neural Network, all the neurons influence each other, and hence, they are all connected. The network can acknowledge and observe every aspect of the dataset at hand and how the different parts of data may or may not relate to each other. This is what Neural Networks are capable of There are two methods in which information moves via a neural network:

Flow-forward networks: Signals in this paradigm only move in one direction, that of the output layer. With zero or more hidden layers, feedforward networks have a single output layer and an input layer. The use of pattern recognition is widespread.

Reaction Networks: In this approach, the recurrent or interactive networks process the input sequence using their internal state (memory). They have loops (hidden layer(s)) in the network that allows for signal transmission in both directions. They are frequently applied to sequential and time-series jobs. Finding extremely complex patterns in vast volumes of data.

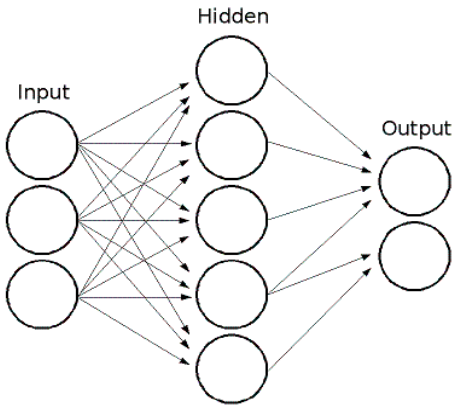


Figure 2-7 Components of Neural Network.

The learning (or training) process in a neural network is started by separating the data into three groups:

- Training dataset - With the help of this dataset, the neural network can comprehend the weights that connect its nodes.
- Validation dataset - This dataset is used to optimize how well the neural network performs.
- Test dataset - The accuracy and margin of error of the neural network are evaluated using this dataset.

Neural network algorithms are then applied to them for training the neural network after the data has been divided into these three categories. The process of facilitating training in a neural network is known as optimization, and the optimizer is the method that is applied. There are various optimization algorithm types, each with particular properties and features including memory needs, numerical accuracy, and processing speed.

Some of the Deep Learning core Neural Networks are the following.

2.1.6.2.Multi-Layer Perceptron

Feed-forward neural networks are supplemented by multi-layer perceptron's. There are three different kinds of layers in it: an input layer, an output layer, and a concealed layer. The input layer is where the input signal for processing is received. The output layer completes the necessary task, such as classification and prediction. The real computational engine of the MLP consists of an arbitrary number of hidden layers that are sandwiched between the input and output layers. Data flows from the input to the output layer of an MLP in the forward direction, much like a feed forward network. With the help of the back propagation learning algorithm, the MLP's neurons are taught. MLPs can resolve issues that are not linearly separable because they are made to approximate any continuous function. Pattern categorization, recognition, prediction, and approximation are the main applications of MLP. (Pethuru Raj, n.d.)

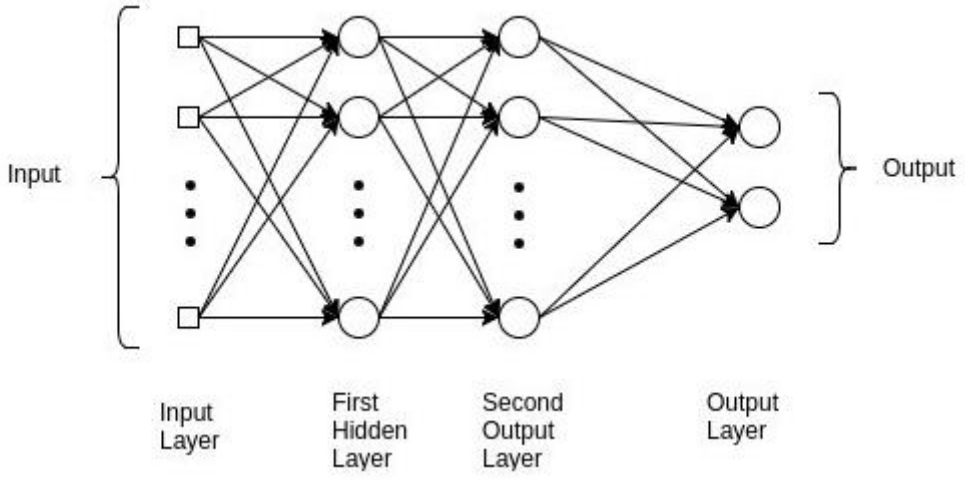


Figure 2-8 the Structure of A Multilayer Perceptron Neural Network (Faghfouri & Frish, 2011)

2.1.6.3. Radial Basis Network

Radio frequency network The Radial Basis Function Network, or RBFN for short, is a type of neural network that relies on integration to solve non-linear classification issues. RBFNs differ from standard multilayer perceptron networks in that they do not just take an input vector and multiply it by a coefficient before summing the outcomes. Instead, RBFNs evaluate each input vector, compare the input to the stored training value, and generate a similarity score using Radial Basis Function Neurons. The sum of each similarity value after being multiplied by weights is then the output layer. (*Radial Basis Function Networks Definition _ DeepAI*, n.d.).

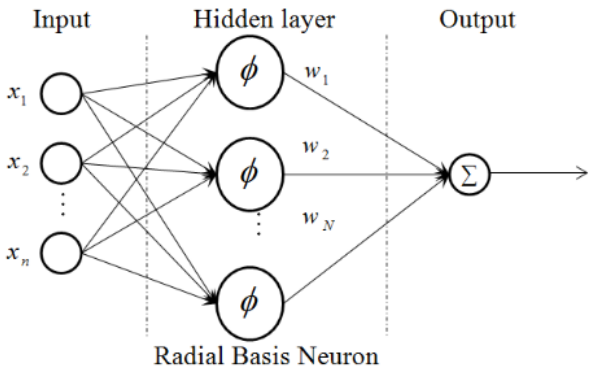


Figure 2-9 Architecture of the RBF Neural Network. (He et al., 2019)

2.1.6.4. Back propagation neural network

A supervised learning technique for multilayer feed-forward networks in the field of artificial neural networks is the backpropagation algorithm. The information processing of one or more neurons, also known as neural cells, serves as the model for feed-forward neural networks. A neuron's dendrites receive input signals before transmitting them to the cell body. The signal is sent from the cell's axon to synapses, which are where one cell's axon connects to another cell's dendrites. The backpropagation approach's fundamental idea is to simulate a given function by altering the internal weightings of input signals to generate an anticipated output signal. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is presented to the system and used to modify its internal state. In a multilayer feed-forward neural network, the backpropagation algorithm is a technique for training the weights. As a result, it necessitates the definition of a network structure consisting of one or more levels, each of which must be fully connected to the previous layer. One input layer, one hidden layer, and one output layer make up a typical network topology.

The network performs best in classification challenges when it has one neuron in the output layer for each class value. A two-class or binary classification issue, for instance, with the class values A and B. It would be necessary to convert these anticipated results into binary vectors with a column for each class value. For example, for A and B, respectively, [1, 0] and [0, 1]. It's known as a one-hot encoding. (Brownlee, n.d.)

2.1.6.5. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN or ConvNet) are sophisticated feed-forward neural networks used in machine learning. Because of its great accuracy, CNNs are utilized for picture categorization and recognition. Yann LeCun, a computer scientist, first proposed it in the late 1990s after becoming intrigued by how humans recognize objects visually. The CNN uses a hierarchical model that constructs a network in the shape of a funnel before producing a fully connected layer where all the neurons are connected and the output is processed.

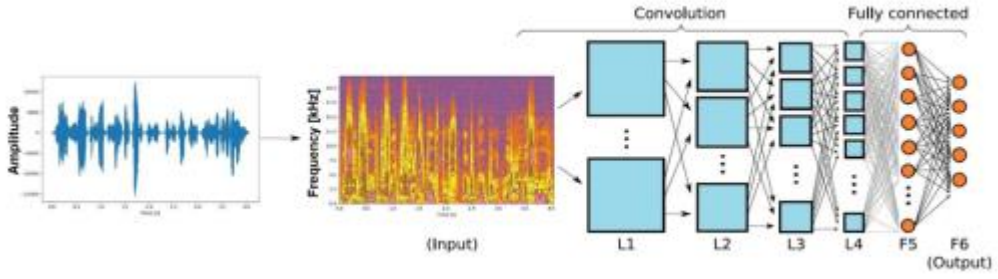


Figure 2-10 Generic Representation of the CNN Architecture (Franti Et Al., 2017)

2.1.6.6. Application of Deep learning

In the past few years, Deep Learning has become a trend. Since deep learning attempts to make better analyses and can learn massive amounts of unlabeled data, deep learning has been applied to several fields. Since 2006, deep structured learning, more commonly called deep learning or hierarchical learning, has emerged as a new area of machine learning research. Automatic Speech Recognition, Image Recognition, Natural Language Processing, Drug Discovery, and Toxicology, Customer Relationship Management, Recommendation Systems, Bioinformatics, and speaker recognition(Hordri et al., 2016).

2.2. Related Works

(A. Mengistu & Alemayehu, 2017) the first attempted Amharic language speaker recognition research. It presents a text-independent speaker identification system for the Amharic language. Speech signals are collected from different speakers including both sexes as well as different age groups. MFCC had been used to extract features from the speech signals and to generate feature vectors. VQ and GMMs had been used for training and identification purpose. The researcher intended to see which modeling approach is better for text-independent speaker identification. For a total of 50 speakers, 74.2% accuracy was achieved when the VQ approach is used whereas 84.3% accuracy for the GMMs.(A. Mengistu & Alemayehu, 2017)

(A. D. Mengistu, 2017)provides a technique for automatically identifying speakers of the Amharic language in noisy settings without the use of text. 100 distinct speakers of both genders' speech signals are collected. Each individual's speech lasts for 10 seconds. For feature extraction, an MFCC, LPCC, and GFCC combination had been applied. VQ and GMMs have been employed for identification and training purposes. By combining the three feature extraction strategies, the researcher was seeking to determine which modeling strategy is more effective for text-independent speaker identification (MFCC, LPCC, and GFCC). In the first experiment, the researcher used a VQ modeling strategy with a combination of three feature extraction strategies for speakers who were 30, 60, and 90 in number. And separately obtained 77.2%, 70.9%, and 69% accuracy. Two, a GMM modeling strategy combining the three feature extraction methods for speakers of 30, 60, and 90. and obtained an accuracy of 75.2%, 76.9%, and 78%, respectively. (A. D. Mengistu, 2017)

(Luengo et al., 2008) They investigated the benefits of combining short-term intonation and energy information with conventional MFCC features to create a more effective language-independent parameterization for speaker recognition to reduce the error rate of speaker recognition in language mismatch conditions. The tests made advantage of a fresh Basque-Spanish bilingual speech database. 22 bilingual speakers—11 men and 11 women—were recorded in a silent setting for the database. Each speaker wrote down seven eight-digit numerical sequences that they read whatever they liked. Finding the spectrum parametrization and identifying language-dependent aspects of the speaker are made possible by attaching prosodic features to the current Mel frequency cepstrum in each numerical sequence recorded

during the session. When the model is taught in Spanish and tested with Basque, the EER increases from 6.34 to 8.25, and from 6.82 to 8.77 when it is trained in Spanish and tested with Basque.

(Agrawal et al., 2009) Describe a technique that uses a back propagation neural network to identify speakers in many languages (BPA). Created a dataset of 32 people who speak 8 different regional Indian languages. There were 904 total utterances. For data clustering, they employed the Fuzzy C-means (FCM) algorithm. The fundamental characteristics of speech, such as average power spectral density, Cestrum coefficient, and the number of zero crossings, are retrieved using MATLAB. Each word in a voice utterance is taken from a sentence in a specific language during preprocessing. Using BP in a text-dependent context, the models are tested using the same language for both training and testing, yielding 95.4% accuracy. (Pandey et al., 2010)

(Nagaraja & Jayanna, 2013) Using 30 randomly chosen speakers from the IITG Multi Variability Speaker Recognition Database, which is collected in a setup with five different sensors, two different environments, two different languages, and two different styles, the monolingual, cross-lingual, and bilingual speaker identifications were carried out. The 16 kHz sampling rate and 16-bit storage resolution were used for the voice signal. The recording was done in Indian English and the speaker's preferred language, which may be one of the Indian languages such as Hindi, Kannada, Tamil, Oriya, Assami, Malayalam, and so on. By combining the evidence from LPR, LPRP, and multitaper MFCC features, it was attempted to improve bilingual speaker recognition in limited data conditions. Utilizing UBM-GMM, speaker modeling is carried out. The findings demonstrated that when combined with LPR/LPRP, the data from the multitaper MFCC offers good speaker identification capability.

(Faundez-zanuy et al., 2014) They have conducted four trials testing and training with Catalan, testing and training with Spanish, training with Spanish and testing with Catalan, and training with Catalan testing with Spanish employing two languages, Catalan and Spanish, to study speaker recognition. There have been two speaker recognition techniques used: One codebook is generated for each speaker using vector quantization and the LBG [5] technique, and the codebooks range from 0 to 8. 2. The calculation of a covariance matrix is implied by the arithmetic-harmonic measure. 49 bilingual readers, reading the same text continuously for

around one minute. With quiet removed, five separate test sentences that are representative of all the speakers were employed. Hamming window and covariance metrics performed better than vector quantization.

(Rozi et al., 2017) These researchers attempted to design bilingual speaker recognition with the awareness of every language as initially. They Used Probabilistic Linear Discriminative Analysis (PLDA). The proposed approach has been evaluated with i-vector/PLDA framework using the CSLT-CUDGT2014 Chinese-Uyghur bilingual speech database. The experimental results show vector/PLDA framework using the CSLT-CUDGT2014 Chinese- Uyghur bilingual speech database. The experimental results show that the language PLDA training brought in a reduction of 15.0% EER in Chinese and 20.00% in the Uyghur test. They used language awareness to their advantage.

(Xia et al., 2019) They introduced an unsupervised Adversarial Discriminative Domain Adaptation (ADDA) to increase the robustness of the cross-lingual speaker recognition system using unlabeled data by reducing the human labeling cost. An unsupervised Adversarial Discriminative Domain Adaptation (ADDA) method to effectively learn an asymmetric mapping that adapts the target domain encoder to the source domain, where the target domain and source domain are speech data from different languages. ADDA, together with a popular Domain Adversarial Training (DAT) approach, are evaluated on a cross-lingual speaker verification task: the training data is in English from NIST SRE04-08, The ADDA method as they mentioned helps to effectively learn an asymmetric mapping that adapts the target domain encoder to the source domain, where the target domain and source domain are speech data from different languages. Showed that with the ADDA adaptation, the Equal Error Rate (EER) of the x-vector system decreases from 9.331% to 7.645%, a relatively 18.07% reduction of EER, and a 6.32% reduction from DAT as well.

(Thienpondt et al., 2020) Describe the top-scoring IDLab submission for the text-independent task of the Short-duration Speaker Verification (SdSV) Challenge 2020. We introduce domain-balanced hard prototype mining to fine-tune the state-of-the-art ECAPA TDNN x-vector-based speaker embedding extractor. They presented HPM as a computationally efficient hard negative mining strategy to fine-tune a speaker embedding extractor towards out-of-domain Farsi data.

(Nawaz et al., 2021) The main aim of this paper is to answer two questions one is “is voice-face association language independent” and “can a speaker be recognized in respective spoken language?” because these two questions are referred to be relevant to the relationship between voice-face and voice language. They demonstrated this by preparing their bilingual audio-visual dataset containing human speech of 154 identities in three languages. 37.5% EER reduction is obtained using VGG-Vox, and SincNet. They have noticed the performance is degraded in unknown languages with the model.

(Science & Zheng, 2016) Present a method of using vowel categories for short utterance speaker recognition (SUSR). Defined vowel categories considering Chinese and English languages. After recognition and extraction of phonemes, the obtained vowels are divided into VC’s, which are then used to languages. After recognition and extraction of phonemes, the obtained vowels are divided into VC’s, which are then used to develop Universal Background VC Models (UBVCM) for each obtained vowels are divided into VC’s, which are then used to develop Universal Background VC Models (UBVCM) for each VC. Conventional GMM-UBM system is used for training and develop Universal Background VC Models (UBVCM) for each VC. Conventional GMM-UBM system is used for training and testing. The proposed categories give minimum EERs of 13.76%, 14.03% and 16.18% for 3, 2 and 1 second respectively.

(Elnaggar & Arelhi, 2020) Present an approach for text-dependent speaker phoneme-based (< 1sec) SI with parametric t-distributed stochastic neighbor embedding (pt-SNE) for dimensionality reduction of features to provide 3D-visualization. Total of 1 second long recording of a speaker is recorded each of which divided by 25 micro seconds divided by 10ms hamming window frame rate. Classification of the silence and voice segments are performed using a K-nearest neighbor classifier. Extracting salient features is performed using Gamma tone frequency Cepstral Coefficient (GFCC). The t-Distributed Stochastic Neighbor embedding (t-SNE) is in use to visualize 3d maps. GMM-based speaker modeling is utilized with the k-means++ algorithm. Unsupervised parametric t-SNE with a deep neural network was proposed. Finally, they were able to identify with 75% accuracy with voice data of less than 1-sec duration.

2.2.1. Summary of Related Work

To summarize the research made by the list of researchers and the research conducted by them have had a great impact on the state of art bilingual speaker recognition. The researchers in this

state art agreed on the impact of language mismatch on speaker recognition tasks and some of them tried to prove the existence and extent of the problem while the others researched to come up with a solution. On the contrary, the solutions stated by those practitioners have different undiscovered perspectives to research again. Of the papers mentioned above only (Tadele et al., 2022) mentioned one of the languages in Ethiopia though it is known that more than 80 different languages are available and more than half of the people in Ethiopia are bilingual. The conducted researches have mentioned their model would not be as useful to the languages that were not mentioned in the training but the Afaan Oromo and Amharic are mentioned as a pair nowhere. On the other hand, even is the issue of bilingual SR itself is a huge task handling the scarcity of data is the other factor most practitioners proved to be critical to work on(Science & Zheng, 2016).(Elnaggar & Arelhi, 2020). Therefore the proposed study is targeting bilingual speaker recognition using the three most spoken languages in Ethiopia Amharic and Afaan Oromo, In addition, the proposed method emphasizes on short utterance-based data to calculate the error rate and by combining different algorithms in the state of art.

Table 2-2 Summary of Related Works

Author	***Title***	***Language***	***EER/Accuracy***	***Feature extraction and Modeling Method***	***Gap***
Atklti Niguse	*Text dependent speaker recognition using artificial neural networks for Tigrigna language utterance*	*Tigrinya*	*96.6% and 93.7%*	*MFCC based on MLP NN*	• *Done on one language* • *Only 10 speakers* • *Text dependent*
A.Mengstu and Alemayehu	*Text-independent speaker identification system for Amharic language,*	*Amharic*	*74.2% & 84.3% accuracy had been achieved for 50 speakers*	*VQ & GMMs*	• *Done for one language* • *Traditional machine learning*
Luengo	*Text independent speaker identification in bilingual environments*	*Basque-Spanish bilingual*	*71 and 73 (%Accuracy)*	*Prosodic values with MFCC features.* *GMM modeling*	• *Traditional machine learning model* • *Afaan Oromo and Amharic are not in the context*

Abreham Debasu	*Text-Independent Amharic Language Speaker Recognition in Noisy Env't Using Hybrid Approaches of LPCC, MFCC and GFCC*	*Amharic*	*77.2%, 70.9%, and 69% accuracy had been achieved for 30, 60, 90 speakers respectively*	*LPCC, MFCC, and GFCC with VQ and GMM*	• *Done for one language* • *Traditional machine language*
Gizachew Belayneh	*Artificial neural network based Amharic language speaker recognition*	*Amharic*	*97% accuracy using 20 hidden layers ANN.*	*MFCC and ANN*	• *10 speakers* • *Done for one language*
Tadele et al.	*Effect of Language Mixture on Speaker Verification: An Investigation with Amharic, English, and Mandarin Chinese*	*Amharic, English and Chinese*	*Combination of Amharic Mandarin yielded, relative performance degradation of 17.1% compared to English-Mandarin*	*MFCC*	• *Only effect analysis of language*

CHAPTER THREE

3. METHODOLGY

3.1. An Overview of the Chapter

The methods, approaches, tools, and procedures used to accomplish the study's goal are all explained in this chapter. In this section, various methods such as data collection and preparation, data preprocessing, feature representation techniques, designing a model, different evaluation techniques to evaluate the performance of the designed ASR model, and tools used to develop a proposed model are discussed.

3.1.1. Framework of Methodology

Activities involved in this study framework of methodology are speech collection, preprocessing of speech signals, feature extraction of preprocessed data, feature modeling of extracted features, feature matching during identification, and performance evaluation. All activities have been done using appropriate techniques and tools. Figure 3-1 shows the overall detailed framework of the methodology.

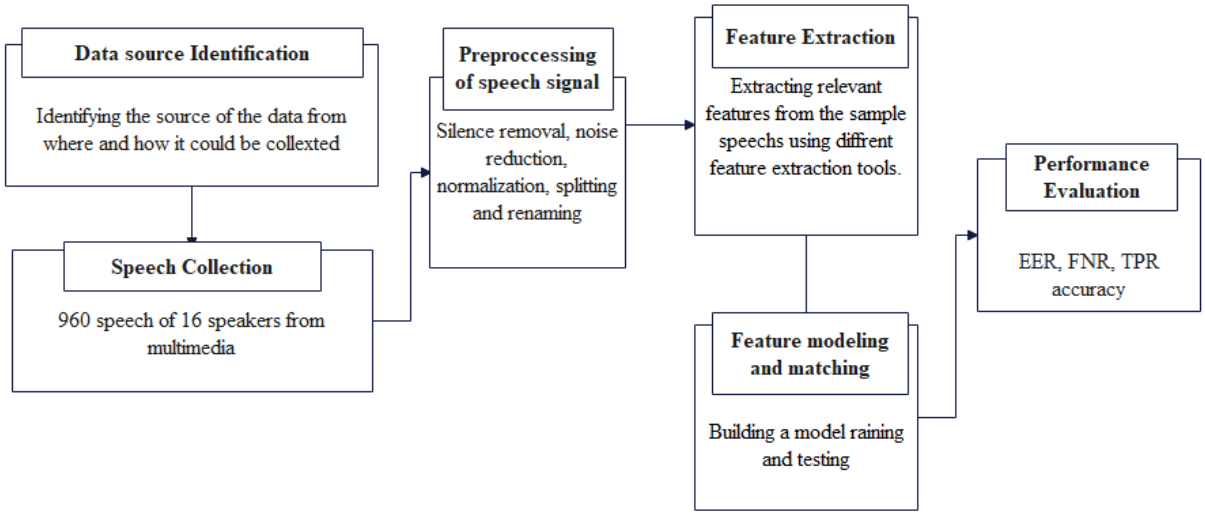


Figure 3-1 Framework of Methodology

3.2. Research Design

Research design is defined as a framework of methods and techniques chosen by a researcher to combine various components of research in a reasonably logical manner so that the research problem is efficiently handled. Selecting an appropriate research design has a great impact on the research's success. The decisions made at this stage of the research is very essential to determine the quality of the conclusions that can draw from research results. Because the expected outcome from this study is performance in percentage, the overall process was numerical. And since the internal key performance indicator selected for this study is the *number of languages*, six experiments have been conducted. Therefore, we chose a quantitative and experimental research design to address the research questions mentioned in chapter one Research Questions 1.4

3.3. Data Source Identification

Bilingual Speaker recognition studies require the speech of bilingual speakers. The performance of the model is affected by the environment the data is taken from, the health of the speaker, the speaker's mood at the time of recording, the language spoken, very short speech length, and noise. The actual problem exists in those aforementioned however we could not get sufficient data from regard especially since people are not willing to give those recordings. Therefore, we used multimedia as a source of data.

Even if there are many social media platforms like Tiktok, Instagram, More, and Snapchat all supporting video. Those social media platforms mostly get used to short-length videos and videos with background music playing. Therefore, YouTube is come into being an adequate data source because it has millions of users all over the world and it is easy to get a long and short recording of different speakers.

To reduce the influence of recording equipment, every speaker's voice was collected from the same situation, that is, from the same YouTube channel and on the same day in both languages. The main reason for this is to use recordings recorded in both languages in the same studio while the person is in the same health and mood. If noise enters it, the same noises will be added to it. Considering this, the qualitative data obtained is from 16 people. To make this clear, if a person named Mr. A were to speak in Amharic in a studio named B, and Mr. A himself was speaking in Oromo in Studio C, the difference in the quality of the recording devices and the health and mood of the speaker at the time of interview affects the main target of this study. To reduce this,

they have tried our best to find and use the recording of the specific person given in a similar describe condition.

We used purposive sampling. It is a type of non-probability sampling in which researchers pick individuals from the public to take part in their surveys based on their judgment. There the criteria's set to be

- A speaker must speak both Afaan Oromo and Amharic
- A video of the selected speaker must be available on YouTube
- The available recording must be in a relatively less noisy condition.
- The recordings must be taken in the same environment and with the same recording

3.4. Speech Collections

The main problem with doing speaker recognition for under-resourced languages like Amharic and Afaan Oromo is the lack of previously available professional datasets. Therefore, we prepared our dataset by taking 16 selected bilingual speakers from YouTube. The speakers are selected based on the criteria set. The multimedia speeches are taken in a different environment while the speakers give interviews, and make a press conference. But the similarity of situations is taken into consideration. All were uploaded to YouTube at different times. The multimedia files are downloaded by *4k downloader* as *.mkv* converted to *.wav* using *.ffmpeg*. Since the speeches in the specific files might not be only from the target speaker (i.e. the interviewer and other unnecessary voices which are not considered silence), therefore those irrelevant recordings are then cut out by *audacity audio trimer* software. A total of 30-second speech is prepared for both languages. All of the speakers are considered to be native Amharic and Afaan Oromo speakers and their accent is not taken into consideration.

Then after, the audio files were changed to ***.Wav*** file format. A WAV file is a lossless audio format that does not compress the original analog audio recording. The high sample rate and bit depth of WAV files, which are a lossless format, enable them to contain all of the frequencies heard by humans. The most basic of the popular file types for audio sample storage are WAV files. WAVs hold samples "in the raw," requiring no pre-processing other than formatting the data, unlike MPEG and other compressed formats. There are three "chunks" of data in the WAV file itself: Specifically, the RIFF chunk that indicates the file is a WAV file, For example, the FORMAT chunk indicates the sampling rate and the DATA chunk which contains the actual data (samples).(*3GP File_*

What A , n.d.)(Wave, n.d.)(*WAV Vs*, n.d.). The conversion is undertaken by ***ffmpeg. Ffmpeg*** is a very fast video and audio converter that can also grab from a live audio/video source.(*Ffmpeg Documentation*, n.d.)

3.5. Preprocessing

The second and most important phase in the speaker recognition process is preprocessing spoken utterances. The speech that has been gathered will be prepared in this stage so that it can be used in the following steps. To get the minimal EER, we utilized several criteria in line with their importance. Thus, this study includes speech file format conversion, formatting sampling rate, signal splitting, noise reduction, normalization, and silence removal.

3.5.1. Silence Removal

Silence removal is the omission of an unvoiced signal from a given speech. Since in unvoiced voice signal there is no relevant feature value to be used for feature matching. Since silence is a characteristic shared among all speakers, its identification is unnecessary. Silence removal relies on identifying periods of the audio therefore, silence ought to be implemented from voice signals that do not contain audio events when compared to periods that are associated with cough events. (Cohen-mcfarlane et al., 2019). Because understanding where speech and silence or unvoiced should be located helps in both reducing the length of the dataset to be processed as well as it helps to improve speaker recognition system accuracy. Voice is the method of communication that humans use most naturally. Speech is produced by the mouth, nose, larynx, pharynx, and other structures in the respiratory system. Speech can generally be divided into vocal, unvoiced, and silent. A frame would be considered silent if all of its samples were zero, which would suggest that the energy level was also zero and that there were no zero-crossings at all. However, there can be a tiny fluctuation if the environment or the microphone make noise when the recording is being done. The energy would be lower than both the voiced and the unvoiced frames, and the zero-crossings would likewise be lower than the other two cases due to the insufficient variety.(Cohen-mcfarlane et al., 2019) Software called an ***audio silence trimmer*** is used for this specific research because of three reasons. The first one is because it is freely available, secondly it is easy and user friendly to use and finally it works successfully on removing the unvoiced speech signal. To add more the absence of silence aids to minimize the

length of the data which in turn provide as a data with only relevant features that best describe the target class.

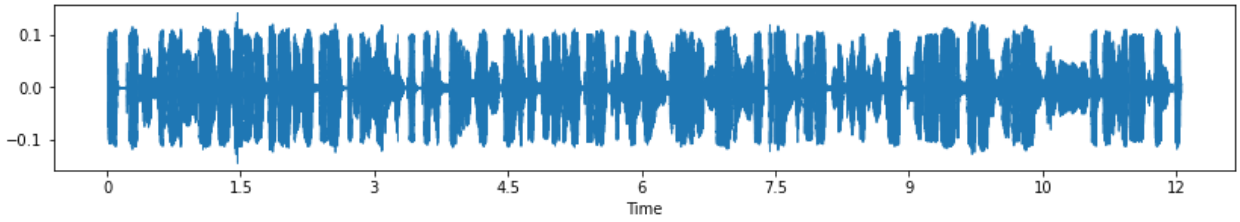


Figure 3-2 AA.wav after Silence Removal

3.5.2. Noise Reduction and Normalization

There is background noise with every utterance since the speech signals are collected from many environments and they are public talks. To precisely extract the special characteristics of the speakers, the surrounding noise should be eliminated. As a result, when silence in the voice signals has been eliminated, noise reduction and normalization come next. In order to bring the amplitude of an audio recording to a desired level, normalization involves applying a consistent amount of gain (the norm). Because the spoken signal's volume is reduced during noise reduction. Normalization restores the volume to its original level (as it was before noise reduction). The speech signal is normalized and the background noise is reduced using the ***Audacity software***. The justification for choosing Audacity because it is free source and ranked third among the top 10 audio editing programs, audacity was chosen. Splitting of Audio Speech Signals.

3.6. Feature Extraction

In many procedures that include computer vision, object detection and location, image processing, and other related fields, the use of feature extraction approaches has become obvious. In order to decode relevant features from the speech signal this study get use to MFCC, Spectral Centroid, Zero Crossing Rate, and Spectral Roll Of, Root Mean Square and Chroma gram as explained below.

3.6.1. Mel Frequency Cepstral Coefficients (MFCC)

One of the most reliable approaches for feature extraction on Speaker Recognition, was employed in this study. The method for obtaining spectral characteristics for speech utilizing the perception-based Mel spaced filter bank processing of the Fourier Transform signal is known

as Mel Frequency Cepstral Coefficients, and it is the most well-liked and often used audio feature extraction methodology. Additionally, the Python programming language's Librosa, which contains a function to read audio files and support the extraction process using MFCC, is used in the feature extraction process.

Fast Fourier Transform (FFT) of a windowed short time signal yields Mel-Frequency Cepstral Coefficients (MFCC), a representation of the real cepstral of that signal. The employment of a nonlinear frequency scale, which simulates the behavior of the auditory system, distinguishes it from the actual cepstral. These coefficients are also resistant to changes caused by the speakers and the recording environment. MFCC is an audio feature extraction method that deemphasizes all other information while extracting speech characteristics that are similar to those utilized by humans to understand speech. A time frame is first created from an arbitrary number of samples using the spoken signal. Most systems use frame overlapping to create a smooth transition from one frame to the next. Then, using a Hamming window, each time frame is windowed to remove discontinuities at the boundaries.

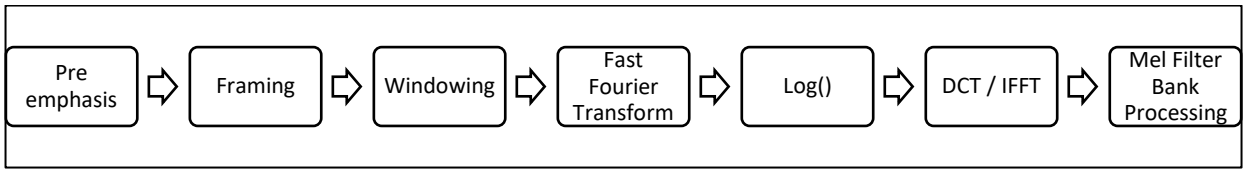


Figure 3-3 MFCC Feature Extraction Steps

Mel Frequency Cepstral Coefficients (MFCC), one of the most reliable approaches for feature extraction on Speaker Recognition, was employed in this study. The method for obtaining spectral characteristics for speech utilizing the perception-based Mel spaced filter bank processing of the Fourier Transform signal is known as Mel Frequency Cepstral Coefficients, and it is the most well-liked and often used audio feature extraction methodology. Additionally, the Python programming language's Librosa, which contains a function to read audio files and support the extraction process using MFCC, is used in the feature extraction process.

Mel-scale frequency cepstral coefficients are the most often utilized acoustic characteristics. The following is an explanation of the MFCC step-by-step computation:

1. **Pre-Emphasis-**

In this step isolated word sample is passed through a filter which emphasizes higher frequencies. It will increase the energy of signal at higher frequency. Filtering that emphasizes higher frequencies is referred to as pre-emphasis. Its goal is to balance the spectrum of spoken sounds with a sharp roll-off in the high-frequency range. The glottal source has a slope of around 12 dB/octave for spoken sounds(Mfcc, 2017). When acoustic energy is emitted by the lips, the spectrum is boosted by around +6 dB/octave. As a result, a voice signal captured with a microphone from a distance has a 6 dB/octave downward slope when compared to the genuine spectrum of the vocal tract. As a result, pre-emphasis eliminates some glottal effects from the vocal tract parameters. The following transfer function describes the most often used pre-emphasis filter.

2. **Framing**

The voice signal has a slow temporal variation or is quasi-stationary. Speech must be evaluated for a sufficiently short period of time to have stable acoustic properties. As a result, speech analysis must always be performed on short segments when the speech signal is believed to be stationary. Short-term spectral measurements are often performed across 20ms windows, with 10ms intervals(*Rabiner.Pdf*, n.d.)(*Neural Networks and Deep Learning*, n.d.). The signal must be divided into brief time frames. This step is necessary because a signal's frequencies fluctuate over time. If we applied the Fourier transform to the full signal, we would lose track of the signal's changing frequency contours. Signal frames should last 20–40 ms. Standard is 25 ms. Accordingly, a 22 kHz signal's frame length is 552 samples, or 0.025 * 22080, with a sample hop length of 160 samples.

3. **Windowing**.

A window is used on each frame to taper the signal towards the frame borders.(*Rabiner.Pdf*, n.d.) In most cases, Hanning or Hamming windows are utilized. This is done to improve the harmonics, smooth the edges, and lessen the edge impact while performing DFT on the signal(Mfcc, 2017). Windowing is essentially applied to notably counteract the assumption made by the Fast Fourier Transform that the data is infinite and to reduce spectral leakage.

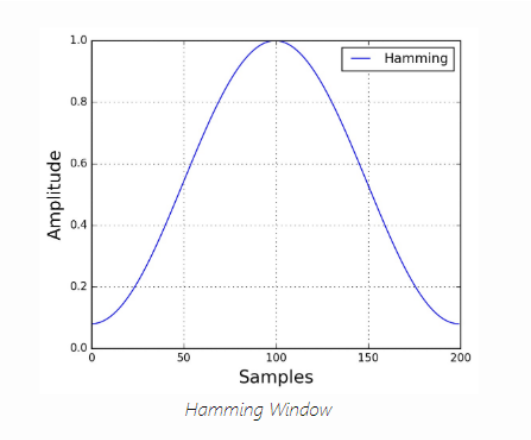


Figure 3-4 Hamming Window

4. Fast Fourier Transform

Now, we may compute the frequency spectrum using a NN-point FFT on each frame, often known as a short-time fourier transform (STFT), where NN is generally 256 or 512, and NFFT = 512, and then compute the power spectrum (periodogram[13]).

5. Mel Filter Bank Processing

Our filterbank consists of 40 vectors, each of which is 257 length (assuming the FFT settings fom step 2). Each vector consists mainly of zeros, but for a portion of the spectrum, it contains non-zero values. We multiply each filterbank by the power spectrum to compute the filterbank energies, then we add the coefficients. 40 digits are left after this process, each of which indicates how much energy was in each filterbank.

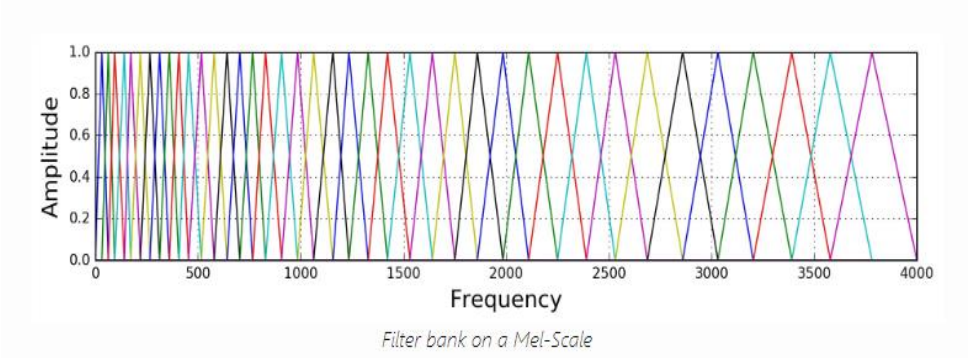


Figure 3-5 Mel Filter Bank

In order to obtain the filter banks depicted in Figure 3-5

[13] Periodogram: An estimate of the spectral density of a signal.

We must first select a lower and upper frequency. For the lower frequency, a good value is 300 Hz, and for the upper frequency, 8000 Hz. Then transform the higher and lower frequencies into Mels. In our example, 8000Hz is 2834.99 Mels, and 300Hz is 401.25 Mels. Since every formula is provided, converting from frequency to MEL is really simple. This provides us with 40 coefficients, which can be any number depending on the requirements, between the chosen ranges. After that, these coefficients are changed back to hertz. These points are then used to plot the filter bank!

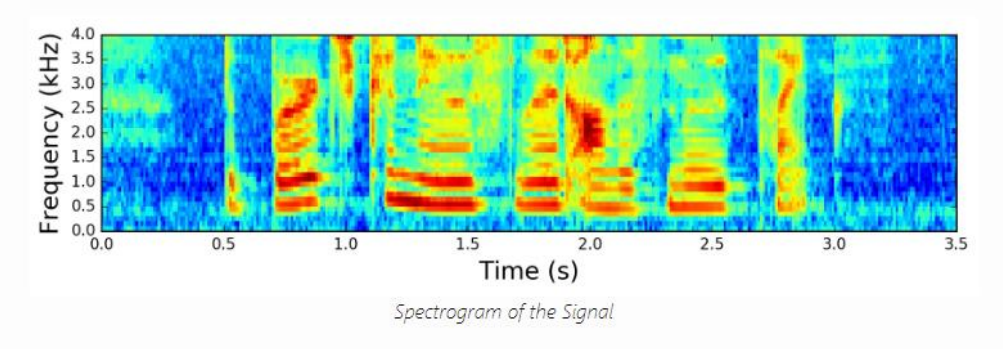


Figure 3-6 Spectrogram of the Signal

6. Discrete Cosine Transform

Discrete cosine transform is the process to convert the log Mel spectrum into time domain. The result is called Mel Frequency Cepstrum Coefficients.

All of the speech signals has been converted in to their respective feature vector by following the above five MFCC feature extraction technique steps. The number of feature coefficient to be extracted is 21. Finally, the feature vectors of the speech signals has been stored in to database in CSV format.

Librosa is an effective and easy to use python library. It is used mostly for music processing, for the reason that it provides various feature extraction tools. In this study the MFCC other described features are extracted from the speeches using this library.

3.6.2. Spectral Centroid

The spectral centroid is the center of ‘gravity’ of the spectrum. The spectral centroid of a signal is the curve whose value at any given time is the centroid of the corresponding constant-time cross section of the signal’s spectrogram. A spectral centroid provides a noise-robust estimate

of how the dominant frequency of a signal changes over time. As such, spectral centroids are an increasingly popular tool in several signal processing applications, such as speech processing.

In digital signal processing, the spectral centroid is a measure used to characterize a spectrum. It indicates the location of the spectrum's center of mass. It has a strong perceptual connection with the perception of brightness of a sound. The spectral centroid is a statistic that can be useful in characterizing the spectrum of the audio file signal because the audio files are digital signals.

Each frame of a magnitude spectrogram is normalized and treated as a distribution over frequency bins, from which the mean (centroid) is extracted per frame.

More precisely, the centroid at frame t is defined as

$$centroid[t] = sum_k S[k,t] * \frac{freq[k]}{(sum_j S[j,t])} \qquad \textit{Equation 2}$$

Using librosa library the spectral centroid is calculated as

$$cent = librosa.feature.spectral\_centroid(y=y, sr=sr) \qquad \textit{Algorithm 1}$$

3.6.3. Spectral Bandwidth

The difference between the higher and lower frequencies in a continuous range of frequencies is known as the bandwidth.

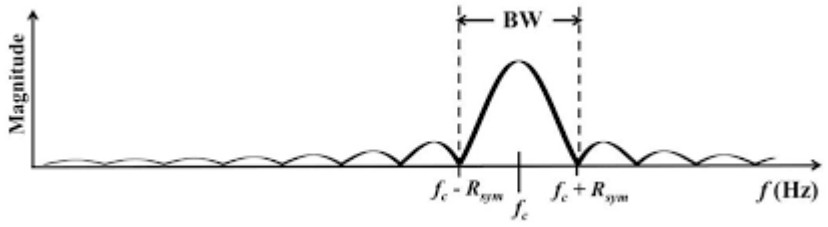


Figure 3-7 Bandwidth [14]

The difference between the higher and lower frequencies in a continuous range of frequencies is known as the bandwidth. Signals, as we are aware, oscillate about a point; therefore, if the point is the centroid of the signal, the bandwidth of the signal at that time can be calculated as the sum of the maximum deviations of the signal on both sides of the point(*A Tutorial on Spectral Feature Extraction for Audio Analytics* -, n.d.)

[14] https://www.usna.edu/ECE/ec312/Lessons/wireless/EC312_Lesson_23_Digital_Modulation_Course_Notes.pdf

The spectral bandwidth at frame t is computed by:

$$sum\_k\ S[k,t]\ *\ (freq[k,t]\ -\ centroid[t])\ **\ p)\ **\ (1/p)$$ *Equation 3*

Using librosa library the spectral centroid is calculated by

$$spec\_bw\ =\ librosa.feature.spectral\_bandwidth(y = y, sr = sr)$$ *Algorithm 2*

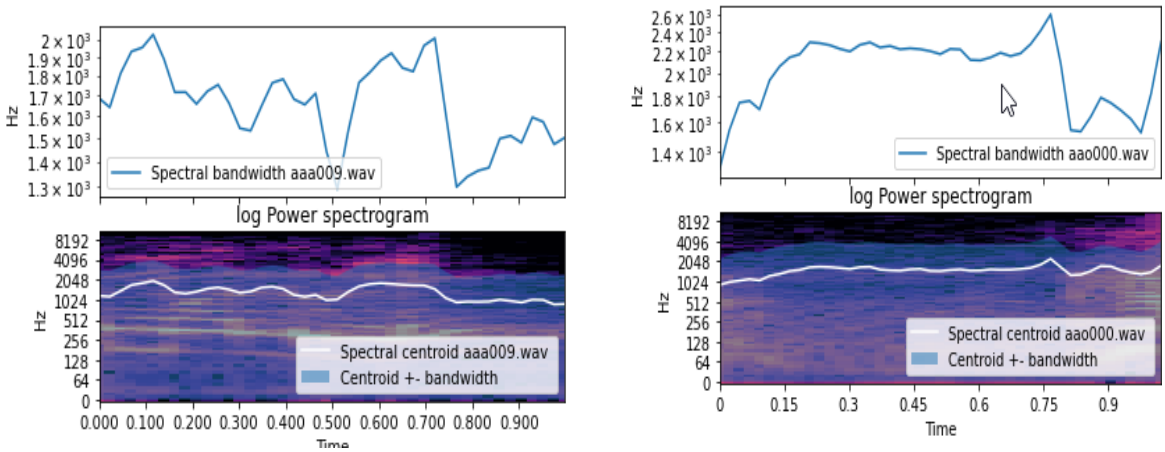


Figure 3-8 Plot Diagram Showing Spectral Bandwidth

3.6.4. Spectral Roll-off

The frequency under which a portion (cutoff) of the spectrum's total energy is stored is known as the roll-off frequency. Harmonic (below roll-off) and loud sounds can be distinguished using the roll-off frequency (above roll-off).

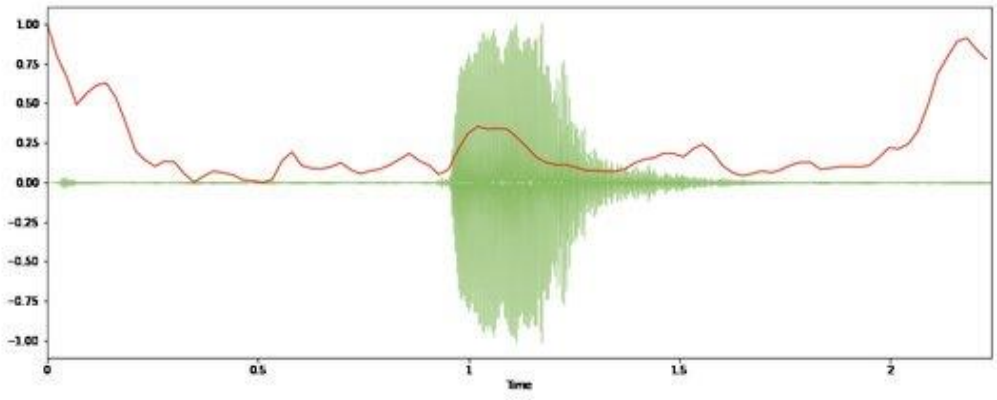


Figure 3-9 Spectral Roll-off(Hussain et al., 2021)

It can be described as the result of a certain kind of filter that is intended to roll off frequencies outside of a particular range. We refer to the process as "roll-off" because it is progressive. Both

hi-pass and low-pass filters have the ability to roll off the frequency of a signal that crosses their thresholds. Formally, we could state the percentage of power spectrum bins at which lower frequencies account for 85% of the power is known as the spectral roll-off point.

The maximum and minimum can be calculated with this by setting the roll percent to a number between 1 and 0.

3.6.5. Zero Crossing Rate

The rate at which a signal changes from positive to zero to negative or from negative to zero to positive is known as the zero-crossing rate (ZCR).

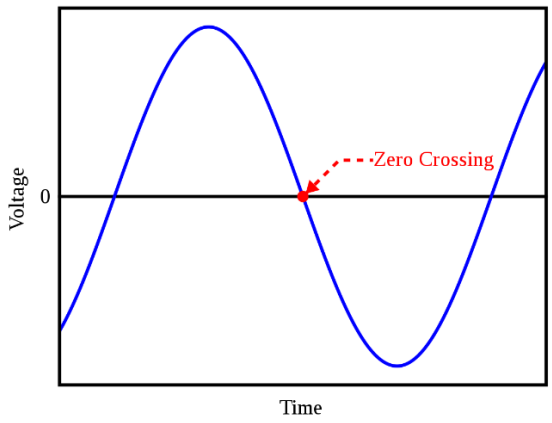


Figure 3-10 Zero Crossing Rate

3.6.6. Chroma Gram

Chroma gram is defined as the whole spectral audio information mapped into one octave. Each octave is divided into 12 bins representing each one semitone.

3.6.7. Root Mean Square

Statistically, the root mean square (RMS) is the square root of the mean square, which is the arithmetic mean of the squares of a group of values.

3.7. Model Selection Techniques

Advanced Speaker recognition models make use of several deep learning algorithms. Since there is no any perfect criteria set to each model for a particular problem, making experiment is the optimum way to select the one fits the problem. This study has set some criteria to pick the models.

The nature of the dataset: - some models work on text and some others perform well on image while some work best on speech. While some of them have best relative performance based on the dataset given. Since this research used a feature vector dataset as an input to the neural network, and the model has to learn the continuity and similarity of the feature vectors in the different utterance of every speaker.

The nature of the problem: - This study by its nature is a multi-class classification problem. And classification problems are best handled in deep learning than machine learning. The features are classified in to 16 speaker classes based on the given dataset.

The third criterion is the most widely used and well-performed model in previous studies for processing NLP tasks. Various deep learning algorithms got success in speaker recognition. The resent works in speaker recognition shows that deep learning models out performed machine learning models greatly. (Et. al., 2021)(Hourri et al., 2021)

The last criterion is to make experimentation between the proposed deep learning models using a prepared dataset to select the outperformed model for this study. To our knowledge, one model isn't said to be better without experimenting with them. Once the best model is known, we can work with that model for further performance improvements.

3.8. Feature Modeling and Matching Technique

The features extracted from the speech signal are vectors stored to a database in **.csv** file format. Each of the features are used for modeling. For modeling the feature vectors neural networks were used. Even though a neural network has many architectures MLP and CNN is the architecture selected for this system.

Multilayer feed forward neural network is a network of perceptron's in which information and computations travel from the input data to the outputs in a single path. The number of perceptron layers equals the number of layers in a neural network.

Convolutional neural network (CNN) is an advanced deep neural network. It is mostly used in computer vision for image classification. Due to its convolving architecture CNN is have shown best performance in classification tasks.

To design MLP and CNN models we used Keras, it is an easy to use but powerful deep learning library for python. Keras is an open-source, user-friendly deep learning library created by

Francois Chollet, a deep learning researcher at Google and it is built on top of Tenser Flow. It have a built in method which allows addition of layers as well as witch enables as to save and reload model values**.** Keras have 3 model type Sequential, functional API and Model sub classing[15]. In this study we used sequential model since the data set is continuous feature vectors.

Keras have Layers API to build different MLP and CNN layers

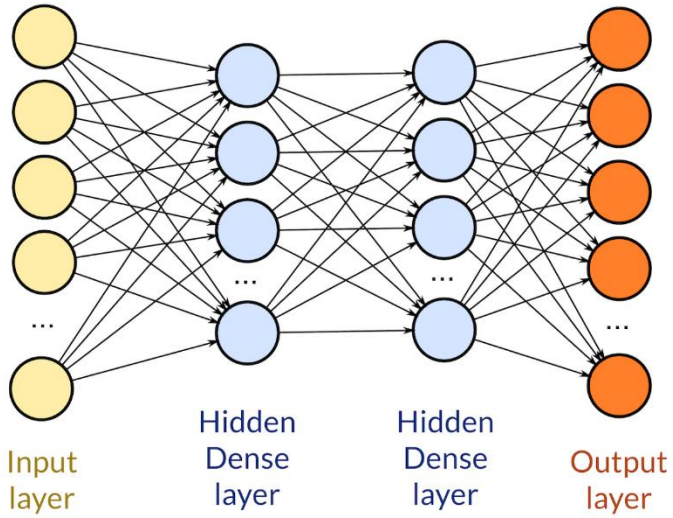


Figure 3-11 Some of the Layers in Keras Layers API

3.9. Evaluation Experiments

The experiments have been prepared from a language perspective.

1. **Experiment 1:** Training Afaan Oromo and Amharic test Amharic [AM-OR]- [AM]
2. **Experiment 2:** Training Afaan Oromo and Amharic test Afaan Oromo[AM-OR]- [OR]
3. **Experiment 3:** Training Afaan Oromo and Amharic test Afaan Oromo[AM-OR]- [AM-OR]
4. **Experiment 4:** Train and Test with Afaan Oromo [OR]-[OR]
5. **Experiment 5:** Train and test with Amharic [AM]-[AM]
6. **Experiment 6:** Train with Afaan Oromo and Test with Amharic. [OR]-[AM]
7. **Experiment 7:** Train Amharic Test with Afaan Oromo [AM]- [OR]

[15] https://keras.io/api/models/

3.10. Implementation Tools and Techniques

The programing language used to develop the experiment is python 3.10. Python is a fully fledged programing language consisting of various built-in libraries like Numpy, Torch, Keras, Matplotlib, Scipy and tensor flow which highly assists research and experiment on various fields. Python is simple to code and understand an already written code, it has multiple frameworks and built in API's to assist deep learning, it can be integrated flexibly with a lot of other frameworks.

Python is preferable tool to work with files. The python library Numpy is used to perform mathematical calculation, array and vector manipulations. Librosa is a library containing the feature extraction algorithms, and Keras is provided as a simple user-friendly neural network development framework.

The speech files loaded for feature extraction and saved as **.csv** file as a vector of 36 features. The feature extraction is done with the help of libraries of librosa. The model is building training and testing is done with Keras.

The main reason to prefer python is because of its advancement in AI especially for speaker recognition it have two basic features i.e. Keras and Librosa. In addition to this, in general, it has a vast library, interactive environment, and built-in graphics for visualizing data and tools for designing with custom GUI.

3.10.1. System Design Tools

E-draw Max: - E-draw max is a software diagramming tool that can help us to design any diagram in this study. This software is a very popular software to design UML diagrams, organizational charts, system design, workflow diagrams, etc. It appears to support over 280 different types of diagrams[16].

3.10.2. Implementation Environments

Anaconda 4.10.3:- It is a Python and other programming language distribution environment for scientific computing such as machine learning, data science, data processing, etc. The primary goal of this environment is to simplify the package and implementation. The anaconda

[16] https://www.g2.com/products/edraw-max/reviews

distribution originated with 250 packages and 7,500 extra open-source packages[17]. It has different applications like Jupyter notebook, Spyder, JupyterLab, etc.

Jupyter Notebook 6.1.4:- Jupyter Notebook is a web-based interactive computational environment that allows you to create notebooks, code, and data. Jupyter notebook allows data scientists and machine learning developers to organize workflows.

Colab: - Colab is a Google research product that allows anyone to write and run Python code, particularly machine learning, in the browser. It is a free Jupyter notebook environment that runs in the cloud to write and execute code in python language. One benefit of using Colab is it enables the different teams to edit the code simultaneously. Colab is a Jupyter notebook service that does not require any configuration and is free of resources such as the Tensor Processing Unit (TPU) and Graphical Processing Units (GPUs)[18]. We used it for our speaker recognition development because it supports deep learning libraries and provides free access to resources.

3.10.3. Programming Language Tools

Python 3.10.6

Python is a high-level, object-oriented, well-structured, and readable programming language. The python programming language was created by Guido van Rossum from 1985- to 1990[19]. It is a general-purpose language designed to be used in various ranges of applications such as automation, data science, etc. We use the 3.8.5 version of the python programming language because of its compatibility with many packages. The language uses English keywords and has rarer syntactical constructions for this reason it is easy to use and understand by programmers.

3.13.4. Important libraries

Keras 2.6.0:- Keras is a Python-based open-source high-level neural network API and runs on Tensor Flow, CNTK, Theano, etc. Keras is designed to offer rapid experimentation with DL to allow for easy and fast prototyping and running seam on CPU and GPU (Choudhury, 2019). Unlike tensor flow, Keras only provides a high-level application programming interface, but tensor flow provides both high-level and low-level API.

[17] https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)
[18] https://research.google.com/colaboratory/faq.html
[19] https://www.tutorialspoint.com/python/index.htm

Matplotlib: - Matplotlib is a cross-platform data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it provides a viable open-source alternative to MATLAB. Developers can also embed charts in GUI applications using Matplotlib's API (application programming interface)[20].

Pandas: - Pandas is the most popular open source Python package for data science/data analysis and machine learning tasks. It is based on another package called **Numpy** that provides support for multidimensional arrays. One of the most popular data wrangling packages, Pandas works well with many other data science modules in the Python ecosystem, ranging from those typically shipped with the operating system to more popular ones like Active State's Active Python. It's included in all Python distributions, up to and including commercial providers.

Librosa: - Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems

Scikit-learn (Sklearn):- It is a powerful open-source machine learning package for analyzing data in the Python programming language. The library has commonly been written in Python and works with NumPy, SciPy, and Matplotlib. NumPy is a Python library for manipulating large, multi-dimensional arrays and matrices.

3.10.4. Data Preparation Tools

Audacity

Audacity is an easy-to-use, multi-track audio editor and recorder for Windows, macOS, GNU/Linux, and other operating systems. Audacity is free, open-source software

4k Video Downloader

4K Video Downloader 64 bit for PC allows downloading videos, playlists, channels, and subtitles from YouTube.

Audio Silence Trimmer pro

Audio Silence Trimmer Pro helps you remove silence from multiple audio files together. You can just give it a list of files to trim instead of doing them one by one. It comes with a

[20] https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/

minimalistic, user-friendly interface that integrates simple controls, making it very accessible and efficient at the same time. Depending on preference, it can either trim the silence at the beginning, end, beginning, and end, or perform a full audio trim, which removes all silence in your recording. You can also add silence at the beginning and end.

Ffmpeg

Ffmpeg is a very fast video and audio converter that can also capture from a live audio/video source. It can also convert between arbitrary sample rates and resize video on the fly using a high-quality multi-phase filter. ffmpeg reads from an arbitrary number of input "files" (which can be regular files, channels, network streams, capture devices, etc.) specified by the -i option and writes to an arbitrary number of output "files" specified by a simple URL- output address[21].

3.10.5. Hardware Tools

To install and work with the above tools, the personal computer that is well-found with an internal hard disk of 1TB, 8GB of physical memory, processor intel® core™ i7-7500U CPU @ 2.70GHz 2.90Hz, Windows 10 pros, and system: 64-bit operating system, the x64-based processor is used.

3.11. Evaluation Methods

Performance and error rate must be assessed for the related result. Since the majority of speaker recognition studies analyze results based on their False Positive and True Negative, EER (equal error rate) is mostly used in this study as a measurement approach. Along with comparing the results to those of other studies that used accurate, precise, and recall measurements, such measurement procedures are also assessed as supplementary metrics. Speech from speakers that has been prepared for testing in the trained neural network is tested in real time. Here, the trained neural network has been tested in relation to the prepared speech of the speakers. Each speaker's proportion of correctly identified objects has been calculated.

3.11.1. Confusion matrix

Confusion matrix show 4 parameters

1. The True negative [TN] shows the incorrect identification of the correct speaker

[21] https://www.ffmpeg.org/ffmpeg.html

2. The True Positive [TP] tells the correct identification of the incorrect speaker
3. The False Negative [FP] is when the wrong speaker is identified incorrectly
4. False positive [FP] when the wrong speaker is identified correctly.

3.11.2. Accuracy

This is a metric that commonly explains how the model performs through all classes. It is a vital metric in every industry to measure the performance of machine learning models. Describes what percentage of our test data is correctly classified. When all classes are equally significant, this metric is adequate. The model accuracy can be calculated as the number of correctly classified instances per the total number of classifications. It is computed using the mathematical formula in Accuracy is calculated with

$$Classification\ Acuracy = \frac{(TP+TN)}{(TN+TP+FP+FN)} \quad \textit{Equation 4}$$

$$Precission = \frac{TP}{(TP+FP)} \quad \textit{Equation 5}$$

$$Recall = \frac{TP}{(TP+FN)} \quad \textit{Equation 6}$$

3.11.3. Equal Error Rate

The threshold parameters for a biometric security system's false acceptance rate and false rejection rate are predetermined using the equal error rate (EER) technique. The common value is referred to as the equal error rate when the rates are equal. The result shows that the proportion of false acceptances and false rejections is equal. The accuracy of the biometric system is higher the lower the equal error rate value.

Where TP is a true positive that is correctly predicted as actual, FP is a false positive that is a negative value predicted as positive, TN is a true negative that is correctly predicted as negative, and FN is a false negative that is a positive value predicted as negative.

3.11.4. Loss

Loss is an evaluation metric that is used to measure the training loss and validation loss of the model. The training loss measures how the model fits on the training dataset, whereas validation loss evaluates the model on the validation dataset[22]. Validation and training loss

[22] *https://www.baeldung.com/cs/training-validation-loss-deep-learning*

is computed from the summation of errors for each instance in the validation set and training set, respectively. This metric was used in this work to notify us whether the model requires additional tuning or not. Thus, we employ this metric to see how the model fits on the validation and training datasets, respectively by visualizing them on the graph.

3.12. Summary

Table 3-1 Summary of Techniques and Tools And Their Purpose

No	Techniques and tools	Purpose
1	4k Video downloader	To download the speech of target speakers
2	Ffmpeg	To convert the downloaded **.mkv** file to .**wav** To crop the speech to 1 **sec** long and rename it consecutively To format file extension
3	Audacity	To trim the selected 20-30 second part of the speech
4	Audio silence Trimmer pro	To remove the front middle and end silence from the selected speech
6	Librosa	To call the feature extraction libraries MFCC
7	MFCC	Feature extraction tool
8	Matplotlib	To display the spectrogram of sample speech's
9	Python	Programing language
10	Keras	To call the Model building training and testing
11	MLP	Model building and matching algorithm
12	Google Colab	IDE

CHAPTER FOUR

4. DESIGN AND IMPLEMENTATION

4.1. Chapter Overview

The chapter describes the proposed model's design and implementation. This section will explain various points such as the general architecture of the designed model, proposed data preprocessing, proposed feature extraction techniques, proposed deep learning model, implementation of data preprocessing, implementation of word embedding methods, and proposed model implementation.

4.2. The Architecture of the Proposed Model

The main objective of this study is to develop a bilingual speaker recognition model for Afaan Oromo and Amharic language speakers that provides a better recognition performance. The dataset will be collected and prepared from multimedia sources in such a way that it could accomplish the objective of the study. The recordings are collected manually from YouTube. The model does not make direct use of the collected dataset. It was filtered and prepared in a format suitable for the deep learning model. And then, the dataset is prepared using several software and python libraries. Downloading, cutting the sample from the total downloaded multimedia file, changing the frequency to 48 kHz, dividing the sample into a second long sample, and organizing it to a folder after preparation the preprocessing will be done. Silence removal preprocessing techniques used in the study.

Once the dataset is prepared and preprocessing is complete, the next step is converting speech data into feature vector. The speech data must be converted to feature vector by using python library called librosa the features like spectral centroid, zero crossing rate, and MFCC are calculated and saved to a .csv file

The feature vector data is ready for the models, and then we built the deep learning model with the correct hyper parameter and trained it with a training dataset. Building the model with the right network hyper parameter and training the model is the core task in this work.

After preprocessing which includes noise reduction, and silence removal the collected speech is now loaded in two separate folders the (**“oro”**) folder contains the Afaan Oromo speech samples of each speaker and the (**“ama”)** folder contains the Amharic speech samples of each 16 speakers.

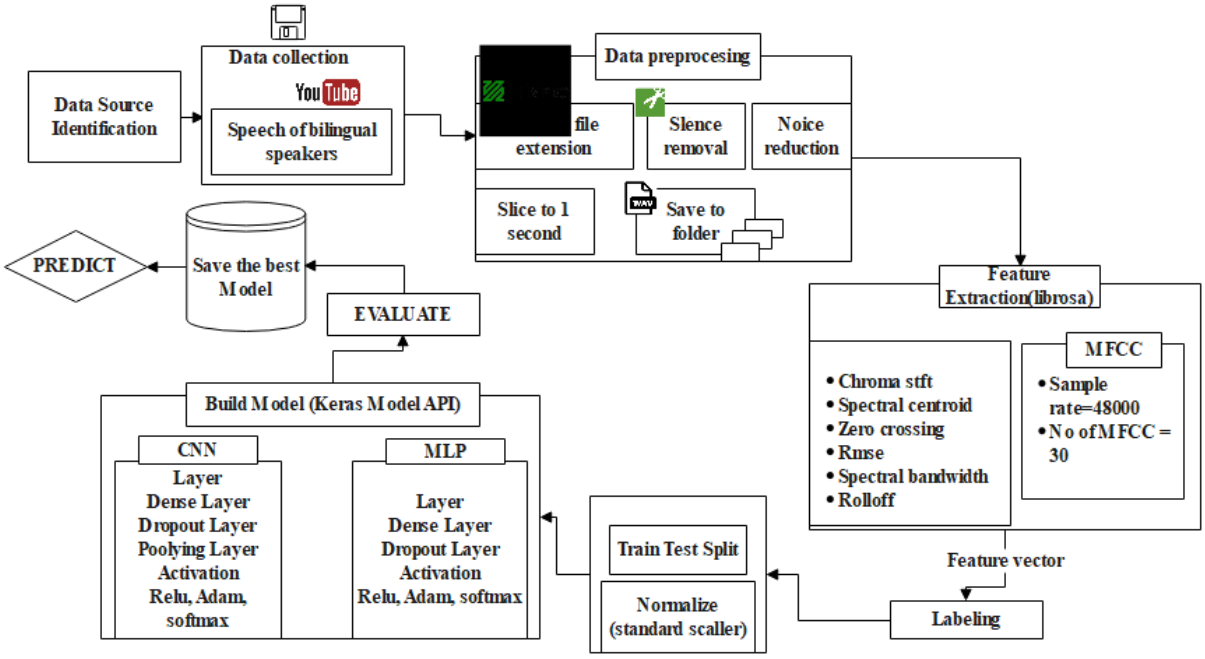


Figure 4-1 Proposed Model Architecture

4.3. Proposed Data Preprocessing

The collected raw speech data is from YouTube. Those speeches are noisy, recorded in a wild randomly with different recording material, and too large to extract features. So when designing a deep learning model, do not always encounter clean, complete, and well-organized data from the real world. Therefore, it is mandatory to work with data and put it in an organized way according to study use cases. Various data preprocessing approaches were utilized in the study, including noise removal, silence removal, resampling, and file extension shifting.

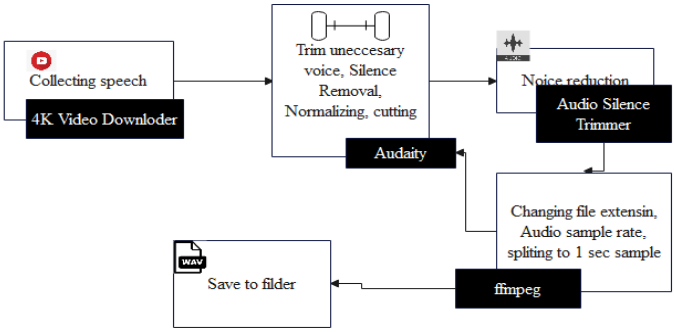


Figure 4-2 Proposed Data Preprocessing Architecture

The data processing is done using different software like audacity, .ffmpeg and audio silence trimer. The 1-sec long speech samples are labeled with a unique tag using the combination of letters from the speaker's name.

4.4. Proposed Feature Extraction

When designing a deep learning model, do not always encounter clean, complete, and well organized data from the real world. Especially in low resource country's like Ethiopia. Therefore, it is mandatory to work with data and put it in an organized way according to study use cases. After loading the data must be converted to feature vector

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. The manageable groups in speech processing are the spectral centroid, the spectral bandwidth, roll of, zero crossing rate and MFCC coefficients. Each of which hold unique identifying value. Many speech feature information extraction tasks are done using built in python libraries we used A Python package for analyzing music and audio called librosa. It offers the components required to build music information retrieval systems(*Librosa — Librosa 0*, n.d.). Librosa have a built in methods to calculate the required features from the signal.

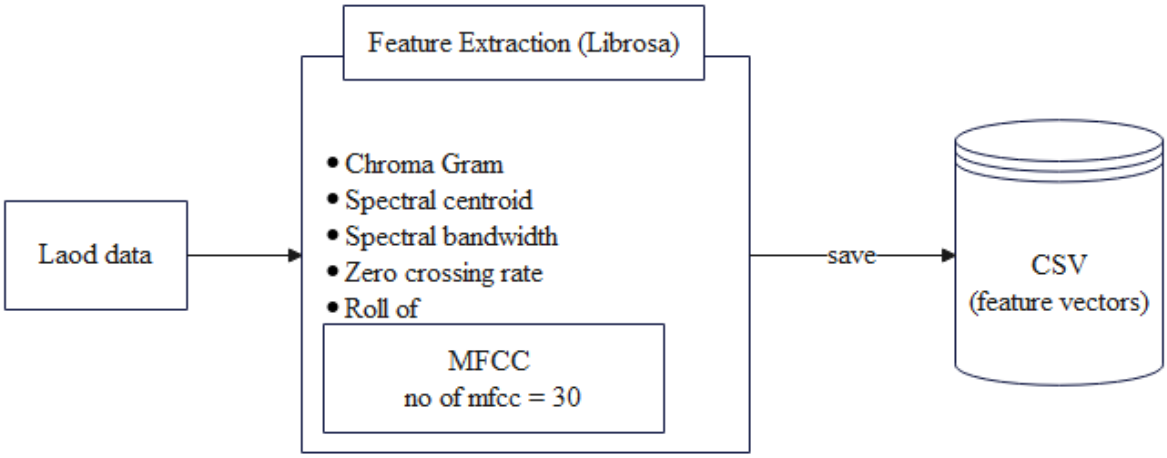


Figure 4-3 Proposed Feature Extraction

4.4.1. MFCC

Mel Frequency Cestrum Coefficient are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip. Cepstral features contain data about the rate changes in the various spectrum bands because they are computed by taking the Fourier transform of the warped logarithmic spectrum. Cepstral characteristics are advantageous because they can distinguish between the effects of the source and filter on a voice signal. In other words, because the low-frequency excitation and formant filtering of the vocal

tract are located in different parts of the cepstral domain, it is possible to distinguish between the influence of the vocal cords (source) and the vocal tract (filter) in a signal.

Since the majority of the spectral energy in sonorant sounds is focused in the low-frequency ranges, a cepstral coefficient with a positive value denotes a sonorant sound. On the other hand, since the majority of the spectral energy in fricative sounds are concentrated at high frequencies, a cepstral coefficient with a negative value denotes a fricative sound. The distribution of spectral energy between low and high frequencies is represented by the first-order coefficient. The average power of the input signal is represented by the zero-order coefficient. The majority of the details on the overall spectral structure of the source-filter transfer function are found in the lower order coefficients. Depending on the sampling rate and estimate technique, 12 to 20 cepstral coefficients are often ideal for speech analysis, even if higher order coefficients imply increasing degrees of spectral details. The models become more complex when a lot of cepstral coefficients are chosen

$$dsmfcc \;=\; librosa.feature.mfcc\left(y = y, sr = sr, n_{mfcc} = 31\right) \qquad \textit{Algorithm 3}$$

Figure 4-4 MFCC Feature Extraction

Bandwidth, roll of, zero crossing rate in addition to 30 MFCC is calculated from each speech samples

filename	chroma_stft	rmse	spectral_centroid	spectral_bandwidt	rolloff	zero_crossing_rate	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8
aaa000.wav	0.360182256	0.10906287	2792.047928	2301.592254	5315.83	0.163237847	-167.926	76.3923	-4.39276	24.6582	-17.97	18.8739	-40.7995	-12.2708
aaa001.wav	0.394392073	0.07087082	1849.344337	1780.822916	3121.58	0.102849787	-234.939	130.023	-12.7522	8.24834	-16.3808	3.33958	-15.5563	3.55318
aaa002.wav	0.318947524	0.13777071	2050.8016	2069.060626	4016.68	0.106367631	-148.594	110.445	-7.25659	18.1057	-26.4684	5.97187	-33.6068	-8.34488
aaa003.wav	0.341570228	0.11988035	1265.865593	1622.089713	2492.96	0.050625888	-212.088	144.564	-15.9932	17.4928	-14.3523	-5.34457	-22.2867	-16.3596
aaa004.wav	0.397361964	0.09992153	3021.138593	2238.060035	5242.6	0.208662553	-172.614	81.0587	7.58688	19.4599	-33.0627	9.77219	-36.6074	0.16011
aaa005.wav	0.347466618	0.1086162	1420.405523	1801.130422	2809.35	0.06069114	-211.756	137.849	-4.15382	13.6147	-30.4878	-3.03683	-25.0872	-12.6117
aaa006.wav	0.240343869	0.16437003	2107.408366	2002.63658	3967	0.106822621	-154.317	93.9544	-21.9662	25.8454	-15.7586	6.40083	-30.8563	-6.70776
aaa007.wav	0.270026267	0.1354699	1464.317547	1770.369295	2898.66	0.049937855	-198.683	123.412	-19.6665	7.44256	-21.7996	-0.46109	-19.6648	-4.72679
aaa008.wav	0.306940764	0.12205204	3211.713894	2334.00466	5571.72	0.224109996	-146.921	68.4422	3.89382	11.0746	-33.016	4.34355	-33.2359	0.11969
aaa009.wav	0.225167066	0.13836323	1339.185157	1666.490428	2685.53	0.041814631	-220.901	121.097	-20.5415	30.9227	-1.17736	-2.90953	-24.9031	-4.08944
aaa010.wav	0.32729724	0.07445841	1895.86082	1934.778121	3646.7	0.079612038	-258.535	102.188	-16.4182	42.7867	4.60377	3.62393	-20.8727	-10.2388
aaa011.wav	0.309152395	0.13717614	1540.123713	1643.698614	2895.97	0.066861239	-178.948	130.826	-21.7282	26.5085	-22.3629	-1.15177	-29.595	-16.47
aaa012.wav	0.284058273	0.12521668	2264.950148	2108.511108	4221.73	0.113591974	-185.16	96.47	-2.56542	21.3635	-14.3125	6.22062	-37.0028	-13.9817
aaa013.wav	0.282537192	0.14090215	1586.876073	1725.278762	3124.03	0.056429776	-172.643	113.228	-33.2937	41.079	-14.0082	-10.3164	-28.9419	-13.0044
aaa014.wav	0.288325846	0.15937485	1397.258494	1627.354198	2649.07	0.04897239	-153.515	133.954	-38.1516	18.2378	-14.9652	-4.84526	-31.7664	-15.5324
aaa015.wav	0.347607017	0.09655952	1544.48686	1765.502372	3015.61	0.073231337	-215.976	132.71	-13.6478	17.6515	-18.38	-10.4722	-26.4292	-14.1228
aaa016.wav	0.306225926	0.14513384	1814.386137	1807.385573	3364.07	0.082108931	-145.888	121.034	-29.7206	16.0734	-21.7675	0.96297	-30.2808	-15.2437
aaa017.wav	0.319327712	0.09278565	2228.558454	2036.934195	4083.48	0.133456143	-199.921	106.744	-8.85125	3.46134	-29.9391	5.80877	-30.2626	-10.6495
aaa018.wav	0.335226387	0.11982322	2396.242954	2098.637818	4548.16	0.109097567	-183.871	88.6221	-15.3099	30.1456	-11.9829	12.6569	-29.3602	-13.3426
aaa019.wav	0.309582561	0.13638909	1797.654118	1816.716526	3368.48	0.081076882	-193.876	113.105	-9.32499	29.9733	-8.53698	3.60403	-27.9354	-13.6436

Figure 4-5 CSV File of Extracted Features

4.5. Proposed Models

4.5.1. Proposed MLP Model

This study proposed two deep learning models. Proposed MLP model hyper parameters used for all experiments are listed below. The Hyper parameters are selected based on experiment and kerasTuner [23]

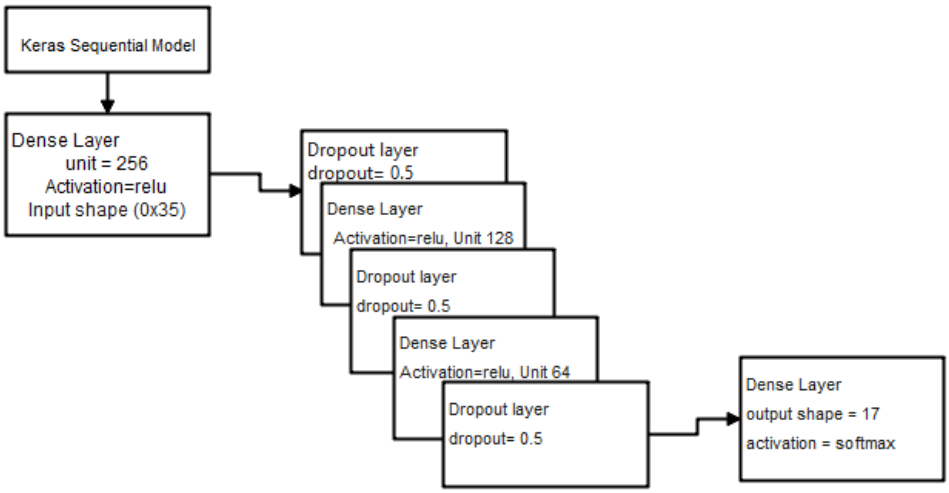


Figure 4-6 Proposed MLP Model

- Model = Sequential Keras. Model
- Layer = 7 layers
- Dropout = 0.5
- Number of Input Neurons = 35
- Number of dense layers = 4 (256, 128, 64 and 18 units[24])
- Training Ratio=70%, Validation Ratio=30%
- Test data 20%
- Epochs=186
- Bach size = 128
- Activation function = relu, softmax
- Optimizer= Adam,
- Loss function ='sparse_categorical_crossentropy', metrics = accuracy

4.5.2. Proposed CNN Model

The proposed CNN Model is a Keras Sequential model with two Conv1D layers each having unit of 512 and 128 respectively. The shape of the input is the size of the Training set and the first Conv1D layer have 4 dimension while the second is set to have 3 dimension reduced in the

[23] https://keras.io/keras_tuner/
[24] dimensionality of the output space

middle by MaxPooling1D[25] pooling layer[26] with size 1. The following are layers and hyper parameters, the hyper parameters are set based on experiment and using the keraTuner.

- Model = Model = Sequential Keras.model.Conv1D
- Pooling = MaxPooling 1D
- Number of Input Neurons = 35
- Number of dense layers = 3 (64, 64 and 18 units)
- Training Ratio=70%, Validation Ratio=30%
- Test data 20%
- Epochs=30
- Batch size = 128
- Activation function = relu, softmax
- optimizer= Adam,
- loss function ='sparse_categorical_crossentropy', metrics = accuracy

4.5.3. Proposed LSTM Model

The proposed LSTM layer is keras sequential model from Keras Layers API. LSTM layer with input shape of training data. The dense and drop out layers are added to the sequential model. The hyper parameters selected from kerasTuner and based on experiment are listed as follows.

- Model = Model = Sequential Keras.Model
- Layer = Keras.Layesrs.LSTM
- Number of Input Neurons = 35
- Number of dense layers = 3 (128, 64 and 18 units)
- Training Ratio=70%, Validation Ratio=30%
- Test data 20%
- Epochs=80
- Batch size = 128
- Activation function = relu, softmax
- optimizer= Adam,
- loss function ='sparse_categorical_crossentropy', metrics = accuracy

4.6. Proposed data processing Implementation

4.6.1. Loading Dataset and Preparing for Feature Extraction

Dataset is loaded into the python environment before any data is processed. To load the dataset into the python environment, the *librosa.load ()* method is used to locate the path of the audio data in the google drive and load to google colab iteratively and transform it into metadata. The audio dataset is loaded as shown in the following sample code.

[25] Down samples the input representation by taking the maximum value over a spatial window of size
[26] A pooling layer is a new layer added after a convolutional layer.

```
y, sr = librosa.load(speech,
mono=True, duration=30)
.
```

Algorithm 4 Loading Speech Samples to Jupyter

The librosa load() method the extracts the audio time series[27] 'y' and 'sr' which stands for sample rate. By default librosa sample rate is initialized to 22050 hz. After the data is loaded librosa resamples the sample rate to 48000hz, and set the chanell to mono. Then remaining trailing and leading silences are romoved.

```
y, index = librosa.effects.trim(y)
```

Algorithm 5 Removing Remaining Silence

4.7. Proposed Feature extraction implementation

After the prepared files are located and loaded the next step is to extract the features and save the scores. 30 MFCC feature vectors and mean score of zero crossing, spectral roll-off, Chroma gram, spectral centroid, and spectral bandwidth are extracted as follows

```
mfcc = librosa.feature.mfcc(y=y, sr=sr n_mfcc=30)
chroma_stft = librosa.feature.chroma_stft(y=y, sr=sr)
spec_cent = librosa.feature.spectral_centroid(y=y, sr=sr)
spec_bw = librosa.feature.spectral_bandwidth(y=y, sr=sr)
rolloff = librosa.feature.spectral_rolloff(y=y, sr=sr)
zcr = librosa.feature.zero_crossing_rate(y)
rmse = librosa.feature.rms(y=y)
```

Algorithm 6 Feature Extraction

The extracted features are then exported to .CSV file.

4.7.3. Loading the CSV and Labeling

The data set features are extracted and saved to .csv file in the drive, know we load the csv file for further processing. The features are now labeled to their corresponding speaker. The sample of data of a single speaker is 30 each one second long know the features are extracted and saved with their given tag. We must organize the tag by matching the tag with their label. The following sample code shows how the labeling is implemented.

[27] Audio sounds can be thought of as an one-dimensional vector that stores numerical values corresponding to each sample

```
for i in range(len(filenameArray)):
    speaker = filenameArray[i][0] + filenameArray[i][1]
    #print(speaker)
    if speaker == "aa":
       speaker = "0"
    elif speaker == "bg":
       speaker = "1"
    elif speaker == "di":
```

Algorithm 7 Labeling Speakers Tag to Corresponding Number

Finally, the class ID is given based on the alphabetical order from 1 up to 16 and 17 is given for an "unknown".

Table 4-1 Labeling The Speech Samples To Some Corresponding Number.

No	**Speaker Name**	**Tag**	**Wav File Name**	**Id**
1	Abiy Ahmed	aa	aa001.wav --	0
2	Bekele Gerba	bg	bg001.wav --	1
3	Dawud Ibsa	di	di001.wav --	2
4	Fantu Demissew	fd	fd001.wav --	3
5	Hachalu Hundesa	hh	hh001.wav --	4
6	Hangasa Ibrahim	hi	hi001.wav --	5
7	Henok Gabisa	hg	hg001.wav --	6
8	Hiskiel Gabisa	Hg	Hg001.wav --	7
9	Jambo Jote	jj	jj001.wav ---	8
10	Jawar Mohamed	jm	jm001.wav ---	9
11	Kejela Merdasa	km	km001.wav ---	10
12	Lencho Leta	ll	ll001.wav ---	11
13	Mihretu Shanko	ms	ms001.wav ---	12
14	Shimelis Abdisa	sha	sha001.wav ---	13
15	Takele Uma	tu	tu001.wav ---	14
16	Taye Dende'a	td	td001.wav ---	15
17	Unknown	-	-	16

The tag of the speakers is given using two small letters from the first letter of their first name and last name.

4.8. Proposed Model Implementation

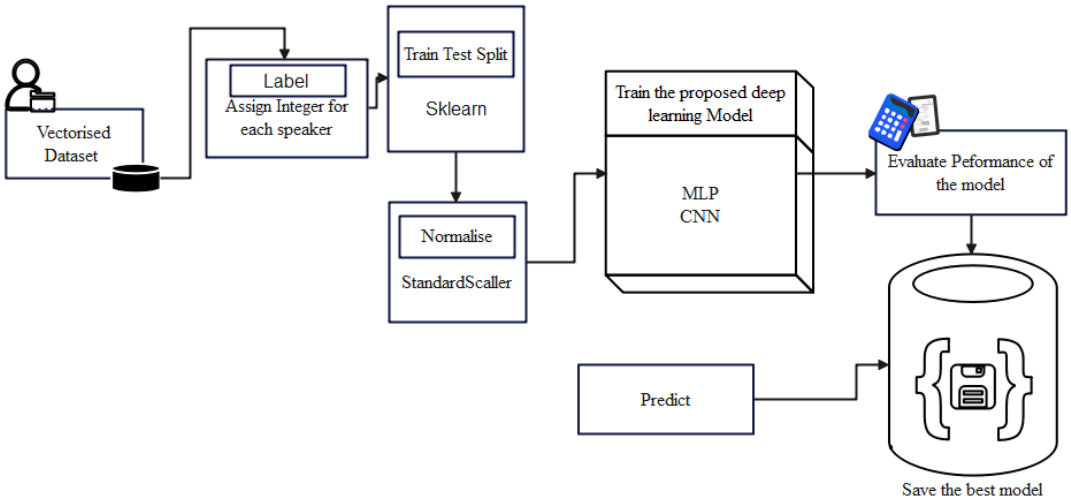


Figure 4-7 Proposed Model Implementation

After the dataset is preprocessed and represented as a vector, the next step is building the deep learning algorithms and training the model with the training dataset. As discussed in section 3.8 this is a study proposed with Keras Models API.

The *train_test_split ()* function of the sklearn_model_selection library is used to divide the model into train test split. In this thesis the T

For the first two experiments, the languages are used interchangeably for training and testing. When the Afaan Oromo data is used for training the Amharic data set is used for test purposes. In either case, we used a 70/30 training and validation proportion. And 20% for test

The random state parameter controls how the data is shuffled before doing the split. For the third and fourth experiments when training and testing are done with the same language. We used 70 training and 30 validation.

The Test split with done manually by calculating the ratio of the dataset 15% is used. And taken by the sklearn test split. The Train test and validation data are then divided into two dimensions X and Y. X Train, Y Train, and X validation are the first 35 columns, and the last column is the label column for Y Train, Y Validation, and Y Test. Sklearn standard scaler is used as internal data normalization after the dataset is converted to pandas Numpy Array.

The following sample code shows the normalization done by sklearn standard scalar

```
from sklearn.preprocessing import StandardScaler
import numpy as np
scaler = StandardScaler()
X_train = scaler.fit_transform( X_train )
X_val = scaler.transform( X_val )
X_test = scaler.transform( X_test)
```

Algorithm 8 Normalizing the Dataset

4.8.1. Implementation of MLP Model

During the feature extraction stage, the speech signal is transformed into feature vectors in a manner suitable for training a neural network. The next step is training and testing the neural network. After setting the training parameters, a set of inputs with corresponding target outputs is fed sequentially to the neural network. A sequential model is a linear stack of layers. It can be created a sequential model by passing a list of level instances to the constructor.

```
model = models.Sequential()
model.add(layers.Dense(256, activation='relu',
                       input_shape=(X_train.shape[1],)))
model.add(layers.Dropout(0.5))
model.add(layers.Dense(128, activation='relu'))
model.add(layers.Dropout(0.5))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dropout(0.5))
model.add(layers.Dense(20, activation='softmax'))
```

Algorithm 9 Implementation of MLP

4.8.2. Implementation of CNN model

CNN is a highly performing model especially in the area of computer vision. Recently many researchers are using CNN for different areas including speaker recognition. With the help of high per parameters that resulted good performance the model is built as shown in the sample code bellow

```
cnn_model = Sequential()
cnn_model.add(Conv1D(512,1,1,
        input_shape=(X_train.shape[1],1),activation='relu'))
cnn_model.add(MaxPooling1D(pool_size=1))
cnn_model.add(Conv1D(128, 1, 1, activation='relu'))
cnn_model.add(MaxPooling1D(pool_size=1))
cnn_model.add(Dropout(0.25))
cnn_model.add(Flatten())
cnn_model.add(Dense(64, activation='relu'))
cnn_model.add(Dropout(0.5))
cnn_model.add(Dense(64, activation='relu'))
cnn_model.add(Dense(18, activation='softmax'))
```

Algorithm 10 Implementation of CNN

4.8.3. Implementation of LSTM model

Long short-term memory (LSTM) is an artificial neural network used in the fields of artificial intelligence and deep learning. With the help of high per parameters that resulted good performance the model is built as shown in the sample code bellow

```
lstm_model = Sequential()
lstm_model.add(layers.LSTM(512,  input_shape=(X_train.shape[1],1)))
lstm_model.add(Dropout(0.2))
lstm_model.add(Dense(128, activation='relu'))
lstm_model.add(Dropout(0.2))
lstm_model.add(Dense(64, activation='relu'))
lstm_model.add(Dense(18))
lstm_model.add(Activation('softmax'))
```

Algorithm 11 Implementation of LSTM

After the models are built and compiled they are fitted to be trained on the prepared data.

4.9. Implementation Evaluation Methods

We evaluated the performance of the model using different evaluation metrics that are listed and described in section 7 different methods are designed to calculate the EER, and accuracy also to print the confusion matrices. The implementation of the Evaluation metrics is available in

CHAPTER FIVE

5. RESULT AND DISCUSSION

5.1. Chapter Overview

In this chapter, the results are discussed based on the experiments done concerning the prior research done on bilingual speaker recognition and speaker recognition developed for Afaan Oromo and Amharic. The results are taken from experiments done on the prepared dataset. The experiments done on different scenarios and using different models built are discussed in terms of the evaluation metrics mentioned in section 7. The study also discusses a comparison of the present bilingual speaker recognition and prior works on the state of the art. The study compares the speaker recognition models that are done with a single language for the languages selected for this study however, there is no prior research available on a public source that has been done on Afaan Oromo speaker recognition but there are few for Amharic therefore those prior study's will also be compared for further briefing.

5.2. Results

The dataset is collected from bilingual speakers whose recordings are available on YouTube, it is prepared to hold 30 utterances each 1 sec for one language and equal size for the second language. Total of 60 utterances from 16 speakers which is a total of 960 utterances. Experiments done and results obtained on the selected models are discussed below.

5.2.1. Result of MLP

A supplement of feed-forward neural network Multilayer perceptron LSTM and CNN models are presented in this study. The data set is given to those models built with their optimal hyper parameter and tested accordingly. The MLP model was built with optimum hyper parameters listed in section 4.5. Trained with a bilingual dataset consisting of equal utterance Amharic and Afaan Oromo speech utterances and tested with both and achieved 95.46% test accuracy. The model is well fit on 175 epoch.

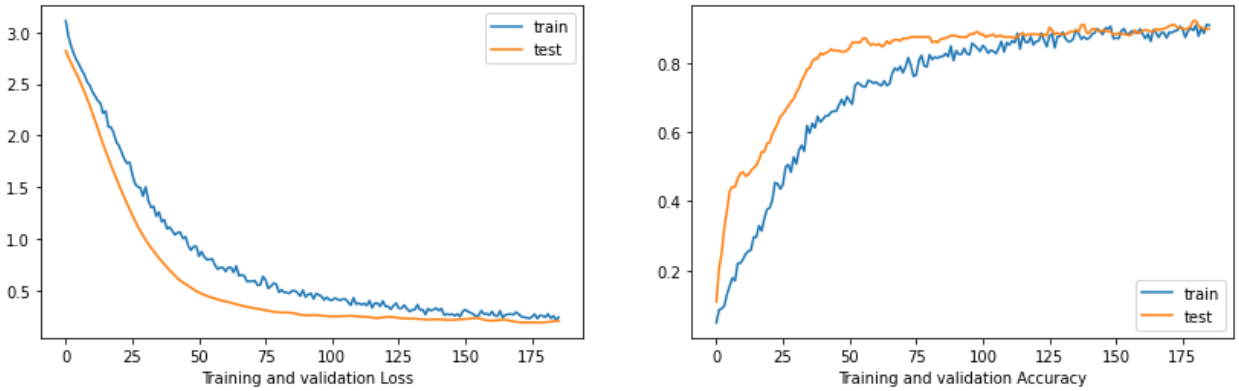


Figure 5-1 Learning Curve Diagram of MLP

The EER of the MLP model is 0.43 while the validation loss is 0.09.

Averaged F1 score is measured using the precision and recall values. It is measured between 0-1 therefore if the result is closer to 1 the model is performing well. In this study f1-score result is 0.88. The result of all 16 speakers is also presented.

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10
1	0.90	0.90	0.90	10
2	0.91	1.00	0.95	10
3	1.00	1.00	1.00	10
4	1.00	0.90	0.95	10
5	0.91	1.00	0.95	10
6	1.00	0.80	0.89	10
7	1.00	1.00	1.00	10
8	0.91	0.91	0.91	11
9	0.75	0.90	0.82	10
10	0.91	1.00	0.95	10
11	1.00	1.00	1.00	21
12	1.00	0.92	0.96	12
13	1.00	0.88	0.93	16
14	0.45	0.62	0.53	16
15	0.40	0.25	0.31	16
16	0.91	1.00	0.95	10
accuracy			0.87	202
macro avg	0.89	0.89	0.88	202
weighted avg	0.87	0.87	0.86	202

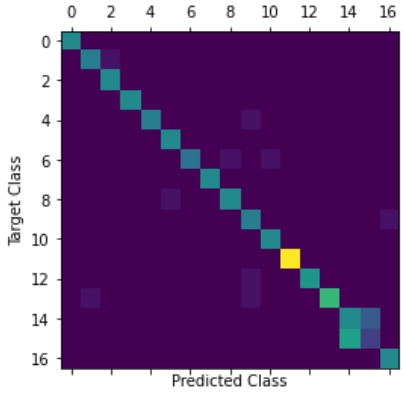


Figure 5-2 Shows the result of the f1-score for all speakers using MLP

The model is tested using different scenarios by interchanging the languages and the following result was presented as follows.

Table 5-1 Comparison of MLP Bilingual speaker recognition test result

Training language	Test language	Accuracy %
[AM and OR]	[AM and OR]	95.16
[AM and OR]	[AM]	94.89
[AM and OR]	[OR]	97.73

5.2.2. Result of CNN model

CNN is one of the most popular models used today. In this study 1D convolution, a neural network is deployed with the Keras Models API. The model is designed with the optimum hyper parameter to be trained bilingual dataset consisting of equal utterance Amharic and Afaan Oromo speech and achieved 91.02% test accuracies. The model is well fit on 30 epochs.

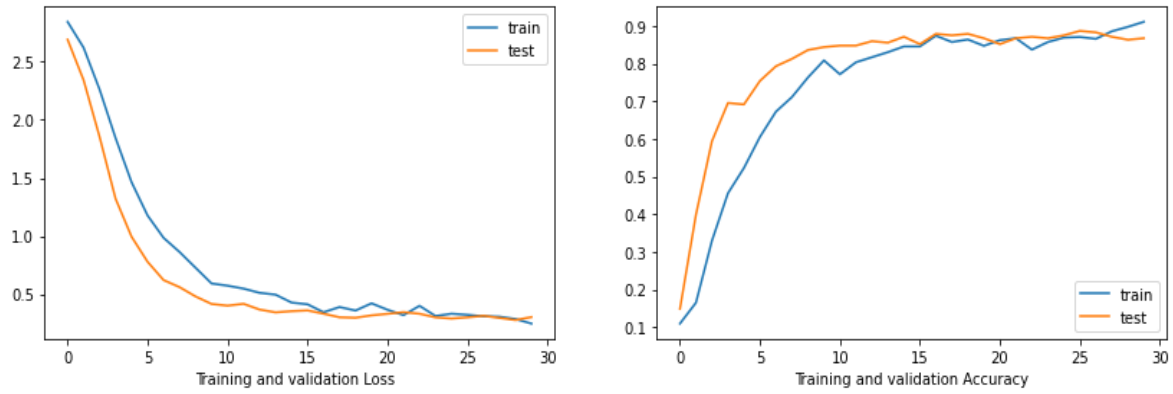


Figure 5-3 Learning Curve Diagram of CNN

The EER of the CNN model is 0.445 while the test loss is 0.3297.

Averaged F1 score is measured using the precision and recall values. It is measured between 0-1 therefore if the result is closer to 1 the model is performing well. In this study, the f1-score result is 0.85. The result of all 16 speakers is also presented.

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10
1	0.90	0.90	0.90	10
2	0.91	1.00	0.95	10
3	1.00	1.00	1.00	10
4	0.90	0.90	0.90	10
5	1.00	0.90	0.95	10
6	1.00	0.70	0.82	10
7	1.00	1.00	1.00	10
8	0.69	0.82	0.75	11
9	1.00	0.90	0.95	10
10	0.91	1.00	0.95	10
11	1.00	1.00	1.00	21
12	1.00	0.92	0.96	12
13	0.88	0.94	0.91	16
14	0.43	0.62	0.51	16
15	0.33	0.19	0.24	16
16	0.91	1.00	0.95	10
accuracy			0.85	202
macro avg	0.87	0.87	0.87	202
weighted avg	0.85	0.85	0.85	202

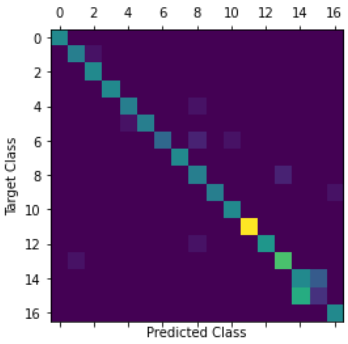


Figure 5-4 Shows the result of f1-score for all speakers using CNN

The result of the CNN model is summarized as follows

Table 5-2 Summary of CNN Bilingual speaker recognition test result

Training language	Test language	Accuracy (%)
[AM – OR]	[AM – OR]	91.02
[AM – OR]	[AM]	90.91
[AM – OR]	[OR]	98.30

5.2.3. Comparison of CNN, LSTM and MLP

This study aims to prepare a better-performing model for bilingual speaker recognition, as such in this section we provide an experimental result on how the CNN, LSTM and MLP speaker recognition models are affected by bilingual speakers. To show that we prepared 6 experiments by combining two languages Afaan Oromo and Amharic and testing every possibility. On the selected two models. The combinations show the result of monolingual bilingual and cross-lingual speaker recognition. In the experiments, the hyper parameter that resulted in an optimum result on the models is used to fit all combinations.

Result of Experiment 1: In this experiment, the model is first trained with Afaan Oromo and Amharic speech samples of each speaker and tested with Amharic.

Table 5-3 Result of experiment 1- [AM and OR] tested by [AM]

Model	Accuracy (%)	EER	Loss
CNN	90.91	0.420	0.59
LSTM	68	0.62	2.2
MLP	94.89	0.427	0.40

Result of Experiment 2: In this experiment, the dataset is bilingual for training and the model is first trained with Afaan Oromo and Amharic speech samples of each speaker and tested with Afaan Oromo.

Table 5-4 Result from experiment 2: [AM and OR] tested by [OR]

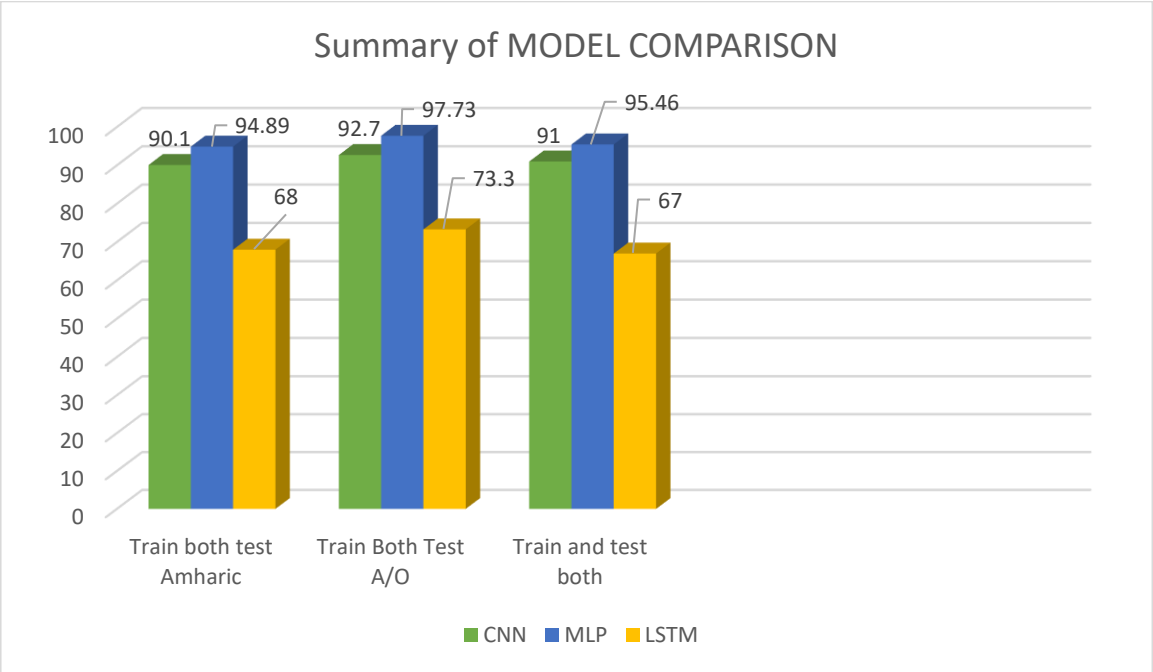
Model	Accuracy (%)	EER	Loss
CNN	98.30	0.43	0.24
LSTM	73.3	0.62	0.83
MLP	97.73	0.43	0.09

Result of Experiment 3: In this experiment, the dataset is bilingual for both training and testing. I.e. the model is first trained with Afaan Oromo and Amharic speech samples of each speaker tested with both parallel.

Table 5-5 Result of experiment 3: [AM and OR] tested by [AM and OR]

Model	Accuracy (%)	EER	Loss
CNN	91.02	0.44	0.59
LSTM	67	0.62	2.2
MLP	95.46	0.44	0.4474

5.2.3.1. Summary of Experiments



MLP out performed CNN and LSTM according to the experiments. Therefore based on the conducted experiments MLP is the best performing model for bilingual speaker recognition of Afaan Oromo and Amharic.

5.2.4. Effect of Monolingual, Bilingual, and Cross-lingual Tests

It is clear that the main objective of this research work is to design an optimum deep learning model for bilingual speakers of specified language. Even though both models have achieved good recognition performance MLP has resulted comparatively good result.

So by taking the best performed model we have conducted some other experiments to brief out the performance of the model on other language combinations, and to analyses the impact of language variation on the proposed speaker recognition model. As far as bilingualism is concerned in this study the necessity of this study is what we wanted to show in these additional experiments.

The combinations are classified as Monolingual, Bilingual, and Cross lingual. We have prepared 4 additional experiments, and generalized the effect of our proposed model.

Monolingual means using one language. This is a much easier task to produce good performance accuracy since the speaker is trained and tested in only one language the language effect is

relieved. The monolingual combinations are the once taken in Experiment 3 and 4 by training and testing the proposed model with the same language.

Cross lingual is using different language for training and test. Training with Afaan Oromo and testing to identify the same speaker while he/she is speaking Afaan Oromo. So the combinations are tested and reported in result of 5 and 6.

Result of Experiments 4 and 5

In these experiments, the dataset is trained and tested as monolingual speaker recognition. Experiment 4 is Train and test with Afaan Oromo only and Experiment 5 is Train and test with Amharic only, the result is presented in Table 5-6

Table 5-6 Result of Monolingual Testing Experiments

TEST	**Accuracy** (%)	**EER**	**Loss**
[OR] →[OR]	95.45	0.43	0.14
[AM]→ [AM]	96.59	0.43	0.14

Result of Experiments 6 and 7

In these experiments, the dataset is trained and tested as cross-lingual speaker recognition.

Table 5-7 Result of Cross-Lingual Experiments

TEST	**Accuracy** (%)	**EER**	**Loss**
[AM] → [OR]	33.40	0.83	8.87
[OR] → [AM]	39.77	0.70	4.54

The result of cross lingual test is very poor. Hence the proposed model cannot able to identify speakers in language mismatch conditions. The experiments showed that the performance of monolingual data is highly positive while the performance of cross-lingual test is poor.

Percentage change calculation:

Performance comparison of bilingual speaker recognition with respect to cross-lingual = Average accuracy of Experiment 1-3 – accuracy of Experiment 6&7

$$\text{(percentage change calculation)} \quad \frac{(V1 - V2)}{V1} * 100$$

$$\frac{(95.93 - 36.58)}{95.93} * 100 = 61.86\% \text{ decrease}$$

Performance degradation of multilingual to cross lingual = average accuracy of bilingual - Average accuracy of Cross-lingual test

$$\frac{(96.02 - 95.93)}{96.02} * 100 = 0.094\% \text{ decrease}$$

In terms of EER the multilingual speaker recognition showed 4% better EER than cross lingual recognition

Equation 7 Performance Drop Comparison

Bilingual to Monolingual	Bilingual to Cross lingual
0.094% drop	61.86% drop

The performance of the proposed model varies highly positively depending on language variation. Especially when the model is trained and tested in a cross-lingual context. This shows that having a bilingual model that is aware of both languages is the optimum way to handle bilingual speakers.

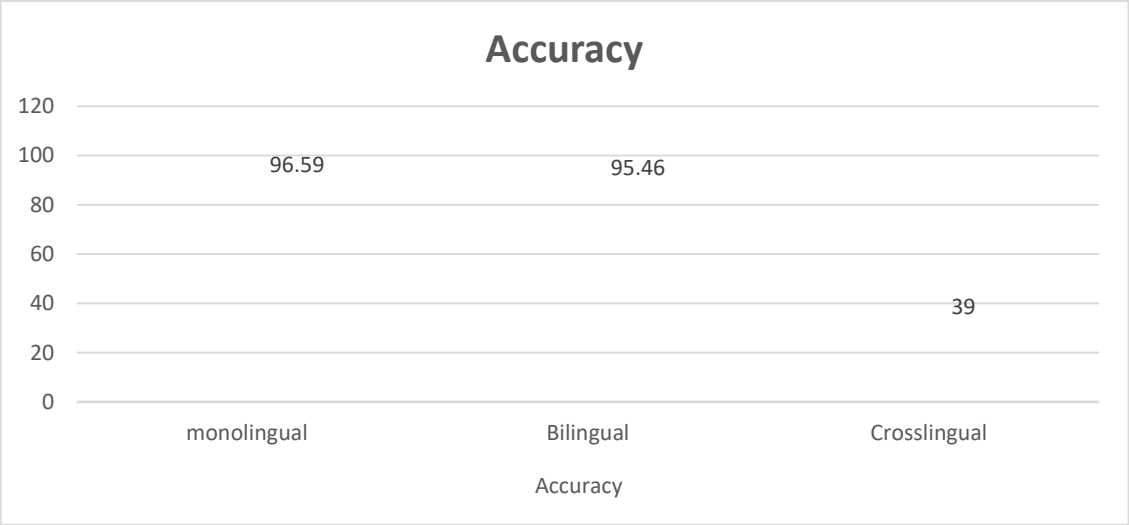


Chart 5-1 Performance Comparison of Monolingual Cross-Lingual and Bilingual Experiment Based On the Accuracy

According to experiment 4 and 5 above when the same model is trained with different language the performance rate shows significant degradation. Therefore for the proposed model it is

generalized that a speaker recognition model is highly affected by language. Hence it is necessary to consider researching on a models that can alleviate these problems.

5.3. Discussion

To our knowledge, this study is the first of its kind to be conducted on Afaan Oromo and Amharic bilingual speaker recognition. The study undertakes several experiments with a prepared dataset on three proposed models taking various language combination experiments to design the first-of-its-kind speaker recognition model supporting two Ethiopian languages. The CNN, LSTM and MLP models have been used and evaluated in this study. From the three models, MLP outperformed CNN in all experiments. 30-order MFCC is used to extract feature vectors from speech.

The study conducted by (Et. al. & Gizachew Belayneh Gebre, 2021) on ANN-based speaker recognition for the Amharic language. The researcher used Feed Forward Neural Network with MFCC and achieved 97.7 % using one hidden layer and 20 neurons. Therefore when this study is compared to our proposed models, the proposed model has achieved relatively less accuracy. The main reason is the researchers used only one language for both training and testing. On the other hand, they used record of 10 people while ours is of 16 people. The proposed model achieved 96.9% accuracy when trained and tested in a single language using 16 people's record.

Another study presented by (A. Mengistu & Alemayehu, 2017). This paper presents the performance analysis of text-independent speaker identification system for the Amharic language in noisy environments. VQ (Vector Quantization), GMM (Gaussian Mixture Models), BPNN (Back propagation neural network), MFCC (Mel-frequency Cepstrum coefficients), GFCC (Gammatone Frequency Cepstral Coefficients). According to their resulting experiment they have achieved 84.7% High classification accuracy using Back propagation neural network. When compared to our study we have done text independent recognition and used two languages. And achieved 96.28% average accuracy for monolingual test and 91.19%average accuracy for multilingual tests.

Table 5-8 Comparison of Related Studies

Parameters	**Previous study 1**	**Previous study 2**	**Previous study 3**	**This study**
Author	Abreham Debasu	Atklti Nigusie	Gizachew Belayneh	
No of speakers	50 speakers, 500 utterance	10 Speakers 40 utterances	10 speakers 20 utterance	16 speakers 30 sec Amharic and 30 sec Afaan Oromo
Implementation tool	MATLAB	MATLAB	MATLAB	Python
Feature Extraction	MFCC	MFCC	MFCC	MFCC
Speech language	Amharic	Tigrinya	Amharic	Afaan Oromo and Amharic
Classification algorithm	VQ and GMM	ANN	ANN	CNN and MLP
Approach	Text Independent	Text Dependent	Text In dependent	Text independent
Evaluation	Accuracy and subjective evaluation	Accuracy and confusion matrix	Accuracy and confusion matrix	Accuracy, EER and confusion matrix
Best result achieved	74% using VQ and 84.3 using GMM	96.6% for word and 93.7% for phrase	96%, 96.7%, 97.3% using 15, 20, and 25 number of hidden neurons respectively	• Average 95.93 using MLP, for bilingual test • 96.59 AM-AM, MLP • OR-OR 95.45% Afaan Oromo MLP

Finally, the research provided an answer to the research questions in this study based on the experiments, the study poses four research questions and those questions are answered in this section.

Research Question 1:- What is the impact of bilingual speakers on speaker recognition models? This research question deals with evaluating the performance impact of a model that performed well when trained and tested with a single language with respect to training both languages and testing one. The comparison is done by training a MLP model in three different experiments in experiments 5 & 6 as stated in Table 5-6 the MLP model performed achieved 96.02% average accuracy in both languages monolingual test. The same model was proposed for two languages using training and testing at the same time. As presented in Table 5-5 the model performance drop resulting 4% increase in equal error rate and 0.094% performance drop. This justifies that giving a training dataset for a well-performing monolingual model could not result in similar identification performance. Therefore bilingual speakers affect the performance of the speaker recognition model.

Research question 2: - Which model is best to handle the bilingual Afaan Oromo and Amharic languages, with better performance? This question intends to pick the best-performing model out of the proposed models. As depicted in Table 5-5 MLP outperformed CNN and LSTM in all experiments. Both proposed models are trained in 2 languages with their optimum hyper parameter. In the first experiment as presented in Table 5-3, when tested with Amharic MLP achieved 94.89% accuracy while CNN accomplished 90.91% and LSTM 68%. In the second experiment Table 5-4 when the model is tested with Afaan Oromo LSTM 73.30% CNN achieved 98.30% while MLP accomplished 97.73% accuracy. In the third experiment when the models are tested with the dataset containing both languages CNN achieved 91.02% while MLP accomplished 95.46%. Therefore MLP is the best performing than CNN and LSTM according to the experimental results.

Research Question 3: - What is the effect of cross-lingual tests in speaker recognition? This question gives analytical answer to one of the biggest problems in the bilingual speaker recognition area. Speaker recognition models face performance degradation when they are trained with one language and tested with other language. For this we have prepared cross lingual experiments and compared them with the performance of that of bilingual and

monolingual. Therefore the cross lingual speaker recognition test resulted in 13% increase in EER.

Research Question 4: - What is the effect of monolingual, and bilingual tests on the bilingual ASR? This question requires a performance evaluation answer from experiments done with a model trained in both languages and tested in all the monolingual and bilingual possibilities. The monolingual possibilities are tested with Afaan Oromo and Amharic one at a time. With Amharic the bilingual trained model, and both at the same time. Thus, the experiments are done and the results shown in Table 5-3, and Table 5-4, show the result of testing the model in Amharic and Afaan Oromo ensuing 0.42 EER while the test done with both languages at a time delivers shows an increase in 2%EER. Therefore testing a bilingual model with bilingual dataset results in an average rise of 2%EER.

Conclusion and Recommendation

Conclusion

Speaker recognition is interesting and wide area of research, in real time application it have several application areas. Those applications available in the developed countries can easily be adopted with the help of more research like this. To get the benefits of the speaker recognition plenty of researches have to be done to end up with reliable result, but recently that is not what is happening. Yet the researches proposed to date are not enough. There is no even single Afaan Oromo ASR research, as well as bilingual ASR. The more the researchers contribute the easier it could be to get more professional and conventional data set. This study is presented to contribute in the huge gap of research and to provide Ethiopians with the merit of ASR. This thesis work presented a deep learning bilingual speaker recognition model explicitly for two highly spoken languages in Ethiopia

The ultimate aim of this work was to enable a bilingual text-independent speaker recognition model that handles Afaan Oromo and Amharic speakers optimally using a novel modeling approach. It addressed the problem of having only a single language model for bilingual speakers by training both languages and testing both languages, together and each at a time. Amharic speaker recognition has been done by a few researchers while in Afaan Oromo there is none to mention as our knowledge. So there is a questions like why do we need a bilingual model? What is the impact of bilingual speakers on speaker recognition model? Which model can be the best to handle the bilingual SR of Afaan Oromo and Amharic? What is the impact of monolingual, bilingual, and cross-lingual SR test? So this study answered these questions statistically based on experiment.

To prepare the study in this area we must first prepare a dataset since there is no any public professional and conventional corpus prepared. We have prepared 960 speeches from 16 people who can meet certain criteria from YouTube. We have tried to reduce side effects of recording device, health status and mood variation of the target speaker between his/her Amharic and Afaan Oromo speech. Since the speech taken is a recording from multimedia the noisy, non-target speakers' voice and other unnecessary noises are removed and preprocessed with different software and tools because the clarity of the dataset has a direct impact on the model. MFCC

feature extraction is used with LSTM, CNN and MLP to satisfy the goal. 30 MFCC coefficients are employed to change the speech to a feature vector. So that it could be appropriate to train the neural network. The 1D convolutional neural network long short term memory, and Multilayer perceptron is trained with Keras Models API. We have prepared 3 major experiments that have a direct impact on the goal of our research by training both languages and testing both, testing Afaan Oromo and testing Amharic. An average of 95.93% recognition rate has achieved with MLP.

Finally, the experimental results showed that a speaker recognition model have significant performance variation and we witnessed the selected convolutional neural network is effective and suitable for identification processes since MLP outperformed CNN and LSTM. This study contributed a novel attempt at modeling techniques for bilingual Amharic and Afaan Oromo language speaker recognition. In addition to the main objective, this research has conducted an ice-breaking speaker recognition model for Afaan Oromo, prepared a multilingual dataset, and examined the effect The availability of this research will create a chance for the coming researchers who have a willingness in this research area to conduct comparative analysis on modeling techniques, especially for Amharic and Afaan Oromo language speaker recognition. Though it is not the ultimate goal the conducted dataset and code is uploaded as a GitHub repository for further study.

Recommendation

Because of our technical limitation and time constraint to learn, this research has left various productive tools and techniques that resulted good performance like Khalid, deep speaker embedding, Time Delay Neural Net (TDNN) and NEMO are kept aside. Therefore, since we have seen those tools but never used or tested them in this study, we recommend interested researchers to use these tools with MSRAMOR[28].

[28] GitHub repository link to our prepared dataset https://github.com/millionsime/MSRAMOR-Afaan-Oromo-and-Amharic

Future Work

The coming researcher can extend this research by adding the following features

- Work on cross-lingual speaker recognition in Ethiopian languages.
- Prepare a professional large dataset
- Measure the performance of speaker recognition in scarce data set and prepare a method to come up with better solution.
- Develop a bilingual speaker recognition with additional languages in Ethiopia

References

Agrawal, P., Shukla, A., & Tiwari, R. (2009). Multi lingual speaker recognition using artificial neural network. *Advances in Intelligent and Soft Computing*, *61 AISC*(May), 1–9. https://doi.org/10.1007/978-3-642-03156-4_1

Anukriti, Tiwari, S., Chatterjee, T., & Bhattacharya, M. (2013). Speaker independent speech recognition implementation with adaptive language models. *Proceedings - 2013 International Symposium on Computational and Business Intelligence, ISCBI 2013*, *August*, 7–10. https://doi.org/10.1109/ISCBI.2013.9

Ashok, J., Shivashankar, V., & Mudiraj, P. V. G. . (2010). An overview of biometrics. *International Journal on Computer Science and Engineering*, *02*(07), 2402–2408. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:An+overview+of+biometrics#0

BasuMallick, C. (n.d.). *Top 10 Speech Recognition Software and Platforms in 2022*. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/speech-recognition-software/

Brownlee, J. (n.d.). *How to Code a Neural Network with Backpropagation In Python*. https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., & Cowan, B. (n.d.). *The State of Speech in HCI : Trends , Themes and Challenges*.

Cohen-mcfarlane, M., Member, I. S., & Goubran, R. (2019). Comparison of Silence Removal Methods for the Identification of Audio Cough Events. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1263–1268.

Dave, N. (2013). Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition. *International Journal for Advance Research in Engineering and Technology*, *1*(Vi), 1–5.

Elnaggar, O., & Arelhi, R. (2020). A New Unsupervised Short-Utterance based Speaker Identification Approach with Parametric t-SNE Dimensionality Reduction. *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, *January*, 92–101. https://doi.org/10.1109/ICAIIC.2019.8669051

Et. al., G. B. G. (2021). Artificial Neural Network Based Amharic Language Speaker Recognition. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 5105–5116. https://doi.org/10.17762/turcomat.v12i3.2043

Et. al., G. B. G., & Gizachew Belayneh Gebre, E. al. (2021). Artificial Neural Network Based Amharic Language Speaker Recognition. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 5105–5116. https://doi.org/10.17762/turcomat.v12i3.2043

Faghfouri, A. E., & Frish, M. B. (2011). Robust discrimination of human footsteps using seismic signals. *Unattended Ground, Sea, and Air Sensor Technologies and Applications XIII*, *8046*(May 2011), 80460D. https://doi.org/10.1117/12.882726

Faundez-zanuy, M., Satué-villar, A., Universitària, E., Mataró, P. De, Politècnica, U., & Upc, D. C. (2014). *SPEAKER RECOGNITION EXPERIMENTS ON A BILINGUAL DATABASE*. *January 2006*, 2–6.

ffmpeg Documentation. (n.d.).

Franti, E., Ispas, I., Dragomir, V., Dascalu, M., Zoltan, E., & Stoica, I. C. (2017). Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots. *Romanian Journal of Information Science and Technology*, *20*(3), 222–240.

Furui, S. (1981). Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(2), 254–272. https://doi.org/10.1109/TASSP.1981.1163530

Furui, S. (2009). *40 Years of Progress in Automatic Speaker Recognition*. 1050–1059.

Guodong Guo, West Virginia University, U., Hugo Proenca, University of Beira Interior, P., & Massimo Tistarelli, University of Sassari, I. (n.d.). *Visual Surveillance and Biometrics: Practices, Challenges, and Possibilities*. IEEE Access. Retrieved March 6, 2022, from

https://ieeeaccess.ieee.org/closed-special-sections/visual-surveillance-biometrics-practices-challenges-possibilities/

He, H., Yan, Y., Chen, T., & Cheng, P. (2019). Tree height estimation of forest plantation in mountainous terrain from bare-earth points using a DoG-coupled radial basis function neural network. *Remote Sensing*, *11*(11). https://doi.org/10.3390/rs11111271

Hofmann, N. (n.d.). *Speech recognition for human computer interaction*.

Hordri, N. F., Yuhaniz, S. S., & Shamsuddin, S. M. (2016). Deep Learning and Its Applications: A Review. *Conference on Postgraduate Annual Research on Informatics*, *October*, 1–5.

Hourri, S., Nikolov, N. S., & Kharroubi, J. (2021). Convolutional neural network vectors for speaker recognition. *International Journal of Speech Technology*, *24*(2), 389–400. https://doi.org/10.1007/s10772-021-09795-2

Hussain, M. G., Rahman, M., Sultana, B., Khatun, A., & Hasan, S. Al. (2021). Classification of Bangla Alphabets phoneme based on audio features using MLPC SVM. *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0, ACMI 2021*, *June*. https://doi.org/10.1109/ACMI53878.2021.9528088

Jalata, A. (2010). *TRACE : Tennessee Research and Creative Exchange What is next for the Oromo people ?*

Jones, K., Walker, K., Caruso, C., Wright, J., & Strassel, S. (2022). *WeCanTalk : A New Multi-language , Multi-modal Resource for Speaker Recognition*. *June*, 3451–3456.

Kabir, M. M., Mridha, M. F., Shin, J., Jahan, I., & Ohi, A. Q. (2021). A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. *IEEE Access*, *9*(May), 79236–79263. https://doi.org/10.1109/ACCESS.2021.3084299

Kaur, G., Srivastava, M., & Kumar, A. (2017). Analysis of feature extraction methods for speaker dependent speech recognition. *International Journal of Engineering and Technology Innovation*, *7*(2), 78–88.

Kekre, H. B., & Kulkarni, V. (2013). Closed set and open set Speaker Identification using amplitude distribution of different Transforms. *2013 International Conference on*

Advances in Technology and Engineering, ICATE 2013, *August 2016*. https://doi.org/10.1109/ICAdTE.2013.6524764

Kleynhans, N. T., & Barnard, E. (2005). Language dependence in multilingual speaker verification. *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, Langebaan, South Africa, January 2005*, 117–122. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.8363&rep=rep1&type=pdf

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

librosa — librosa 0. (n.d.).

Luengo, I., Navas, E., Sainz, I., Saratxaga, I., Sanchez, J., Odriozola, I., & Hernaez, I. (2008). Text independent Speaker identification in multilingual environments. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, *2004*, 1814–1817.

Markowitz, J. A. (2000). Voice biometrics. *Communications of the ACM*, *43*(9), 66–73. https://doi.org/10.1145/348941.348995

Matějka, P., Novotný, O., Plchot, O., Burget, L., Sánchez, M. D., & Černocký, J. H. (2017). Analysis of score normalization in multilingual speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2017-Augus*, 1567–1571. https://doi.org/10.21437/Interspeech.2017-803

Mengistu, A., & Alemayehu, D. (2017). Text Independent Amharic Language Speaker Identification in Noisy Environments using Speech Processing Techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, *5*, 109. https://doi.org/10.11591/ijeecs.v5.i1.pp109-114

Mengistu, A. D. (2017). *Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC , MFCC and GFCC*. *6*(5), 8–12.

Mfcc, T. (2017). *MFCC Features*. https://doi.org/10.1007/978-3-319-49220-9

Mohd, R., Isa, K., & Mohamad, S. (2021). A review on speaker recognition : Technology and challenges. *Computers and Electrical Engineering*, *90*(April 2020), 107005. https://doi.org/10.1016/j.compeleceng.2021.107005

Nagaraja, B. G., & Jayanna, H. S. (2013). Multilingual speaker identification by combining evidence from lpr and multitaper mfcc. *Journal of Intelligent Systems*, *22*(3), 241–251. https://doi.org/10.1515/jisys-2013-0038

Nawaz, S., Saeed, M. S., Morerio, P., Mahmood, A., Gallo, I., Yousaf, M. H., & Del Bue, A. (2021). Cross-modal speaker verification and recognition: A multilingual perspective. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1682–1691. https://doi.org/10.1109/CVPRW53098.2021.00184

Neural networks and deep learning. (n.d.).

Pandey, B., Ranjan, A., Kumar, R., & Shukla, A. (2010). Multilingual speaker recognition using ANFIS. *ICSPS 2010 - Proceedings of the 2010 2nd International Conference on Signal Processing Systems*, *3*. https://doi.org/10.1109/ICSPS.2010.5555759

Pethuru Raj, P. E. (n.d.). *Advances in computers*.

Poornima, S. (2016). *Basic Characteristics of Speech Signal Analysis*. *5*(4), 1–5.

Proakis, J. R. D. J. H. L. H. J. G. (2000). *Discrete-Time Processing of Speech Signals*. https://ieeexplore.ieee.org/book/5266102

rabiner.pdf. (n.d.).

Radial Basis Function Networks Definition _ DeepAI. (n.d.).

Rahman, M. (2012). *Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches*. *3*(11), 131–138.

Rashid, S. A. A. and N. K. A. (n.d.). *No Title*. https://www.intechopen.com/chapters/63970

Rosistem Barcode - Biometric Education. (n.d.).

Rozi, A., Wang, D., Li, L., & Zheng, T. F. (2017). Language-aware PLDA for multilingual speaker recognition. *2016 Conference of the Oriental Chapter of International Committee*

for Coordination and Standardization of Speech Databases and Assessment Techniques, O-COCOSDA 2016, *October*, 161–165. https://doi.org/10.1109/ICSDA.2016.7919004

Russo, F. A., Abdoli-e, M., Lu, Z., Jelen, A., Health, P., Players, P. S., & Square, D. (2018). *Language and Linguistics* (Vol. 35, Issue 3, pp. 82–83).

Science, C., & Zheng, F. (2016). *Vowel-Category based Short Utterance Speaker Recognition*. *March*. https://doi.org/10.1109/ICSAI.2012.6223387

Tadele, F., Wei, J., Honda, K., Zhang, R., & Yang, W. (2022). *Effect of Language Mixture on Speaker Verification: An Investigation with Amharic, English, and Mandarin Chinese BT - Artificial Intelligence and Security* (X. Sun, X. Zhang, Z. Xia, & E. Bertino (Eds.); pp. 243–256). Springer International Publishing.

Thienpondt, J., Desplanques, B., & Demuynck, K. (2020). *Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization*. *October*. https://doi.org/10.21437/Interspeech.2020-2662

Trujillo, F. (1994). *The Production of Speech Sounds*.

WAV vs. (n.d.).

Wave, T. (n.d.). *WAVE PCM soundfile format*.

Xia, W., Huang, J., & Hansen, J. H. L. (2019). Cross-lingual Text-independent Speaker Verification Using Unsupervised Adversarial Discriminative Domain Adaptation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2019-May*, 5816–5820. https://doi.org/10.1109/ICASSP.2019.8682259

Zhang, W. E. I., Duffy, V. G., & Linn, R. (1999). *Voice recognition based human-computer interface design Voice Recognition Based Human-Computer Interface Design*. *8352*(November 2018). https://doi.org/10.1016/S0360-8352(99)00080-7

Appendixes

A. Source Dataset and Implementation Code.

The complete source code and dataset is posted in the following GitHub repository https://github.com/millionsime/MSRAMOR-Afaan-Oromo-and-Amharic

B. Evaluation Metrics Sample code

```
def evaluate(sims, labels):
    # Calculate evaluation metrics
    thresholds = np.arange(0, 1.0, 0.001)
    fm, tpr, acc = calculate_roc(thresholds, sims,
        labels)
    eer = calculate_eer(thresholds, sims,
        labels)
    return fm, tpr, acc, eer
def calculate_roc(thresholds, sims, labels):
   nrof_pairs = min(len(labels), len(sims))
    nrof_thresholds = len(thresholds)

    tprs = np.zeros((nrof_thresholds))
    fprs = np.zeros((nrof_thresholds))
    acc_train = np.zeros((nrof_thresholds))
    precisions= np.zeros((nrof_thresholds))
    fms = np.zeros((nrof_thresholds))
    accuracy = 0.0
    indices = np.arange(nrof_pairs)
    # Find the best threshold for the fold

    for threshold_idx, threshold in enumerate(thresholds):
        tprs[threshold_idx], fprs[threshold_idx],precisions[threshold_idx],\
        fms[threshold_idx], acc_train[threshold_idx] = calculate_accuracy(threshold, sims, labels)
    bestindex = np.argmax(fms)
    bestfm = fms[bestindex]
    besttpr = tprs[bestindex]
    bestacc = acc_train[bestindex]
    return bestfm, besttpr, bestacc
def calculate_accuracy(threshold, sims, actual_issame):
    predict_issame = np.greater(sims, threshold)
    tp = np.sum(np.logical_and(predict_issame, actual_issame))
    fp = np.sum(np.logical_and(predict_issame, np.logical_not(actual_issame)))
    tn = np.sum(np.logical_and(np.logical_not(predict_issame), np.logical_not(actual_issame)))
    fn = np.sum(np.logical_and(np.logical_not(predict_issame), actual_issame))
```

```
    tpr = 0 if (tp+fn==0) else float(tp) / float(tp+fn) #recall
    fpr = 0 if (fp+tn==0) else float(fp) / float(fp+tn)
    precision = 0 if (tp+fp==0) else float(tp) / float(tp+fp)
    fm = 2*precision*tpr/(precision+tpr+1e-12)
    acc = float(tp+tn)/(sims.size + 1e-12)
    return tpr, fpr,precision,fm, acc
def calculate_eer(thresholds, sims, labels):
    nrof_pairs = min(len(labels), len(sims))
    nrof_thresholds = len(thresholds)
    indices = np.arange(nrof_pairs)
    # Find the threshold that gives FAR = far_target
    far_train = np.zeros(nrof_thresholds)
    frr_train = np.zeros(nrof_thresholds)
    eer_index = 0
    eer_diff = 100000000
    for threshold_idx, threshold in enumerate(thresholds):
        frr_train[threshold_idx], far_train[threshold_idx] = calculate_val_far(threshold, sims, labels)
        if abs(frr_train[threshold_idx]-far_train[threshold_idx]) < eer_diff:
            eer_diff = abs(frr_train[threshold_idx]-far_train[threshold_idx])
            eer_index = threshold_idx
    frr, far = frr_train[eer_index], far_train[eer_index]
    eer = (frr + far)/2
    return eer
def calculate_val_far(threshold, sims, actual_issame):
    predict_issame = np.greater(sims, threshold)
    true_accept = np.sum(np.logical_and(predict_issame, actual_issame))
    false_accept = np.sum(np.logical_and(predict_issame, np.logical_not(actual_issame)))
    n_same = np.sum(actual_issame)
    n_diff = np.sum(np.logical_not(actual_issame))
    if n_diff == 0:
        n_diff = 1
    if n_same == 0:
        return 0,0
    val = float(true_accept) / float(n_same)
    frr = 1 - val
    far = float(false_accept) / float(n_diff)
    return frr, far
```

C. Labeling Speakers Sample Code

```
def getSpeaker(speaker):
    speaker = str(speaker)
    if speaker == "0":
        return "AbiyAhmed"
    elif speaker == "1":
        return "BekeleGerba"
    elif speaker == "2":
```

```
        return "DawudIbsa"
    elif speaker == "3":
        return "FantuDemissew"
    elif speaker == "4":
        return "henokGabisa"
    elif speaker == "5":
        return "HiskelGabisa"
    elif speaker == "6":
        return "HachaluHundesa"
    elif speaker == "7":
        return "HangasaIbrahim"
    elif speaker == "8":
        return "JamboJote"
    elif speaker == "9":
        return "JawarMohamed"
    elif speaker == "10":
        return "KajelaMerdasa"
    elif speaker == "11":
        return "MihretuShanko"
    elif speaker == "12":
        return "ShimelsAbdisa"
    elif speaker == "13":
        return "TayeDendea"
    elif speaker == "14":
        return "TakalaUma"
    else:
        speaker = "Unknown"

def printPrediction(X_data, y_data, printDigit):
    print('\n# Generate predictions')
    for i in range(len(y_data)):
        prediction = getSpeaker(model.predict(X_data[i:i+1])[0])
        speaker = getSpeaker(y_data[i])
        if printDigit == True:
            print("Number={0:d}, y={1:10s}- prediction={2:10s}- match={3}".format(i, str(speaker), str(p
rediction), speaker==prediction))
        else:
            print("y={0:10s}- prediction={1:10s}- match={2}".format(str(speaker), str(prediction), speake
r==prediction))



def report(X_data, y_data):
    #Confution Matrix and Classification Report
    Y_pred = np.argmax(model.predict(X_data), axis=-1)
```

```
    y_test_num = y_data.astype(np.int64)
    conf_mt = confusion_matrix(y_test_num, Y_pred)
    print(conf_mt)
    plt.matshow(conf_mt)
    plt.show()
    print('\nClassification Report')
    target_names = ["AbiyAhmed","BekeleGerba", "DawudIbsa", "FantuDemissew", "henokGabisa", "
HiskelGabisa", "HachaluHundesa","HangasaIbrahim","JamboJote", "JawarMohamed", "KajelaMerd
asa", "MihretuShanko", "ShimelsAbdisa","TayeDendea", "TakalaUma", "Unknown"]
    print(classification_report(y_test_num, Y_pred))
```