

2021-03

# DESIGNING AN AUTOMATIC TEXT-INDEPENDENT AMHARIC LANGUAGE SPEAKER IDENTIFICATION

METADEL, ADUGNA

---

<http://ir.bdu.edu.et/handle/123456789/12393>

*Downloaded from DSpace Repository, DSpace Institution's institutional repository*



**BAHIR DAR UNIVERSITY**  
**BAHIR DAR INSTITUTE OF TECHNOLOGY**  
**SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES**  
**COMPUTING FACULTY**

**DESIGNING AN AUTOMATIC TEXT-INDEPENDENT AMHARIC**  
**LANGUAGE SPEAKER IDENTIFICATION**

By

**METADEL ADUGNA AGMAS**

**March 2021**

**BAHIR DAR, ETHIOPIA**

DESIGNING AN AUTOMATIC TEXT-INDEPENDENT  
AMHARIC LANGUAGE SPEAKER IDENTIFICATION

BY

Metadel Adugna Agmas

A THESIS SUBMITTED TO THE SCHOOL OF RESEARCH AND  
GRADUATE STUDIES OF BAHIR DAR INSTITUTE OF  
TECHNOLOGY, BDU IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN  
INFORMATION TECHNOLOGY IN THE COMPUTING FACULTY.

**ADVISOR: ABRHAM DEBASU (ASS. PRO.)**

BAHIRDAR, ETHIOPIA

March 12, 2021

## DECLARATION

I, the undersigned, declare that the thesis contains my work. In compliance with internationally accepted practices, I have acknowledged and referred to all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/ fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also suggest penal action from the sources which have not been properly cited or acknowledged.

Student Name: Metadel Adugna Signature \_\_\_\_\_

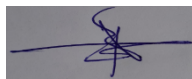
Date of submission: 02/26/2021

Place: Bahir Dar

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

Advisor Name: Abrham Debasu(Ass.Pro.)

Advisor's Signature:



## THESIS APPROVAL SHEET

### Student:

Name Metadel Adugna Signature -----Date 02/26/2021

The following graduate faculty members certify that this student has successfully presented the necessary written final thesis and oral presentation for partial fulfillment of the thesis requirements for the Degree of Master of Science in Information technology.

### Approved By:

Advisor: Name **Abrham Debasu (Ass.Pro.)** Signature  02/26/2021

External Examiner:

Seid Muhie Yimam (Ph.D)



| Name | Signature | Date |
|------|-----------|------|
|------|-----------|------|

Internal Examiner:

| Name | Signature | Date |
|------|-----------|------|
|------|-----------|------|

Chair Holder:

| Name | Signature | Date |
|------|-----------|------|
|------|-----------|------|

Faculty Dean:

| Name | Signature | Date |
|------|-----------|------|
|------|-----------|------|

## ACKNOWLEDGMENT

First of all, I would like to thank the Almighty of GOD for his unending blessings and for his helping through all challenges. Then, I am glad to express my heartfelt appreciation to my advisor Abrham Debasu (Assistant Professor) for his unreserved assistance, patience, outstanding guidance, and constructive comments throughout this thesis. Moreover, I would like to convey thanks to all my course lecturers and staff members of the school of Computing for their help and guidance on many occasions. Finally, I would like to acknowledge my beloved family and entire friends for their inspirational encouragement and moral and even financial support throughout this thesis.

## ABSTRACT

Speaker recognition is the technique of recognizing automatically, who is speaking on the source of individual information built-in in speech waves. It uses the speaker's voice to confirm their identity and provides control access for services such as voice calls, database access facilities, information services, voice mail, and security control for private information areas, and remote access to computers, and for other several fields.

Research within the area of speaker recognition has now crossed over five decades and has shown positive results under controlled conditions and having big data sets.

Besides, almost all of the previous study uses traditional methods like MFCC and LPCC as feature extraction and GMM, VQ and HMM, etc. as classifiers. These features generally require the domain knowledge for researchers to be designed and take a long time to be tested and selected for high classification performance.

And also concerning the requirement of realistic applications, the system should take into account the problems related to memory and computational resource limitation.

So, in this thesis, we tried to address the short utterance data sets in the real circumstance by applying an end-to-end Convolutional Neural Network and CNN with SVM. Since CNN training usually requires large sample data and has the overfitting problem. On the other hand, SVM has good generalization ability and can solve the small sample problem.

For the preparation of the data, audio signals are gathered from the Bahir Dar city by directly recording sounds from Amharic language speakers using an audio recorder. A total of 50 datasets were collected having 10 samples in each speaker. Each speaker had 10 seconds of waiting time to speak over the recorder, after being collected and pre-processed; each pre-processed voice was transformed into a spectrogram image.

Then, the spectrogram images were used as an input to the proposed model to learn speaker-specific spectral features.

To see the effect of gender on speaker identification for the Amharic language, we have experimented with 25 male and 25 female speakers. Then, we have found an accuracy of 76 % and 92 % for males and females respectively.

Finally, we compared the end-to-end and combination of CNN with SVM using a 50 speakers' voice dataset. Then we have achieved an accuracy of 82%, and 95%, for an end-to-end Convolutional Neural Network, and convolutional neural network with a support vector machine model respectively.

**Keyterms:-Convolution neural network (CNN), SVM (support vector machine), speaker identification, text-independent, Amharic language**

## TABLE OF CONTENTS

|                                                                           |     |
|---------------------------------------------------------------------------|-----|
| DECLARATION .....                                                         | II  |
| ACKNOWLEDGMENT .....                                                      | V   |
| ABSTRACT.....                                                             | VI  |
| TABLE OF CONTENTS.....                                                    | VII |
| ABBREVIATIONS LIST .....                                                  | X   |
| LIST OF FIGURE.....                                                       | XI  |
| LIST OF TABLE .....                                                       | XII |
| CHAPTER ONE.....                                                          | 1   |
| 1 INTRODUCTION.....                                                       | 1   |
| 1.1 Background .....                                                      | 1   |
| 1.2 Statement of the problem .....                                        | 3   |
| 1.3 Research question.....                                                | 4   |
| 1.4 Objectives of the study .....                                         | 4   |
| 1.4.1 General objectives .....                                            | 4   |
| 1.4.2 Specific objectives .....                                           | 4   |
| 1.5 Scope of the study and limitations of the study .....                 | 5   |
| 1.6 Significance of the study .....                                       | 5   |
| 1.7 Methodology.....                                                      | 5   |
| 1.7.1 Data collection and preparation.....                                | 6   |
| 1.7.2 Tools and techniques .....                                          | 6   |
| 1.8 Organizations of the study .....                                      | 7   |
| CHAPTER TWO.....                                                          | 8   |
| 2 LITERATURE REVIEW.....                                                  | 8   |
| 2.1 Introduction to Amharic (አማርኛ) Language and its characteristics ..... | 8   |
| 2.1.1 Amharic language Consonant.....                                     | 9   |
| 2.1.2 Amharic language Vowels .....                                       | 10  |

|                     |                                                          |    |
|---------------------|----------------------------------------------------------|----|
| 2.2                 | Human speech production.....                             | 10 |
| 2.2.1               | Introduction.....                                        | 10 |
| 2.3                 | Speaker recognition system .....                         | 20 |
| 2.3.1               | Speaker identification and speaker verification .....    | 21 |
| 2.3.2               | Text-independent and text-dependent .....                | 23 |
| 2.3.3               | Open-set and closed-set.....                             | 24 |
| 2.4                 | speakers modeling.....                                   | 25 |
| 2.4.1               | Generative model .....                                   | 25 |
| 2.4.2               | Discriminative models .....                              | 26 |
| 2.5                 | Feature Extraction .....                                 | 27 |
| 2.5.1               | Deep Learning Speech Feature Extraction Techniques ..... | 28 |
| 2.5.1.1             | convolution neural network (CNN) .....                   | 28 |
| 2.6                 | Evaluation technique.....                                | 37 |
| 2.7                 | Related works .....                                      | 38 |
| CHAPTER THREE ..... |                                                          | 42 |
| 3                   | DESIGN OF AUTOMATIC SPEAKER IDENTIFICATION .....         | 42 |
| 3.1                 | Introduction.....                                        | 42 |
| 3.2                 | Signal Preprocessing.....                                | 44 |
| 3.2.1               | A/D Conversion.....                                      | 44 |
| 3.2.2               | Pre-emphasis.....                                        | 45 |
| 3.2.3               | Framing.....                                             | 46 |
| 3.2.4               | Windowing.....                                           | 48 |
| 3.3                 | Silence Removal of Speech Signal .....                   | 48 |
| 3.3.1               | speech silence removal using frame by frame.....         | 49 |
| 3.4                 | Spectrogram .....                                        | 51 |
| 3.5                 | Resizing .....                                           | 53 |
| 3.6                 | Grayscale .....                                          | 53 |

|                   |                                                                               |    |
|-------------------|-------------------------------------------------------------------------------|----|
| 3.7               | Feature extraction.....                                                       | 54 |
| 3.7.1             | Introduction.....                                                             | 54 |
| 3.7.2             | Review of Feature Representations .....                                       | 54 |
| 3.9.1             | Convolution Neural Network (CNN) As Feature Extraction .....                  | 56 |
| 3.10              | Classification .....                                                          | 58 |
| 3.10.1            | Support Vector Machine (SVM).....                                             | 59 |
| CHAPTER FOUR..... |                                                                               | 60 |
| 4                 | EXPERIMENTAL RESULT AND DISCUSSION .....                                      | 60 |
| 4.1               | Introduction.....                                                             | 60 |
| 4.2               | Data Collection and Preparation .....                                         | 60 |
| 4.3               | Development Environment .....                                                 | 61 |
| 4.4               | Evaluation Techniques .....                                                   | 61 |
| 4.5               | Experiment on the proposed end to end CNN model.....                          | 63 |
| 4.5.1             | Experiment based on Activation function using proposed end-to-end CNN .....   | 63 |
| 4.5.2             | Experiment based on distinct gender group using proposed end-to-end CNN ..... | 64 |
| 4.5.3             | Experiment based on learning rate using proposed end-to-end CNN .....         | 67 |
| 4.6               | Experiment on proposed CNN-SVM model.....                                     | 68 |
| 4.7               | Comparison of the proposed model with the previous work.....                  | 68 |
| CHAPTER FIVE..... |                                                                               | 69 |
| 5                 | CONCLUSION AND RECOMMENDATION .....                                           | 69 |
| 5.1               | Conclusion.....                                                               | 69 |
| 5.2               | Recommendation.....                                                           | 69 |
| Reference.....    |                                                                               | 71 |

## ABBREVIATIONS LIST

|        |                                          |
|--------|------------------------------------------|
| ASR    | Automatic Speech Recognition             |
| VQ     | Vector Quantization                      |
| GMM    | Gaussian Mixture Model                   |
| MFCC   | Mel Frequency Cepstral Coefficient       |
| GFCC   | Gamatone frequency cepstral coefficient  |
| ASI    | Automatic speaker Identification         |
| ANN    | Artificial Neural Network                |
| DNN    | Deep neural network                      |
| SPI    | Speaker identification                   |
| GFM    | Generalized Fuzzy Model                  |
| CMVN   | Cepstral Mean and Variance Normalization |
| FW     | Feature Warping                          |
| UBM    | Universal Background model               |
| CNN    | Convolutional Neural network             |
| EM     | Expectation maximization                 |
| Matlab | Matrix laboratory                        |
| SVM    | Support vector machine                   |
| BPNN   | back propagation neural network          |

## LIST OF FIGURE

|                                                                                                    |    |
|----------------------------------------------------------------------------------------------------|----|
| Figure 1 the Amharic alphabet (adopted from LexiLogos website) .....                               | 9  |
| Figure 2 Speech chain system (adopted from Peter Denes, 2012) .....                                | 11 |
| Figure 3 Anatomical structure of the human vocal system (adopted from J. L. Flanagan, 2016). ..... | 12 |
| Figure 4 Illustration of changing vocal tract shapes (adopted from Ittansa K., 2014) .....         | 13 |
| Figure 5 Vocal tract articulations (adopted from J. L. Flanagan, 2010) .....                       | 14 |
| Figure 6 Spectral envelope and its formants (adopted from Maëva Garnier, 2013) .....               | 15 |
| Figure 7 Discrete-time speech production model.....                                                | 16 |
| Figure 8 Estimated speech production model. ....                                                   | 17 |
| Figure 9 Original signal of speech wave form .....                                                 | 18 |
| Figure 10 Waveform and spectrograms of a speech signal.....                                        | 19 |
| Figure 11: Speaker Recognition fundamental tasks (adopted from Tumisho B., 2018) .....             | 21 |
| Figure 12 Speaker Identification system (adopted from Tom Bäckström, 2015) .....                   | 22 |
| Figure 13 speaker identification system (adopted from Ronak Bajaj, 2018) .....                     | 23 |
| Figure 14 speaker recognition taxonomy (adopted from Tshephisho S., 2019) .....                    | 24 |
| Figure 15 general frameworks of CNN (Adapted from Henning Müller, 2020). ....                      | 29 |
| Figure 16 an example of a convolution operation with a kernel size of $3 \times 3$ .....           | 30 |
| Figure 17 Activation functions commonly applied to neural networks .....                           | 32 |
| Figure 18: Pictorial representation of max pooling and average pooling. ....                       | 32 |
| Figure 19 A routine checks for recognizing overfitting.....                                        | 37 |
| Figure 20 detail a diagrammatic representation of the Speaker identification system.....           | 43 |
| Figure 21: Magnitude responses and Phase response of a first order .....                           | 46 |
| Figure 22 Speech frame blocking and windowing.....                                                 | 47 |
| Figure 23: Comparison of Hanning and Hamming window .....                                          | 48 |
| Figure 24 frame by frame silence removal technique flow chart.....                                 | 50 |
| Figure 25: Original speech .....                                                                   | 50 |
| Figure 26: speech without silence .....                                                            | 51 |
| Figure 27 original signal.....                                                                     | 52 |
| Figure 28 spectrograms of the speech signal.....                                                   | 52 |
| Figure 29 grayscale image.....                                                                     | 54 |
| Figure 30 Training Progress of our end-to-end CNN model.....                                       | 62 |
| Figure 31 Proposed end-to-end CNN model Summary .....                                              | 63 |
| Figure 32 confusion matrix for female modeling.....                                                | 66 |

## LIST OF TABLE

|                                                                                                                                   |    |
|-----------------------------------------------------------------------------------------------------------------------------------|----|
| Table 1 Amharic consonant.....                                                                                                    | 9  |
| Table 2 Vowel Articulations .....                                                                                                 | 10 |
| Table 3 lists of parameters and hyperparameters of CNN .....                                                                      | 31 |
| Table 4 commonly applied last layer activation functions.....                                                                     | 33 |
| Table 5 comparisons of various feature extraction and recognition techniques for speaker identification in Amharic languages..... | 41 |
| Table 6 Spectrogram Image resizing algorithm .....                                                                                | 53 |
| Table 7 the structures of proposed CNN model .....                                                                                | 58 |
| Table 8 CNN model on different activation functions .....                                                                         | 64 |
| Table 9 Gender based CNN model.....                                                                                               | 65 |
| Table 10 compares the learning rate.....                                                                                          | 67 |
| Table 11 Comparison of the proposed model with previous works.....                                                                | 68 |

## CHAPTER ONE

### 1 INTRODUCTION

#### 1.1 Background

Ethiopia is a multi-linguistic country that has 83 different languages with up to 200 different dialects spoken. From these Amhara's, Afan Oromos and Tigrayans are the biggest linguistic groups (Degu 2014). Semitic, Cushitic, Omotic, and Nilo-Saharan are the major language families in Ethiopia.

The Amharic language is categorized under the Semitic language family and it is spoken by 30.3 million people as a first and second language; this is about 29.3% of the total population.

It is the second most-spoken Semitic language in the world next to the Arabic language and the second-largest language in Ethiopia after Afaan Oromo a Cushitic language, and possibly one of the five largest languages on the African continent (Asker 2010). Among these, more than 3 million speakers are living abroad in different countries (Joshua 2019).

The Amharic and Tigrigna languages use the Ethiopic Geez script. The Amharic language is to contain 33 basic and other additional symbols each basic symbol has seven different orders other than most languages there are five vowels /a, ē, i, o, u/. In addition to these five vowels Amharic has two additional vowels /e, ī / the orders are created by the combination of each consonant with the seven vowels (Abebe June 2018).

Language is the engine of civilization and the purpose of language is communication. In our ordinary lives, there are having many forms of communication languages, like body language, textual language, pictorial language, speech, etc (Liu JC 2018).

Speech is a natural way of transforming information between the speaker and the listener and it is always considered as the most powerful form than others because of its rich dimensions characters, Such as gender, attitude, emotion, health situation, and identity of a speaker (Yonas 2014).

Signal processing is an area that studies the analog and digital signal processing methods in the time-frequency domain. Processing a signal means that searching and extracting some useful spectral information from the given signal.

In the real world, signal processing and its techniques are being widely used in many application areas like text-to-speech translation, speech recognition, and speaker recognition (C. D. Acken 2016)

The speaker recognition and speech recognition terms are incorrectly interpreted as similar responsibilities. However, the two related areas are different in their application.

Speech recognition has to differentiate the spoken words or phrases (what is being said) in spoken language and change them into a machine-readable format, whereas the task of speaker recognition is to identify the speaker using his /her unique acoustic spectral features encompassed in voiceprints. Because, each speaker has his or her characteristic style of talking, including the use of a specific accent, intonation style, rhythm, choice of vocabulary, and articulation pattern (Schroeder 2014). And Speaker identification is the process of determining from which of the registered speakers a given utterance comes.

Speaker recognition systems are valuable in different application areas such as for Forensics and surveillance, remote attendance logging, Authentication systems, teleconferences, voice mail, speech recognition, Law Enforcement, voice dialing, mobile banking, online marketing, database access services, information services, security control for private information areas, and remote access to computers (Praveen P S 2017).

Speaker recognition can be divided into two fundamental applications based on its goal which is speaker identification and speaker verification. For the identification duty, an unknown speaker is associated with a dataset of known speakers, and the best matching speaker is measured as the identification result. For the task of verification, the system's purpose is to decide whether a voice sample was produced by the claimed person or not. Both speaker identification and verification applications can be divided into text-dependent and text-independent approaches.

In text-dependent systems, speaker recognition depends on communication channels that introduce additional distortions to the speech. These different noise sources alter speech production so that most speaker recognition systems are exposed to failure in noise corrupted environments (Chang J 2017).

In another way by considering the availability of voice signals in the record, speaker identification is categorized as open-set and closed-set. In closed set identification, the method identifies the talker from a list of records by matching the voice biometrics of the voices with those utterance models kept before in the file box and the system makes a forced decision just by picking the best matching speaker from the speaker's data set. However, in open-set identification, the system identifies the speaker from a list of given utterances using a predefined tolerance level between the unidentified speaker and the best matching speaker (M. M. Thiye Modipa 2019).

To develop a speaker recognition system, analysis, feature extraction, and feature classification components are playing important roles. Until now, most of the speaker identifications are designed using conventional and state-of-the-art Mel Frequency Cepstrum Coefficient feature vectors, Gaussian Mixture Models, vector quantization methods, and fusion methods. However, such methods are underprivileged by their dependency on handcrafted feature engineering, limited processing time, and weakness to noise, and performance degradation for short audio utterances. Because, if there is noise in our data we can't extract the identifiable parts of the speaker so it reduces the performance. Existing speaker modeling techniques, mostly based on clustering (vector quantization (VQ)) (Gray 1984) and statistical modeling (Gaussian mixture modeling (GMM)) work well under sufficient data conditions. Under limited data, the modeling may not be reliable due to the effect of either sparse distribution for clustering or insufficient data for the estimation of statistical model parameters. Further, during testing, the decision may not be reliable due to the limited number of feature vectors. To fill these gaps a deep neural network algorithm handling a sufficient amount of audio data set is proposed (Krishnamoorthy P 2015).

In this thesis, I used a deep learning Convolutional Neural Network approach. By the method, the recorded voices of each speaker have been changed into a time-frequency (spectrogram) representation by using digital processing techniques. Then, the generated spectrogram images are used as input into CNN to learn speaker-specific spectral features.

## **1.2 Statement of the problem**

Human-computer interaction increases in recent times and most devices are becoming 'smart' to connect with humans efficiently. The security demands are also increased, more traditional authentication methods, such as personal ID cards, keys, or passwords are used to identify individuals but they are easily exposed through theft, collusion, or loss and also difficult to remember the password. (Columbus 2020).

To improve these problems using biometrics studies and research to verify users becomes important. The biometrics technique tried to reduce this problem by authenticating and identifying individuals using such methods as the face, fingerprint, iris, voice, hand geometry, and others. In these methods face and iris scanning are suitable but they need the users constantly being in the presence of the camera, whereas fingerprint and hand-geometry involve the handler to hold the device physically. But, Voice avoids such types of limitations and it is a widely accepted method in biometrics for authentication and/or identification (J. Markowitz 2019).

The human voice is a perfect source of data for personal identification in numerous applications. With a growing demand for security in different public places, voice biometrics appears to be a good solution, because voice records can be easily taken. This is one of the reasons for us to do this study because there is no work within the Amharic language.

Also for the last five years, consistent researches have been done in speaker identification. But almost all are designed using traditional and state-of-the-art Mel Frequency Cepstrum Coefficient feature vectors, Gaussian Mixture Models, Vector quantization methods, and combination methods for classification. However, such methods are not good enough by their dependency on handcrafted feature engineering, limited processing time, and exposure to noise and performance degradation for short audio data sets (Chang J 2017)). In contrast, a convolutional neural network (CNN), which has been widely applied to signal to process, can learn to automatically extract features from the training data using convolution and subsampling, but CNN training usually requires large sample data and has the overfitting problem. On the other hand, SVM has good generalization ability and can solve the small sample problem. Therefore, we proposed a CNN-SVM combined model to make use of their advantages.

To this end, this thesis explored and answered the following research questions.

### **1.3 Research question**

1. Which speech processing technique is suitable for automatic text-independent Amharic language Speaker identification?
2. Which activation function is suitable for CNN based on automatic text-independent Amharic language speaker identification?
3. Which learning rate value is suitable for CNN based on automatic text-independent Amharic language speaker identification?
4. To what extent the identification rate is registered?

### **1.4 Objectives of the study**

#### **1.4.1 General objectives**

The main objective of this study is to design an automatic Text-Independent Amharic Language Speaker Identification.

#### **1.4.2 Specific objectives**

Depending on the general objective of this study the following specific objectives are identified to indicate the main goals of the research in details:

- Identifying the speech processing technique suitability for automatic text-independent Amharic language speaker identification.
- To determine a suitable activation function.
- Finding the suitable activation function and learning rate values for CNN based on automatic text-independent Amharic language speaker identification.
- To evaluate the performance of the proposed model.

### **1.5 Scope of the study and limitations of the study**

The study only considered the close-set text-independent speaker identification task in the Amharic language using deep learning methods. As a limitation, this study did not consider the verification, text-dependent, and open data set. Also, the speaker's mood, health condition, and speaker's session variability are not considered in this study.

### **1.6 Significance of the study**

The interaction of humans with computers is increased and voice has become the most accessible and acceptable, low-cost, difficult to forge biometric technique and speaker recognition the thesis has the following significances:

It is preferable to have speaker identification technology that provides good performance using limited speech data for training and testing. To increase the usability of this technology for personal identification as part of multi-factor authentication systems.

**Online Transactions:** In addition to a passphrase to access bank information or to purchase an item over the phone, one's speech signal can be used as an extra layer of security.

**Law Enforcement:** There are practical situations like forensic investigations where the availability of speech data is limited and technology is needed to identify the person using only limited data in Courts, police commissions to keep our society from criminal acts.

**Speech Data Management:** Voicemail services and annotation of recorded or live meetings can use speaker recognition to label speakers automatically.

Hospitals, health centers, and other health service providers can benefit from the study for Medical Remote Monitoring.

I hope it will be the baseline of the researcher.

It increases the usability of the speaker identification technology

### **1.7 Methodology**

In this thesis, as the research methodology approach, the researcher used an experimental research methodology. For an appropriate understanding of the gap related to speaker identification, speech analysis, speaker feature extraction techniques, speaker modeling algorithms, detail-related works of

literature such as books, journals, magazines, conference papers, manuals, and resources from the internet are reviewed. In line with this, the respective methodologies applied in each work are discussed.

### **1.7.1 Data collection and preparation**

We collected spontaneous data set to collect it we used primary sources of data gathering techniques using an audio recorder. There is no prepared data set in the Amharic language and it is the challenging area of speaker and speech recognition. In our dataset collection, we considered the Bahir dar city area, which also consists of both males and females. Having such a data set helps us to decide the potential use of speaker identification on varieties of speech samples.

This dataset contains a total of 50 speakers; all are speaking in the Amharic language. We collect 10 voice samples from each speaker; each voice sample is 10 seconds long since less than 15 seconds categorized into short utterances. Speakers' names are put as the file name of the voice sample name\_number like Metadel\_01. Because it helps us to evaluate which sample data is identified correctly and which is not. All the voice samples are collected according to this format.

The dataset is divided into two portions of training and testing, 75% of voice samples from each speaker are used for training and the remaining one is used for testing. Each sample is taken at a sampling rate of 16 kHz and 16 bit because most of the previous works used it. After that, all these data are properly preprocessed and the necessary features are extracted.

### **1.7.2 Tools and techniques**

To achieve the objective of this thesis we applied the following tools and techniques:

- ✓ We used the MATLAB 2019a 64 bit version for the study because it is a powerful tool for signal processing. MATLAB's functionality can be greatly expanded by the addition of sets of precise functions that delivered more particular functionality called toolboxes. For instance, an excel link allows data to be written in a format recognized by Excel, Statistics Toolbox allows more specialized statistical manipulation of data, etc.
- ✓ Wave Pad Sound Editor software is used to transform multiple Stereo (two-channel) audio signals to Mono (one channel) signals and to translate many mp3 files to wave format spontaneously.
- ✓ And a microphone Audio recorder is used for recording speech files.

To represent the overall architecture, we proposed training and testing phases. In the training phase graphical representation of audio data which is called spectrogram considered as an input sequence.

Appropriate training and testing algorithms were selected, training and testing were performed. Known speakers are classified and the speech feature of each speaker is tested against the available speaker classes to identify the identity of each speaker.

In feature extraction, a convolutional neural network (CNN) algorithm has been used.

Recent advancements are done in using a convolutional neural network for speaker identification tasks because they can easily handle noisy datasets and there is no need for feature engineering, as feature extraction and classification both can be done by CNN. The classification result gives the final method output.

## **1.8 Organizations of the study**

This thesis is organized into five chapters with the existing one. The first chapter discussed the general outline of the thesis. The rest of this thesis is organized as follows:

**Literature Review (Chapter 2)** - in this chapter, reviewed literature about the Amharic language, human speech production, speaker recognition systems, feature extraction, and feature classification models are presented in detail. Finally, and related works also discussed

**Methodology (Chapter 3)**- in this chapter, a method used for the thesis such as the design of the model, discussion about techniques used for dataset collection and preparation, audio signal preprocessing techniques, feature extraction, and classification model is described in detail.

**Experiment Result and Discussion (Chapter 4)** - in this chapter, experiment design and experimental results are discussed.

**Conclusion and Recommendation (Chapter 5)** - in the last chapter, the Conclusions section, sums up the key points of our discussion, the significant outcomes of our study. The recommendation part addresses limitations and suggests how they might be overcome in future work.

## CHAPTER TWO

### 2 LITERATURE REVIEW

#### 2.1 Introduction to Amharic (አማርኛ) Language and its characteristics

According to Ethnologue, 83 individual languages with 200 different dialects are spoken in Ethiopia. One of the ancient languages in Ethiopia is Ge'ez and it was introduced as an official written language during the first Aksumite kingdom when the Sabeans sought refuge in Aksum (Björn 2010). The Aksumites developed Ge'ez, which has a unique script derived from the Sabeian alphabet, and it is still used by the Ethiopian Orthodox Tewahedo Church mainly. Tigrigna and Amharic are the modern languages that are derived from Ge'ez. Amharic is the official national language of Ethiopia and it categorizes into Semitic groups. Generally, Ethiopian languages are divided into four major language groups. Those are Semitic, Cushitic, Omotic, and Nilo-Saharan (EthiopianTreasures.c.uk 2002). Amharic is spoken by about 30.3 million people as a first or second language; this is about 29.3% of the total population. It makes the second most-spoken Semitic language in the world next to the Arabic language and the second-largest language in Ethiopia after Afaan Oromo a Cushitic language, and one of the five largest languages on the African continent (B. Gamback 2018). Among these, more than 3 million speakers are living abroad in different countries (Joshua 2019). It ranks 55 in the number of the first number of speakers in the world. The Amharic language contains 33 basic and other additional symbols each basic symbol has seven different orders other than most languages there are five vowels /a, ä, i, o, u/. In addition to these five vowels Amharic has two additional vowels /e, ī / the orders are created by the combination of each consonant with the seven vowels. (Abebe 2018). Amharic is written (left to right) using ፊደል /Fidel/ or Amharic letter. The Amharic writing system is 231(7x33=231) core characters and there are over forty others which contain a special feature usually representing labialization and also Amharic characters do not use dots and the characters although cursive is not connected in differences with many other non-Latin scripts such as Arabic (Fiaz Hussain 2013)

|    |    |   |    |   |   |   |   |    |
|----|----|---|----|---|---|---|---|----|
|    | ah | u | ee | a | e | i | o | ua |
| h  | ሀ  | ሁ | ሂ  | ሃ | ሄ | ህ | ሆ |    |
| h  | ሐ  | ሑ | ሒ  | ሓ | ሔ | ሕ | ሐ | ሒ  |
| s  | ሠ  | ሡ | ሢ  | ሣ | ሤ | ሥ | ሦ | ሧ  |
| s  | ሰ  | ሱ | ሲ  | ሳ | ሴ | ስ | ሰ | ሲ  |
| Q  | ቀ  | ቁ | ቂ  | ቃ | ቄ | ቅ | ቆ |    |
| b  | በ  | ቡ | ቢ  | ባ | ቤ | ብ | ቦ | ቧ  |
| t  | ተ  | ቱ | ቲ  | ታ | ቴ | ት | ቶ | ቷ  |
| h  | ኀ  | ኁ | ኂ  | ኃ | ኄ | ኅ | ኆ |    |
| n  | ነ  | ኑ | ኒ  | ና | ኔ | ነ | ኖ | ኗ  |
| a  | አ  | ሁ | ሂ  | ሃ | ሄ | ህ | ሐ | ሒ  |
| a  | ዐ  | ቡ | ሢ  | ሣ | ሤ | ሥ | ሦ |    |
| zj | ዠ  | ዡ | ዢ  | ዣ | ዤ | ዥ | ዦ | ዧ  |
| d  | ደ  | ዱ | ዲ  | ዳ | ዴ | ድ | ዶ | ዷ  |
| j  | ጀ  | ጁ | ጂ  | ጃ | ጄ | ጅ | ጆ | ጇ  |
| T  | ጠ  | ጡ | ጢ  | ጣ | ጤ | ጥ | ጦ | ጧ  |
| P  | ጸ  | ጹ | ጺ  | ጻ | ጼ | ጽ | ጾ | ጿ  |
| ts | ፀ  | ፁ | ፂ  | ፃ | ፄ | ፅ | ፆ |    |
| p  | ፐ  | ፑ | ፒ  | ፓ | ፔ | ፕ | ፖ | ፘ  |

|    |    |   |    |   |   |   |   |    |
|----|----|---|----|---|---|---|---|----|
|    | ah | u | ee | a | e | i | o | ua |
| /  | ሰ  | ሁ | ሂ  | ሃ | ሄ | ህ | ሐ | ሒ  |
| m  | መ  | ሙ | ሚ  | ማ | ሜ | ሚ | ሞ | ሚ  |
| r  | ረ  | ሩ | ሪ  | ራ | ሪ | ሪ | ሪ | ሪ  |
| sh | ሸ  | ሹ | ሺ  | ሻ | ሼ | ሸ | ሸ | ሺ  |
| v  | ቨ  | ቩ | ቪ  | ቫ | ቬ | ቨ | ቩ | ቪ  |
| ch | ቸ  | ቹ | ቺ  | ቻ | ቼ | ቸ | ቹ | ቺ  |
| ng | ኘ  | ኙ | ኚ  | ኝ | ኞ | ኘ | ኙ | ኚ  |
| k  | ከ  | ኩ | ኪ  | ካ | ኬ | ከ | ኩ |    |
| h  | ከ  | ኩ | ኪ  | ካ | ኬ | ከ | ኩ |    |
| w  | ወ  | ዉ | ዊ  | ዋ | ዌ | ወ | ወ |    |
| z  | ዘ  | ዙ | ዚ  | ዛ | ዜ | ዘ | ዘ | ዚ  |
| y  | የ  | ዩ | ዪ  | ያ | ዬ | ይ | ዩ | ዪ  |
| g  | ገ  | ገ | ገ  | ገ | ገ | ገ | ገ |    |
| ch | ጸ  | ጹ | ጺ  | ጻ | ጼ | ጽ | ጾ | ጿ  |
| ts | ፀ  | ፁ | ፂ  | ፃ | ፄ | ፅ | ፆ | ፇ  |
| f  | ፈ  | ፋ | ፈ  | ፋ | ፈ | ፋ | ፈ | ፈ  |

|   |    |    |    |    |
|---|----|----|----|----|
|   | uo | uu | ua | ue |
| k | ከ  | ኩ  | ኪ  | ካ  |
| h | ከ  | ኩ  | ኪ  | ካ  |
| g | ገ  | ገ  | ገ  | ገ  |
| h | ከ  | ኩ  | ኪ  | ካ  |

Figure 1 the Amharic alphabet (adopted from LexiLogos website)

### 2.1.1 Amharic language Consonant

The purpose of speech recognition to use Amharic consonants based on their manner of articulation and the presence or absence of voicing and place of articulation.

| Manner of articulation | Voicing                 | Place of Articulation |              |          |         |       |            |         |
|------------------------|-------------------------|-----------------------|--------------|----------|---------|-------|------------|---------|
|                        |                         | Bilabial              | Labiodentals | Alveolar | Palatal | Velar | Labiovelar | Glottal |
| Stops                  | Voiced                  | ብ[b]                  |              | ድ[d]     |         | ግ[g]  | ጓ[gw]      |         |
|                        | Voiceless               | ፕ[p]                  |              | ት[t]     |         | ከ[k]  | ኪ[kw]      | ዕ[ʔ]    |
|                        | Ejective or Glottalized | ጽ[pʼ]                 |              | ጥ[tʼ]    |         | ቅ[kʼ] | ቋ[kʼw]     |         |
| Fricatives             | Voiced                  |                       | ቨ[v]         | ዝ[z]     | ኸ[ʒ]    | ግ[g]  | ጓ[gw]      |         |
|                        | Voiceless               |                       | ፍ[f]         | ስ[s]     | ኸ[ʃ]    | ከ[k]  | ኪ[kw]      | ህ[h]    |
|                        | Ejective                |                       |              | ጽ[sʼ]    |         |       |            | ጓ[hw]   |
| Affricates             | Voiced                  |                       |              |          | ጅ[j]    |       |            |         |
|                        | Voiceless               |                       |              |          | ች[c]    |       |            |         |
|                        | Ejective or Glottalized |                       |              |          | ጭ[cʼ]   |       |            |         |
| Nasals                 |                         | ፖ[m]                  |              | ን[n]     | ኸ[ɲ]    |       |            |         |
| Liquids                |                         |                       |              | ል[l]     | ር[r]    |       |            |         |
| Glides                 |                         | ው[w]                  |              |          | ይ[y]    |       |            |         |

Table 1 Amharic consonant

### 2.1.2 Amharic language Vowels

Vowels are open sounds, made largely by shaping the vocal tract rather than by interfering with the flow of air stream and mostly described in terms of the position of the tongue as they are articulated. A vowel spoken with the body of the tongue relatively forward is classified as a front vowel one made with the body of the tongue relatively high is a high vowel. Vowels produced with the body the tongue neither high nor low are called mid vowels. Vowels produced with the tongue body front are called front vowels while those made with the tongue body back are called back vowels. Those vowels made with the tongue body neither front nor back are called central vowels. Vowels attended by lip rounding as in (u and o) are called rounded vowels while the other vowels are called unrounded vowels (Fantaye 2011).

|      | Front/Unrounded | Central/Unrounded | Back/Rounded |
|------|-----------------|-------------------|--------------|
| High | ፲[i]            | ፳[ɪ]              | ፵[u]         |
| Mid  | ፶[e]            | ፷[ɐ]              | ፸[o]         |
| Low  |                 | ፹[a]              |              |

Table 2 Vowel Articulations

## 2.2 Human speech production

### 2.2.1 Introduction

Speech is a human being's primary means of communication and it contains the meaning of information from a speaker to a listener, individual information representing the speaker's identity, gender, and also sometimes emotions. For a complete explanation of speech production, the properties of both articulators, which produce the sound, and auditory organs, which perceive the sound, should be, involved (Anderson 2020).

Figure 2 shows the speech chain. A spoken message travels from the speaker to the listener (Rodman 2013). The speech chain consists of three events: the production of speech sounds through the vocal apparatus of the speaker, the traveling of the acoustic signal through the air, and, finally, its reception by the ear of the listener. The brain of the speaker controls the production of speech sounds and the brain of the listener analyzes the signal and converts it into meaning.

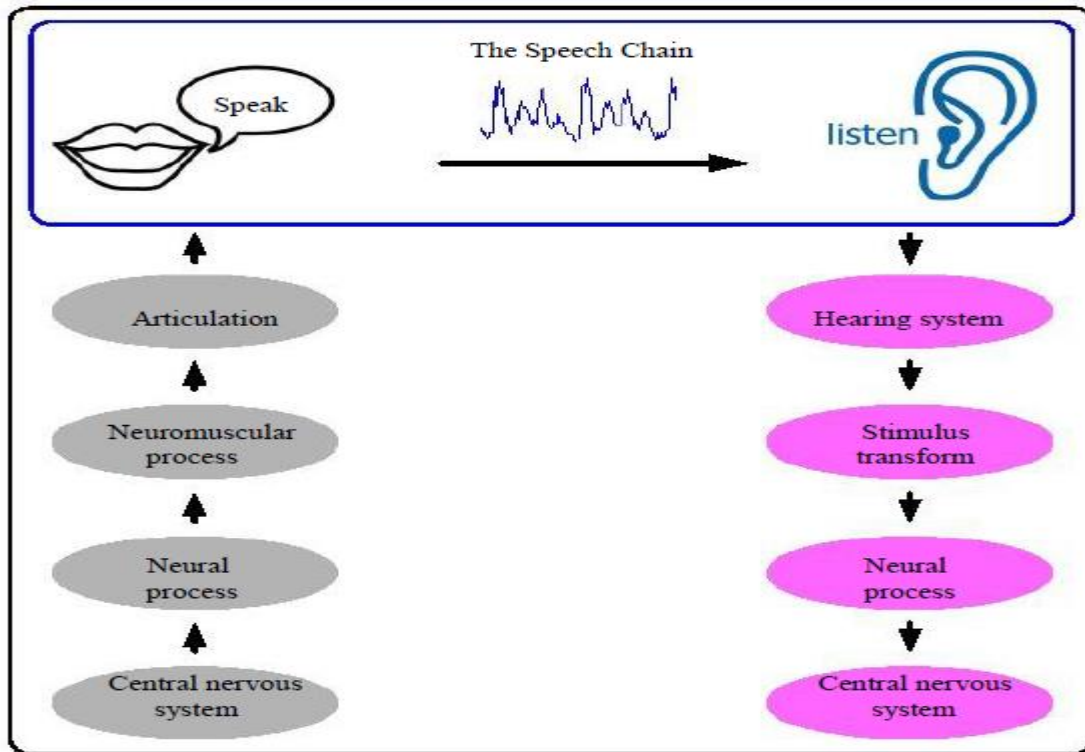


Figure 2 Speech chain system (adopted from Peter Denes, 2012)

To get the required features for the speaker recognition task, it is essential to understand the mechanism of speech production, the properties of the human speech production model, and the articulators that have speaker-dependent characters.

There are two main sources of speaker-specific characteristics of speech: physical traits and learned traits (Hansen 2015).

**Physical traits** refer to the anatomical deference inherent in the vocal tract.

**Learned traits** include speaking rate, timing patterns, pitch patterns, dialect, word/phrase usage, etc.

Although the high-level cues (learned traits) are more robust and are not much affected by noise and channel mismatch, I limit the scope in the low-level cues (physical traits) because they are easy to be automatically extracted and suitable for our purpose.

### 2.2.2 Speech Production Mechanism

The speech production procedure begins with a thought which shows the first communication message. Following the instructions of the spoken language and grammatical structure, words and phrases are selected and ordered. After the thought constructs into language, the brain sends commands utilizing motor nerves to the vocal muscles, which move the vocal organs accordingly to produce sound (Flanagan 2016). Speech production can be divided into three main mechanisms: excitation production, vocal tract articulation, and lip's and/or nostril's radiation.

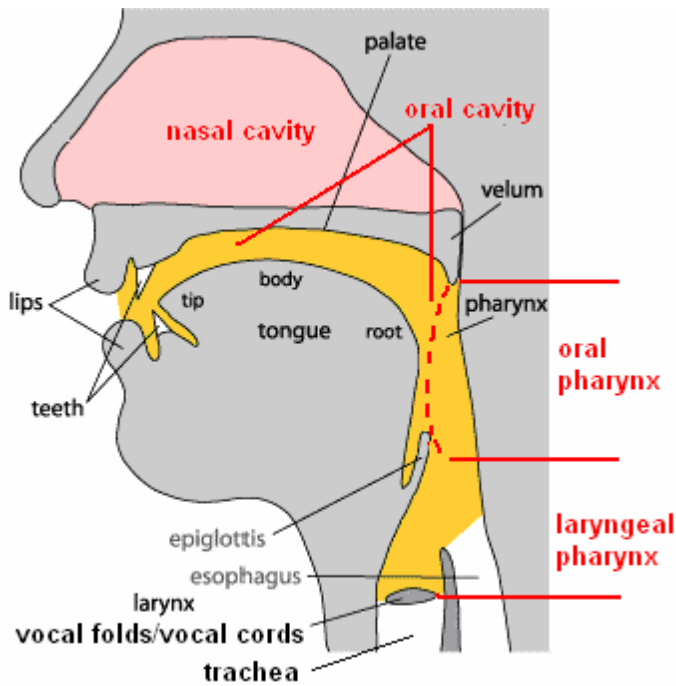


Figure 3 Anatomical structure of the human vocal system (adopted from J. L. Flanagan, 2016).

#### 2.2.2.1 Excitation Source Production

Excitation powers the speech production procedure. It is produced by the airflow from the lungs and then carried by the trachea through the vocal folds. During inspiration, the air is filled into the lungs, and during expiration, the energy will be spontaneously released. The trachea conveys the resulting air stream to the larynx.

The larynx refers to an energy supplier to help inputs to the vocal tract, and the volume of air governs the amplitude of the sound. The vocal folds at the base of the larynx and the glottis triangular-shaped space between the vocal folds are the critical parts from a speech production point of view.

They separate the trachea from the base of the vocal tract. The types of sounds are determined by the action of vocal folds, and we call it excitation. Normally excitations are characterized as phonation, frication, and plosive.

Speech produced by phonated excitation is called the voice, produced by the cooperation between phonation and frication is called mixed voiced, and produced by frication is called unvoiced (L. Siegel 2009).

Voiced speech (periodic) is generated by modulating the air stream from the lungs, and the generation is performed by periodically open and close vocal folds. The oscillation frequency of vocal folds is called the fundamental frequency ( $F_0$ ), or equivalently, the pitch of the resulting sound and it depends on the physical characters of vocal folds. Hence fundamental frequency is an important physical distinguishing factor, which has been found effective for automatic speech and speaker recognition. Vowels and nasal consonants belong to voiced speech.

Mixed voiced speech (impulsive) is produced by the phonation plus frication and closure at some point in the vocal tract, followed by a release of air. Actually, unlike the phonation that is placed in vocal folds (the vibration of vocal folds), the place of frication is inside the vocal tract.

Unvoiced speech (noisy) is produced by a constriction of the vocal tract narrow enough to cause turbulent airflow, which results in noise or breathy voice.

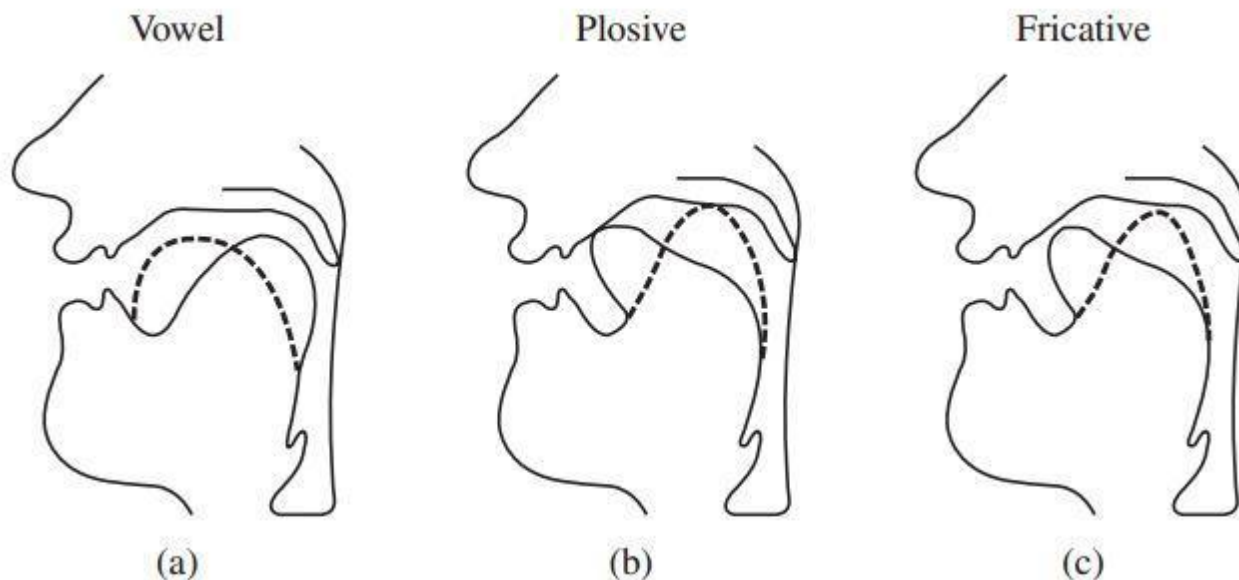


Figure 4 Illustration of changing vocal tract shapes (adopted from Ittansa K., 2014)

(a) Vowels (having a periodic source), (b) plosives (having an impulsive source), and (c) fricatives (having a noise source)

#### 2.2.2.2 Vocal Tract Articulation

The vocal tract is generally considered as the speech production organ above the vocal folds, which is formerly known as vocal cords, and its shape is another important physical distinguishing factor.

It includes both the excitation organs and vocal tract organs. Lungs, trachea, and vocal folds are regarded as organs responsible for excitation production. The combination nasal cavity and oral cavity, in the picture, is referred to as the vocal tract.

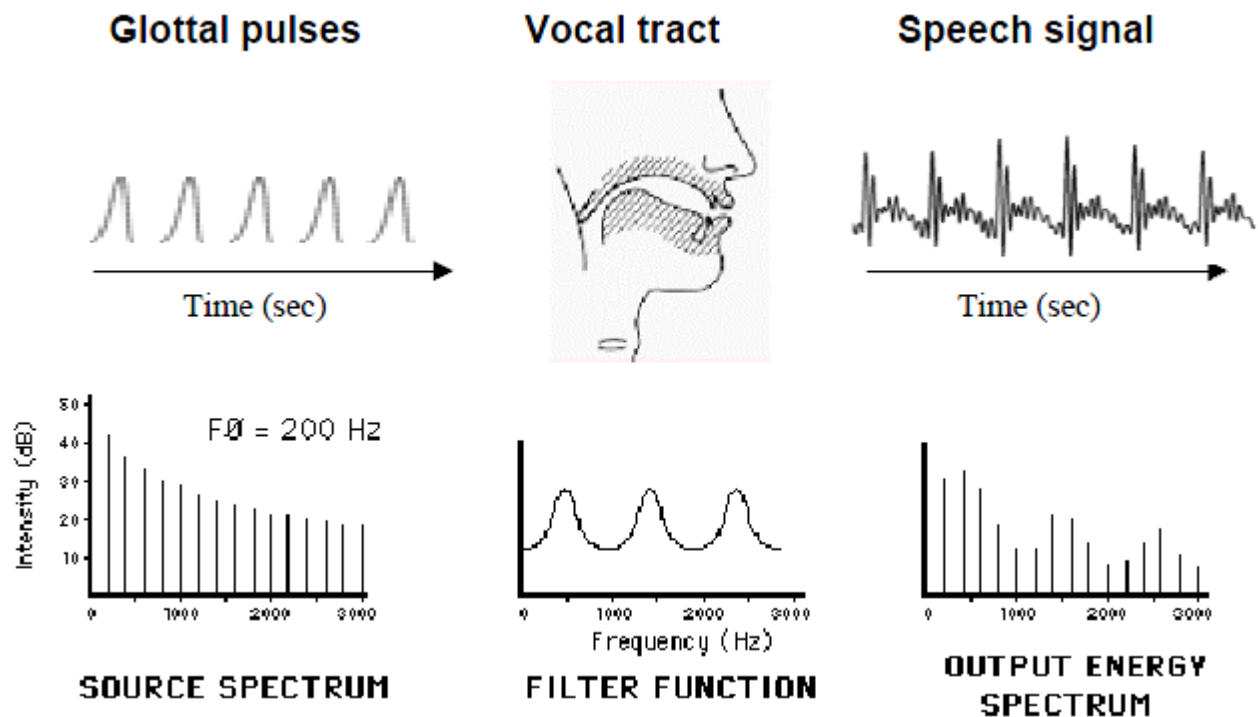


Figure 5 Vocal tract articulations (adopted from J. L. Flanagan, 2010)

While the acoustic wave produced by excitations is passing through the vocal tract, depending on the shape of the vocal tract, the wave will be altered in a certain way and interferences will generate resonances. The resonances of the vocal tract are called formants. The figure 6 shows a spectral envelope and its formants. Their location largely determines the speech sound which is heard (Joe Wolfe 2008).

The vocal tract works as a filter to shape the excitation sources. The uniqueness of the speaker's voice not only depends on the physical features of the vocal tract but also on the speaker's mental ability to control the muscles of the organs in the vocal tract. It is not easy for the speaker to change the physical features intentionally. However, these physical features are possibly being changed with aging.

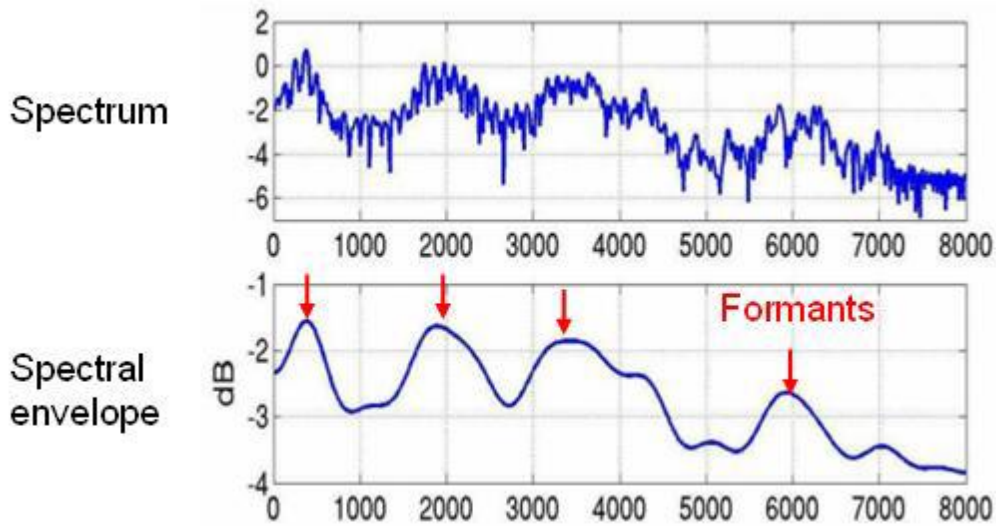


Figure 6 Spectral envelope and its formants (adopted from Maëva Garnier, 2013)

### 2.2.3 Speech production model

As mentioned before, speech production is normally divided into three principal components: excitation production, vocal tract articulation, and lips' and/or nostrils' radiation. As we separate the speech production process into three individual parts, which have no coupling between each other, we assume that these three components are linear, separate, and planar propagation (Joel Trussell 2014). Furthermore, let's think about speech production in terms of an acoustic filtering operation. Consequently, we could construct a simple linear model, discrete-time filter model, for speech production, which consists of excitation production part, vocal tract filter part, and radiation part separately, as shown in the figure 7.

The excitation part corresponds to the vibration of the vocal cords (glottis) causing voiced sounds, or to a constriction of the vocal tract causing a turbulent air-flow and thus causing the noise-like unvoiced excitation. In terms of the LTI system model, the excitation is the input function  $e(n)$ , the vocal tract acts as the system impulse response  $h(n)$ , lip radiation filter impulse response  $r(n)$  and the speech is the output  $s(n)$ . This system is not time-invariant, but for a short time interval of 10-30msec, it can be viewed as a “piecewise” LTI system.

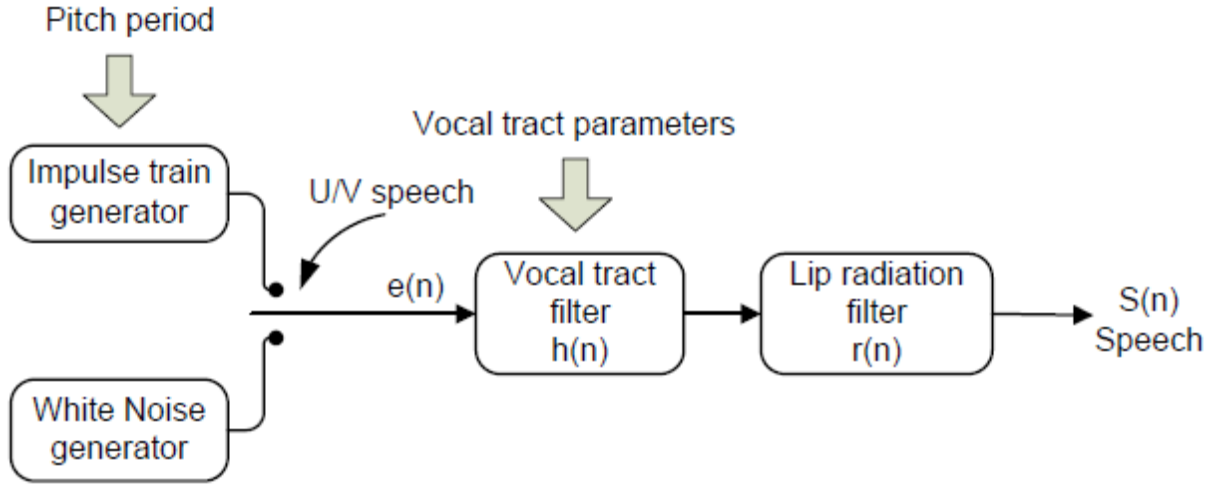


Figure 7 Discrete-time speech production model

Thus, for a specific segment of speech, by using this model a speech can be computed as the product of three respective (Fourier) transfer functions:

$$S(\omega) = E(\omega) H(\omega) R(\omega) \quad (2.1)$$

Where:

$E(\omega)$ : Excitation spectrum

$H(\omega)$ : Vocal tract filter

$R(\omega)$ : Lip radiation filter

In time-domain, relation 2.1 will be presented as a convolution, combination of excitation sequence, the vocal system impulse response, and the speech radiation impulse response:

$$S(n) = e(n) \oplus h(n) \oplus r(n) \quad (2.2)$$

As we know, the magnitude spectrum can be exactly modeled with stable poles, and the phase characteristics can be modeled with zeros. Hence estimation for the true speech production model shown in Figure 8, an all-pole model is valid and useful. LP model (all-pole model) has the correct magnitude spectrum, but the minimum-phase characteristic is compared with the true speech model. Figure 12 shows the estimated model using LP analysis, which is also called the source-filter model.

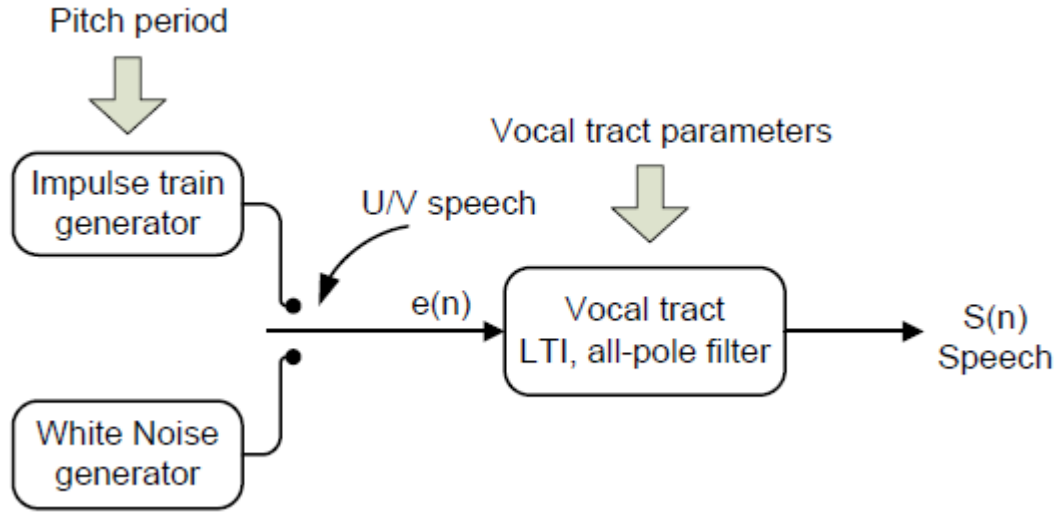


Figure 8 Estimated speech production model.

The transfer function of an all-pole filter is represented by:

$$H(\omega) = \frac{1}{1 - \sum_{k=1}^p a_k(j\omega) - k} \quad (2.3)$$

Where

$p$  is the number of poles;

$a_k$  is the Linear Prediction Coefficients.

As a result of this estimation, the speech signal then can be represented as the product of two transfer functions:

$$S(\omega) = E(\omega) H(\omega) \quad (2.4)$$

Where  $E(\omega)$  is the excitation spectrum and  $H(\omega)$  is represented by (2.3). Consequently, in the time domain, the speech signal is as follows:

$$S(n) = e(n) \oplus h(n) \quad (2.5)$$

#### 2.2.4 Speaker-distinctive Characteristics

There are a variety of voice qualities that characterize a speaker, fundamental frequency, harmonic structure, and intensity.

Speech is a difficult signal produced as a result of several conversions happening at numerous different levels: semantic, linguistic, articulatory, and acoustic (Finch 2017).

Differences in these transformations look as differences in the audio properties of the speech signal.

Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals.

In speaker recognition, all these differences can be used to distinguish between speakers. Human perception of speech sounds is a complex process, to measure the speech waveform into representative features, the consideration of the human auditory properties is essential.

#### 2.2.4.1 Fundamental frequency

The direct result of vocal cord vibration is the fundamental tone of the voice, which determines its pitch. In physical terms, the frequency of vibration as the foremost vocal attribute corresponds to the number of air breaths per second, counted as cycles per second (cps or Hertz).

This frequency is controlled by a combination of effects, both stable and variable factors. The stable determinants of the individual voice range depend on the laryngeal sizes as related to gender, age, and body type.

The smaller a larynx, the higher its pitch range. In this individually fixed range, variables that influence the pitch of a given phonation include the tension of the cord, the force of glottal closure indicated by the glottal resistance, and expiratory air pressure.

During the speech, we repeatedly modify the tension and length of the vocal cords, and the air pressure from the lungs, till we get the desired frequency.

The range of vocal cord frequencies used in normal speech covers from about 60 to 350cps (Jakobson 2014).

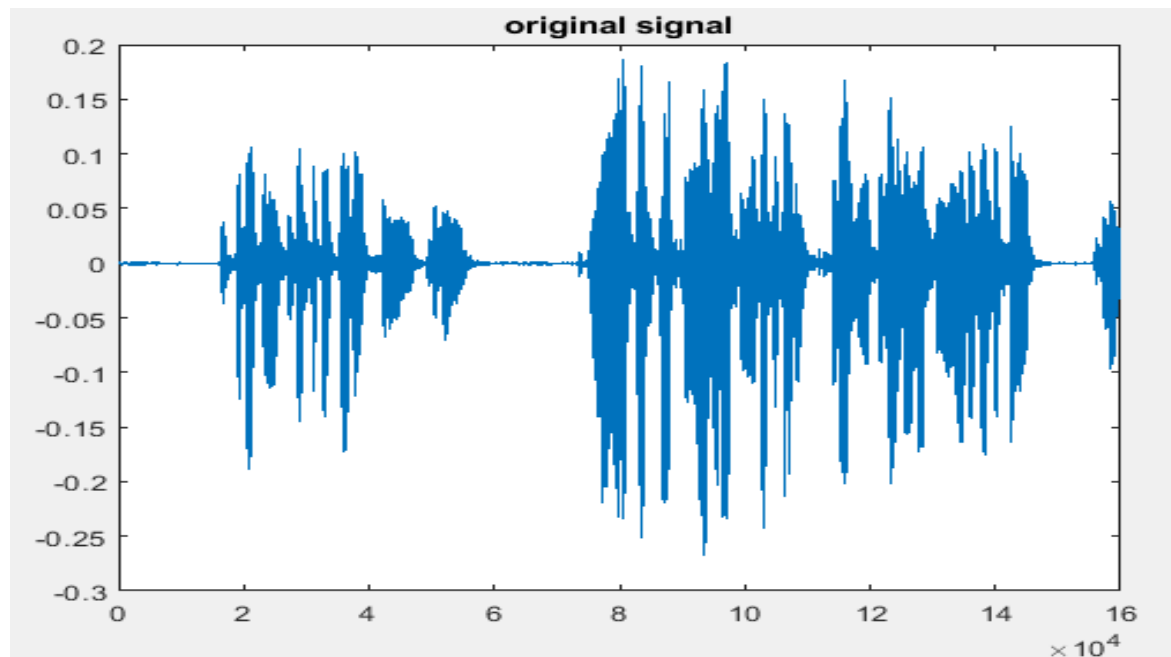
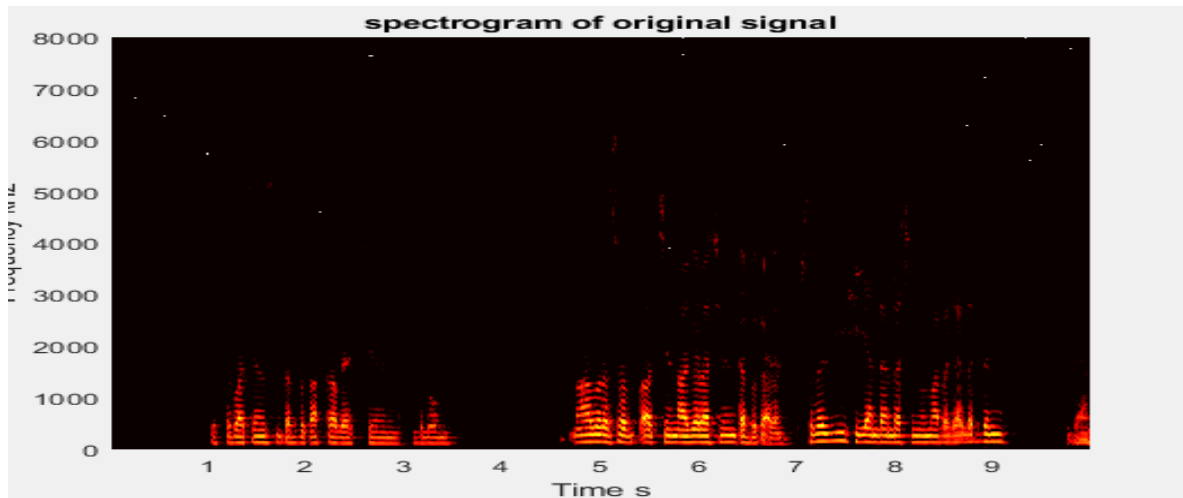
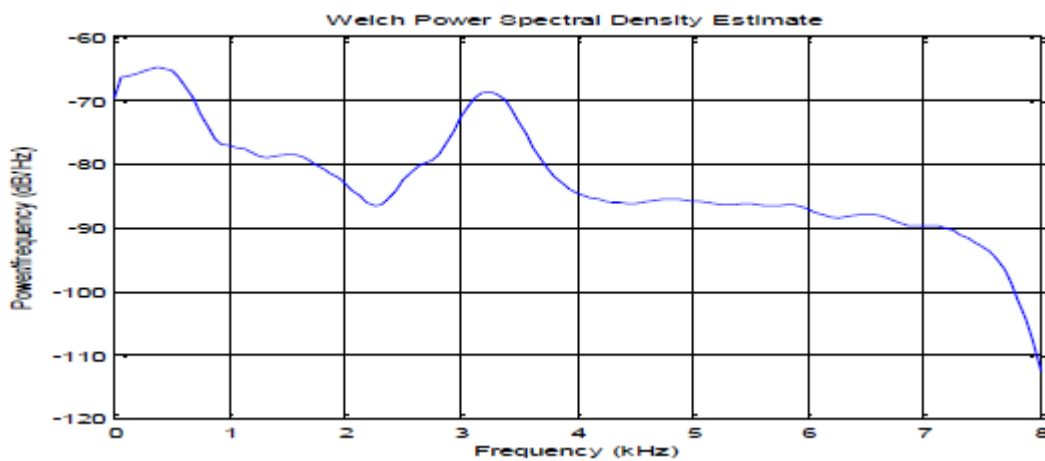


Figure 9 Original signal of speech wave form



(a) Speech waveform and its spectrogram



(b) Speech formant frequencies

Figure 10 Waveform and spectrograms of a speech signal.

#### 2.2.4.2 Harmonic structure

The second quality of vocal sound, harmonic structure, depends on the waveform produced by the vibrating vocal cords.

As a musical instrument, the human voice is not a pure tone rather, it is composed of a fundamental tone (or frequency of vibration) and a sequence of higher frequencies called upper harmonics.

As long as the harmonics are exact multiples of the fundamental, the vocal will sound clear and pleasant.

If non-harmonic components are added, increasing degrees of roughness, harshness, or hoarseness will be perceived about the intensity of the noise components in the frequency spectrum.

The main harmonic structure is radiated into the vocal tract then. The shaping of the vocal tract determines the modulation of the voice through resonance and damping.

As a general rule and wide vocal tract enhances the lower harmonics, producing a full, dark, and resonant voice. Conversely, shortening and narrowing of the vocal tract lead to higher resonances with lightening of the voice and the perceptual attributes ranging from high-pitched and noisy to constricted and guttural.

#### **2.2.4.3 Intensity**

Vocal intensity, the third major vocal attribute, depends primarily on the amplitude of vocal cord vibrations and thus on the pressure of the subglottic airstream.

Another component of vocal intensity is the radiating efficiency of the sound generator and its covered resonator. The larynx has been compared to the physical shape of a horn. This construction is most efficient in acoustical practice, as seen in the shape of wind instruments, car horns, sirens, loudspeakers, etc.

A well-shaped, wide, and flexible vocal tract enhances the projective potential of the voice. Conversely, a morphologically narrow, pathologically constricted, or emotionally tightened throat produces a quiet, constricted sound with poor carrying power.

Generally, the resonant effects of the vocal tract on the harmonic structure are reflected by the spectral envelope of a speech segment. Individual formants usually present themselves with a peak in the spectrum. In the figure 9 (b), the formants are reviewed.

### **2.3 Speaker recognition system**

Speaker recognition is a system used to automatically identify a speaker from a recording of their voice or speech utterance (Mokgonyane 2019).

Speaker recognition has two fundamental tasks, namely speaker verification and speaker identification (Rania Chakroun 2019).

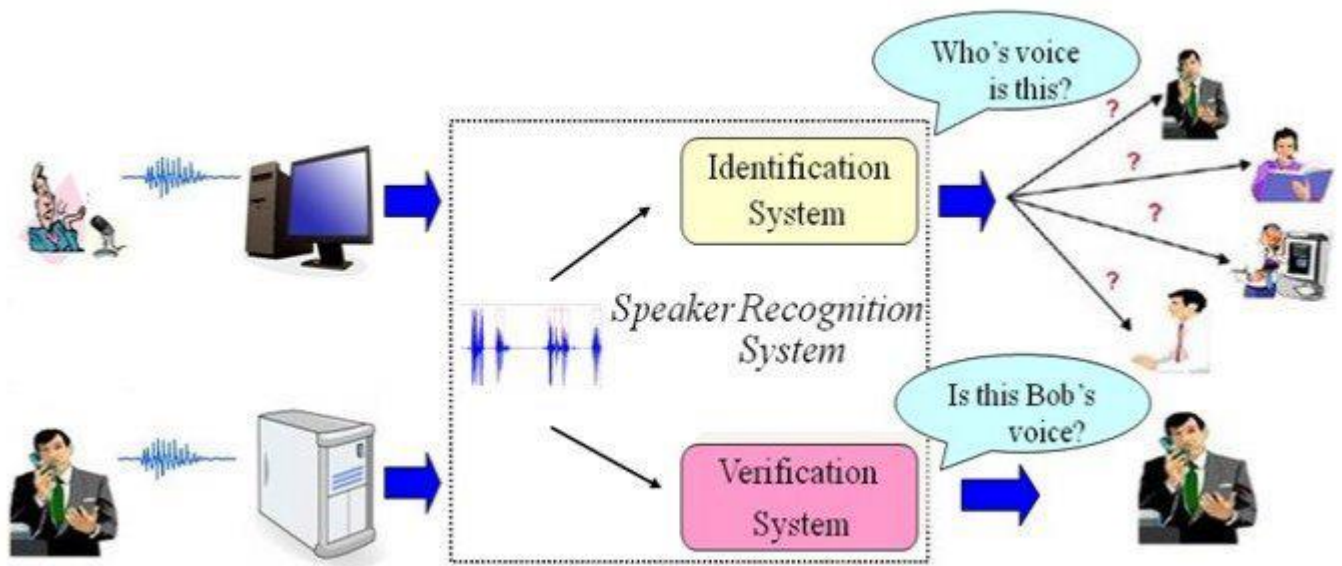


Figure 11: Speaker Recognition fundamental tasks (adopted from Tumisho B., 2018)

### 2.3.1 Speaker identification and speaker verification

As already stated previously, speaker recognition systems can work in two different styles, speaker verification or speaker identification.

Speaker verification is the way of deciding whether accommodating or eliminating the identity claim of a speaker depends on a given voice sample. It could be useful in areas like cell phone or bank account authentication. In a speaker verification system, an identity claim (for example a PIN code) is first input to the system, and the unknown speaker's voice sample is matched against the claimed speaker's voice template. During this, the system performs a 1:1 comparison. If the matching degree between the speech utterances and the template surpasses a predefined decision threshold, then the system produces a single decision either the speaker is accepted or rejected (Praveen P S 2017).

Figure 11 illustrates the major structures of speaker verification systems.

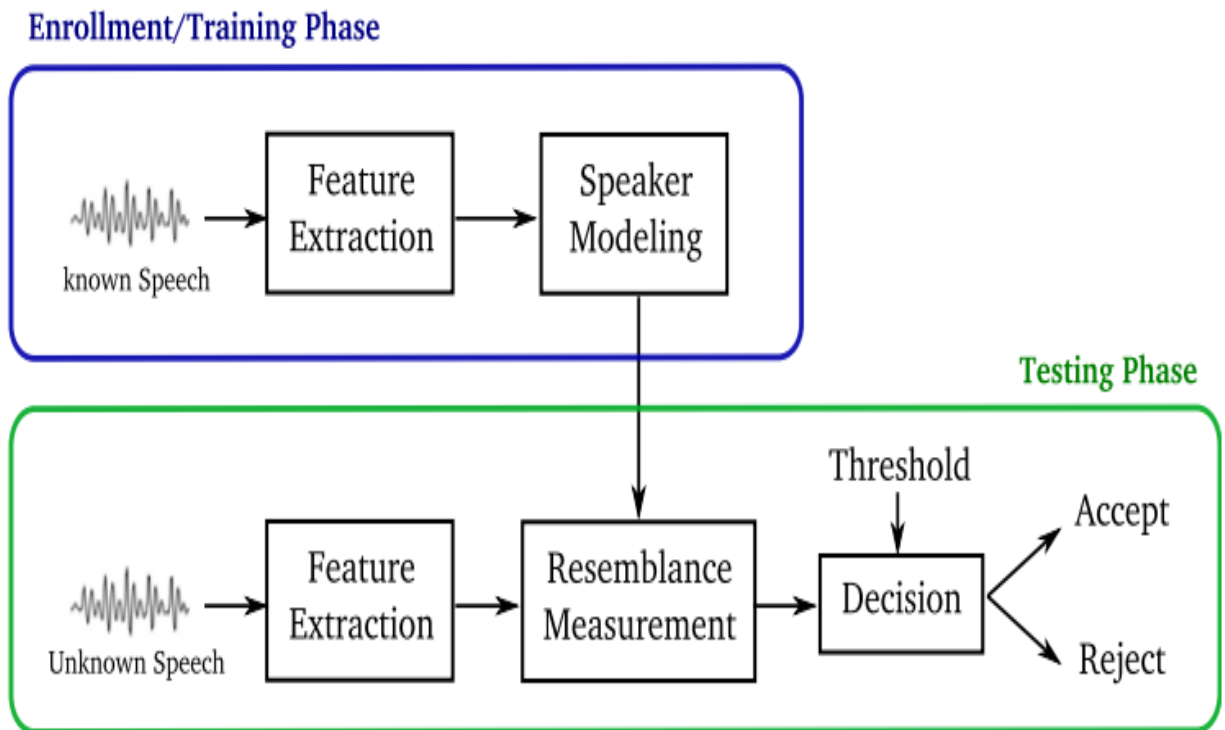


Figure 12 Speaker Identification system (adopted from Tom Bäckström, 2015)

On the other hand, in a speaker identification system, the user will not provide his identity to the system. Instead, the user will input his speech and the system will decide which speaker model better matches the speech input. In this case, the system has to perform  $N$  comparisons,  $N$  being the number of speakers in the system database. Each comparison will produce a likelihood score so the system will choose the identity attached to the most likely speaker model. This means that the kind of decision will be "Speaker  $I$ " with  $i=1\dots N$ . It is clear that the speaker identification problem is more complex than speaker verification and therefore, the results achieved in speaker identification systems will be poorer than in speaker verification (Praveen P S 2017). In this thesis, Speaker Identification was selected.

Figure 12 illustrates the major structures of speaker identification systems.

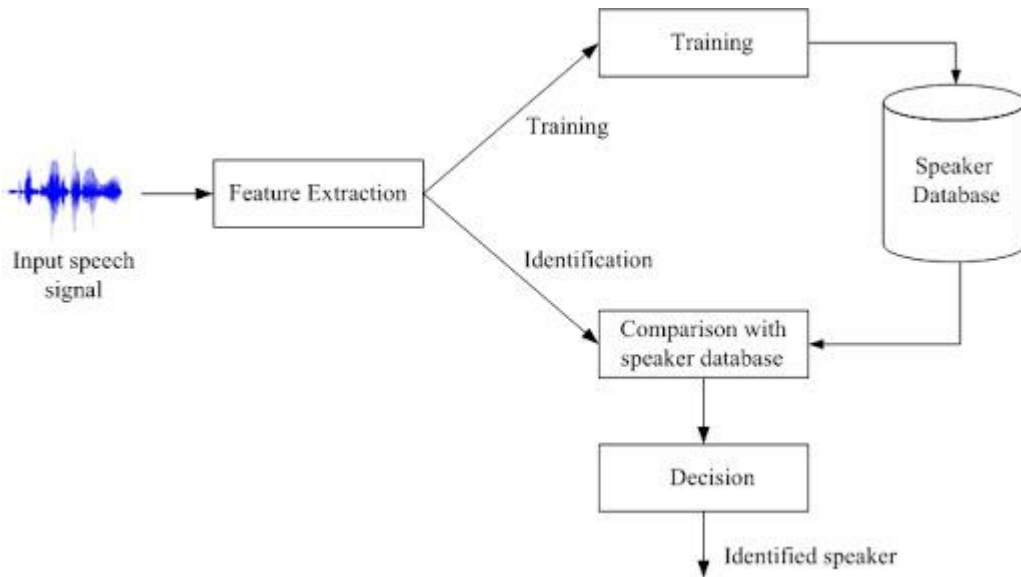


Figure 13 speaker identification system (adopted from Ronak Bajaj, 2018)

### 2.3.2 Text-independent and text-dependent

The second possible classification method for ASRs can be based upon the text uttered by the speaker during the identification process. In the text-dependent domain, a given sentence or word will be always used for training and recognition. In this case, a two leveled security system is obtained. Firstly, the voice has to be produced by an authorized user, and secondly, the user has to provide the proper password (Tolba 2018).

Another kind of system can be included in the group of text-dependent recognizers. These are text-driven systems. In this case, the system will ask the user to pronounce a given code (a sequence of numbers) to prevent a phishing attack. In this case, an attacker should have voice registries of the authorized user including all the possible combinations the system may ask for. The larger the number of keywords used, the more robust system can be built (Praveen P S 2017).

On the other hand, text-independent systems do not care about the message contained in the speech input. They only attend to speaker-dependent attributes of speech and do not rely on the sequence of words pronounced by the speaker.

The recognition and training will be based on any utterance of speech produced by the speakers. The text-independent case is much more flexible since it is not constrained to any kind of password. The applications of this method are not only in biometric security but in any field which requires the answer to the question “who said this?”

This thesis is based on this kind of system because of its flexibility. Text-independent security systems might look unsafe because of the absence of a password. However, this is not true since speech recognition can be combined with speaker recognition to introduce the password level. In this case, speaker recognition will be separated from speech recognition in opposite to the case of text-

dependent recognition (I. Y. Kelbesa 2014). In real-life situations, text-independent recognition systems are more in demand.

Both systems can be defeated if the recorded voice of a registered speaker is played. A possible solution to this condition is to go for text-prompted systems in which a new text is prompted every time one operates a system. So using this technique recorded voice of an authorized user can also be rejected making the systems perform in real-time.

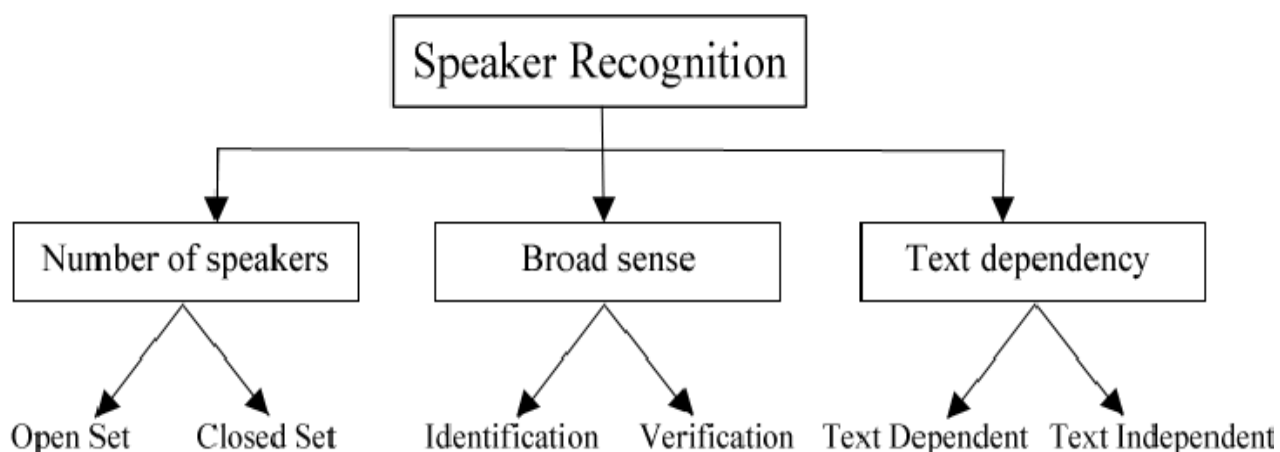


Figure 14 speaker recognition taxonomy (adopted from Tshephisho S., 2019)

### 2.3.3 Open-set and closed-set

The last method for classifying speaker recognition systems distinguishes between the possibility any user in the world can try to enter the system or only a set of them will try to do it.

A closed-set recognizer will consider that the user who is trying to access the system belongs to a group of known users. In this case, if a speaker identification system is being used, the decision taken by the system will just be the most likely speaker to the speech input.

However, an open-set system will consider the possibility that the user who is attempting to enter the system can be unknown. This means that there is not a model associated with this user and therefore the decision taken by the system is that the user is unknown or an impostor.

There are three main methods to deal with this problem. The first one is including a model for unknown speakers obtained from a large voice database including several speakers. The second is setting a likelihood threshold so that if any of the speaker models' score overpasses this threshold, the decision will be “unknown user”. The last one is a proper combination of both methods.

It is clear that open-set systems are more complex and match better the conditions in real life where a listener does not have to be familiar with all the voices he hears every day. But in this thesis, we considered closed-set text-independent (Prasanna 2014).

## 2.4 speakers modeling

The objective of the modeling technique is to produce speaker models using speaker-specific feature vectors. Such models will have enhanced speaker-specific information at a reduced data rate. This is achieved by exploiting the working principles of the modeling techniques (Prasanna 2014). Depending upon type of speech, ease of training, computational and storage requirements, and expected performance the selection of modeling technique is made.

Speaker modeling has two broad categories that are generative and discriminative types.

### 2.4.1 Generative model

Generative models estimate feature distribution within each speaker. They only require training data samples from the target speaker and form a statistical/non-statistical model that describes the target speaker's feature distribution. It includes Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Vector Quantization (VQ) (Pullella 2011) generative model also have sub-categories i.e. parametric, non-parametric.

#### 2.4.1.1 Parametric model

Parametric models are those which assume structure characterized by specific parameters. It consists of GMM, HMM models. The form is limited and less data is needed to specify the density. The advantages of these models are that efficient use of data is there, it is possible to model and understand changes in data through changes in parameters, and statistical summaries can be used rather than complete data (Suma Paulose 2017). The disadvantage is that the structure is restrictive. GMM is the general parametric model that represents each speaker by a pdf (probability density function) governing the distribution of his/her feature vectors. It can form smooth densities of irregular shapes. In GMM for reliable density estimation, the number of required training samples grows exponentially with the number of features. This is also called a Curse of dimensionality (S G Bagul 2013). GMM can also use temporal information i.e. state transition probabilities; in that case, it results in continuous HMM.

**In statistical models**, the pattern matching is probabilistic and results in a measure of likelihood or conditional probability of the observation given the specific model. Classification is based on probabilities /likelihoods rather than distances to average features. Each speaker is modeled as a probabilistic source with a fixed pdf. In training, the parameters of pdf are estimated. There are two models: the HMM (Hidden Markov Model) and GMM (Gaussian Mixture Model) (Nilu Singh 2018). HMM is generally used for text-dependent tasks while GMM is used for Text-independent tasks.

These can be trained using the Forward-backward algorithm and the Expectation-Maximization (EM) algorithm. For text-independent speaker recognition where there is no prior knowledge of what will be said the most successful model is GMM.

#### **2.4.1.2 Non-parametric model**

In non-parametric models, minimal assumptions regarding pdf are made. The most popular methods are the Nearest Neighbor (NN), Vector Quantization (VQ), and Dynamic Time Warping (DTW) (Yadav 2015). In VQ centroid of each cluster of features is computed. The collection of centroids is called a codebook. The codebook for each speaker is generated using Linde Buzo Gray (LBG) algorithm. The identified speaker is the one with minimum distortion. VQ is identical to the NN model except that distances to nearest data representations are measured. Therefore it reduces computation and memory demands as compared to NN.

The advantages of VQ have reduced storage for spectral analysis information. Due to the discrete representation of speech sounds, we can associate a phonetic label with each codebook. Disadvantages of VQ include inherent spectral distortion since finite numbers of codebook vectors are therefore there is a certain level of quantization error. As the size of the codebook increases quantization error decreases (Juang 2017).

**In template** (time-ordered set of features) models pattern matching is deterministic. The test data is assumed to be an imperfect replica of the training data and distance is measured and the minimum distance template is chosen. Many different distance measures are there like Euclidean Itakura, Mahalanobis (J. P. Campbell Jr 2016). This method does not require any model training. DTW is mainly used for text-dependent systems. Template models can be further categorized as time-dependent (e.g. DTW) or time-independent (e.g. VQ). In time-independent models, all temporal variations are ignored. A time-dependent model is more complicated as it must accommodate the variability of the human speaking rate. In VQ larger codebook size is required for better efficiency and also VQ codebook has to be updated from time to time to compensate for intra-speaker variations.

#### **2.4.2 Discriminative models**

Discriminative model models the boundary between speakers. These require training data for both target and non-target speakers and derive an optimal separation between the different speakers. It includes models like Artificial Neural Networks (ANNs) and Supports Vector Machines (SVMs). These modeling techniques can also be termed as soft computing modeling. The decision function

between speakers is trained rather than forming individual speaker models (Kráľ 2010). The advantages of these models include flexible architectures and discrimination training power.

Generally, most current speaker identification systems are either based on Bayes decision (like Gaussian mixture model) or neural network classifier (like radial basis function neural network). The limitation of them is that they need cross-validation to avoid over-training on a limited amount of training data (Fenglei 2017).

This thesis presents a different approach to speaker identification using support vector machines (SVMs). SVM is a learning method based on the statistical learning theory (SLT). SLT is a special theory on machine learning in the case of limited learning data. SLT is based on a set of theories to form a unit frame for limited sample learning problems. It can contain many current learning methods and solve many problems, such as the construction selection of neural networks, local minimum, and overfitting. SVM shows many good properties over other methods in solving limited sample problems and non-linear high dimension pattern recognition problems. SLT and SVM are becoming the new study hot area after the neural network, and will further develop the machine learning theory and technology (W. M. Campbell 2006). SVMs have confirmed to be a powerful technique for pattern classification. SVMs map inputs into a high-dimensional space and then separate classes with a hyperplane.

## **2.5 Feature Extraction**

Feature extraction starts from the first set of measured data and forms derived values (features) proposed to be useful and non-redundant, help the subsequent learning and generality steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction (Philomina Simon 2020).

When the input data to an algorithm is too large to be managed and it is assumed to be redundant (e.g. the same dimension in both feet and meters, or the repetitiveness of images presented as pixels), then it can be changed into a reduced set of features called feature vector.

Determining a subset of the initial features is called feature selection (Mohana 2018). The selected features are normal to contain the relevant information from the input data so that the wanted task can be performed by using this reduced representation instead of the complete original data.

There are several feature extraction techniques, selecting accurate feature extraction, and classification methods appropriate for the data set is very important. Since the performance of the speaker identification system depends on the performance of the selected algorithm feature extraction and its computation time. To extract spectral features from the voice-signal the most

prominent classical state-of-the-art features extraction algorithms are used. These are Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficient (LPCC), and Mel Frequency Cepstral Coefficients (MFCC). However, in noisy environments, all of the classical algorithms are not robust enough and it provides less accuracy (Jyoti B. Ramgire 2016). So, in this thesis, I used deep learning speech feature extraction techniques.

### **2.5.1 Deep Learning Speech Feature Extraction Techniques**

Deep learning is presently an effective research area in machine learning techniques and pattern classification association. This has achieved big success in the areas of application namely computer vision, speech recognition, and NLP (G. E. Hinton 2006) . This thesis gives the impact of feature extraction that is used in a deep learning technique such as Convolutional Neural Network (CNN).

#### **2.5.1.1 convolution neural network (CNN)**

CNN is a category of a deep learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of the animal visual cortex (Chen MC 2018) and designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. CNN is a mathematical concept that is naturally composed of convolution, pooling, and fully connected layers or building blocks.

Convolution and pooling layers perform feature extraction, while a fully connected layer, performed classification.

A convolution layer plays an important role in CNN, which is collected of a load of mathematical operations, such as convolution, a specialized type of linear operation.

In numerical images, pixel values are stored in a two-dimensional (2D) grid, that is an array of numbers, and a small grid of parameters called the kernel, optimizable feature extractor, is applied at each image position, which makes CNN's highly efficient for image processing, since a feature may occur everywhere in the data.

The procedure of normalizing parameters such as kernels is called training, which is performed to minimize the difference between outputs and ground truth labels through an optimization algorithm called backpropagation and gradient descent, among others.

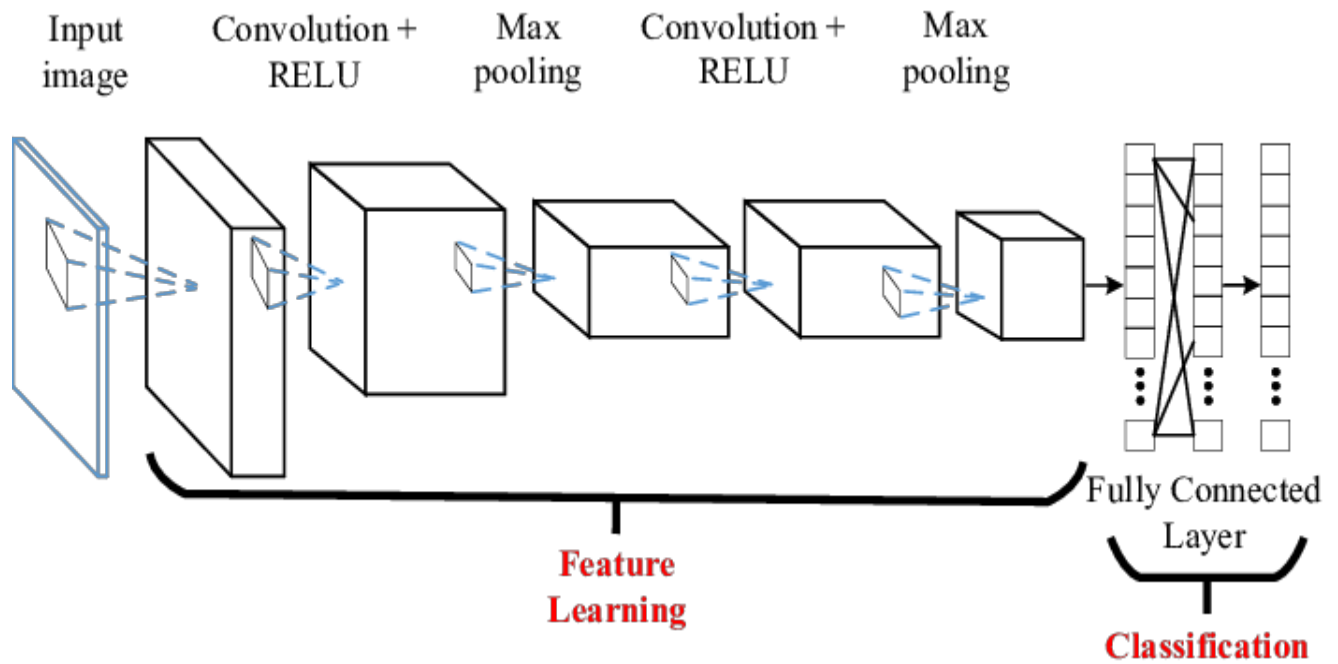


Figure 15 general frameworks of CNN (Adapted from Henning Müller, 2020).

#### 2.5.1.2 Building blocks of CNN architecture

The CNN architecture contains several building blocks as mentioned above such as convolution layers, pooling layers, and fully connected layers.

A typical architecture contains repetitions of a mass of several convolution layers and a pooling layer, followed by one or more fully connected layers.

The stage where input data are changed into output through these layers is called forward propagation. Although convolution and pooling operations defined in this section are for 2D-CNN, similar operations can also be performed for three-dimensional (3D)-CNN (LeCun Y 2015).

##### Image input layer

The input layer is the input of the whole CNN and it is usually a multidimensional array of data. The Input data can be a 2-D or 3-D image pixels or their transformation, patterns, time series, or audio signals. This layer needs the image input size parameters such as width, height, and the number of color channels. For instance, in a grayscale image, the number of channels is represented by 1, and for colored images represented by 3 (Zahir H. 2020).

##### Convolution layer

This layer is the main component of CNN. CNN works on a layer-to-layer operation named convolution. In the convolutional layers the activations from the input layers are convolved with a set of small parameterized filters, frequently of size  $3 \times 3$ , collected in a tensor  $W^{(j,i)}$ , where  $j$  is the filter number and  $i$  is the layer number.

By having each filter share the same weights across the whole input domain, i.e. translational equivariance at each layer, one achieves a drastic reduction in the number of weights that need to be learned. The motivation for this weight-sharing is featured appearing in one part of the image likely also appear in other parts. If we have a filter capable of detecting horizontal lines, say, then it can be used to detect them wherever they appear. Applying all the convolutional filters at all locations of the input to a convolutional layer produces a tensor of feature maps.

The two significant hyperparameters that express the convolution process are the size and number of kernels. So in this work, we use  $3 \times 3$ , but sometimes the other researchers also used  $5 \times 5$  or  $7 \times 7$ .

There is no pre-determined rule about the number of convolutional layers that are used to integrate a CNN model. This is determined based on us' amount of data sets.

However, from the literature, two to four layers have been implemented in different architectures (Madhuri Yadav 2018). Convolution in a two-dimensional input image  $I$  in the first table and two-dimensional kernels  $K$  in the second table is given by the following equation:

$$F(x,y) = (x*k)(x,y) = \sum_m \sum_n x(m,n) k(x+m,y+n) \quad 2.6$$

Where  $x$  and  $y$  are an image  $I$  coordinates whereas  $m$  and  $n$  are kernel  $K$  coordinates

The output of convolution operation after each operation is  $W \text{ output} \times H \text{ output} \times D \text{ output}$  where:

$$W \text{ output} = ((W \text{ input} - F + 2 P) / S) + 1$$

$$H \text{ output} = ((H \text{ input} - F + 2 P) / S) + 1$$

$$D \text{ output} = K \quad 2.7$$

Where  $F$  is the filter size,  $P$  is the number of zero padding,  $S$  is the number of stride size, and  $K$  is the number of filters applied.

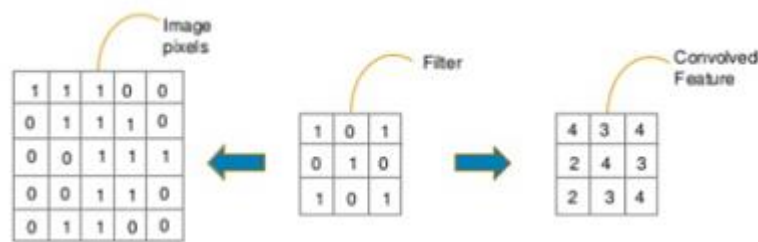


Figure 16 an example of a convolution operation with a kernel size of  $3 \times 3$

Figure16 shows a convolution operation with a kernel size of  $3 \times 3$ , no padding, and a stride of 1. Stride is a parameter of the neural network's filter that modifies the amount of movement over the image or video. Neural network's stride is set to 1 meaning, the filter will move one pixel, or unit, at a time.

The primary convolution layer extracts low-level searching features for case edges, corners, textures, and shapes. The highest-level features are extracted in the last convolution layer.

The convolution process needs the number of filters, the receptive field size or filter size, the stride size, and the amount of zero-padding. The numbers of filters are used to control the depth of the output volume but the receptive field size or filter size determines the size of each filter (kernel). The stride size determines the number of pixels bounced when we slide the filter over the input feature. The last parameter is the amount of zero-padding, which is used to control the size of the output. The stride size of one and padding with “same” are applied when there is no change in the size of an input volume during the convolution processes (Goodfellow 2016).

The process of training a CNN model about the convolution layer is to identify the kernels that work best for a given task based on a given training dataset.

Kernels are the only parameters automatically learned during the training process in the convolution layer; on the other hand, the size of the kernels, number of kernels, padding, and stride are hyperparameters that need to be set before the training process starts.

| Building blocks of CNN | Parameters | Hyperparameters                                                                                                                                |
|------------------------|------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Convolution layer      | Kernels    | Kernel size, number of kernels, stride, padding, activation function                                                                           |
| Pooling layer          | None       | Pooling method, filter size, stride, padding                                                                                                   |
| Fully connected layer  | Weights    | Number of weights, activation function                                                                                                         |
| Others                 |            | Model architecture, optimizer, learning rate, loss function, mini-batch size, epochs, regularization, weight initialization, dataset splitting |

Table 3 lists of parameters and hyperparameters of CNN

Note that a parameter is a variable that is automatically adjusted during the training process but a hyperparameter needs to be set beforehand.

### **Activation functions (Non-linearity)**

This layer is the energetic component of a neural network that decides whether the output of a neuron is activated or not. The output of the activation function is always similar to its input dimension because no parameters are learned from them. After linear layers, it is common practice to apply nonlinear activation functions (Ramachandran P 2018). There are different activation functions; such as sigmoid, tanh, and ReLU activation functions. However, Rectified Linear Unit (ReLU) is preferable because CNNs with ReLU activation functions can be trained several times faster by keeping up the gradient more or less constant at all network layers than the same networks using tanh

and sigmoid functions. Also, ReLUs do not need input normalization to prevent them from saturating a ReLU layer implements a threshold operation to each element of the input, where any value less than zero is set to zero. This mathematical operation is equivalent to:

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \quad 2.8$$

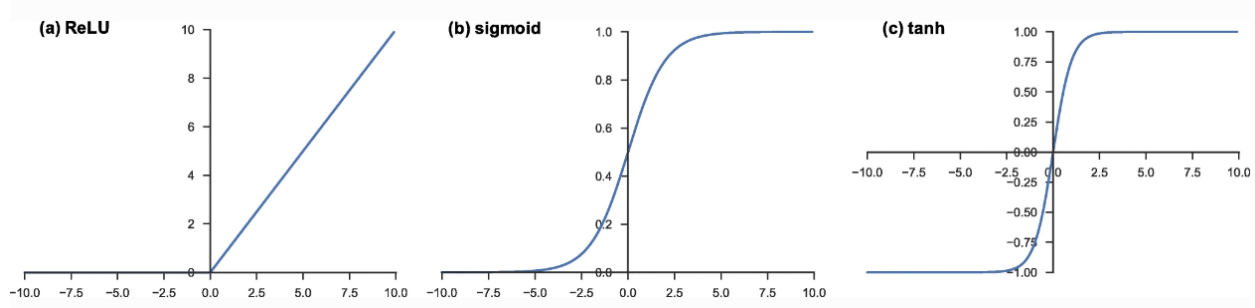


Figure 17 Activation functions commonly applied to neural networks

### Pooling layers

A pooling layer provides a typical pooling operation which reduces the dimensionality of the feature maps to present convolution invariance to small shifts and distortions and decreases the number of successive learnable parameters. There is no learnable parameter in the pooling layers, whereas filter size, and stride is hyperparameters in pooling operations.

There are different types of pooling but the most common pooling layer types are max pooling and average pooling. Max pooling, which extracts areas from the input feature maps, outputs the maximum value in each area, and discards all the other values a max-pooling with a filter of size  $2 \times 2$  with a stride of 2 is commonly used in practice. This downsamples the in-plane dimension of feature maps by a factor of 2. Unlike height and width, the depth dimension of feature maps remains unchanged (LinM 2018)

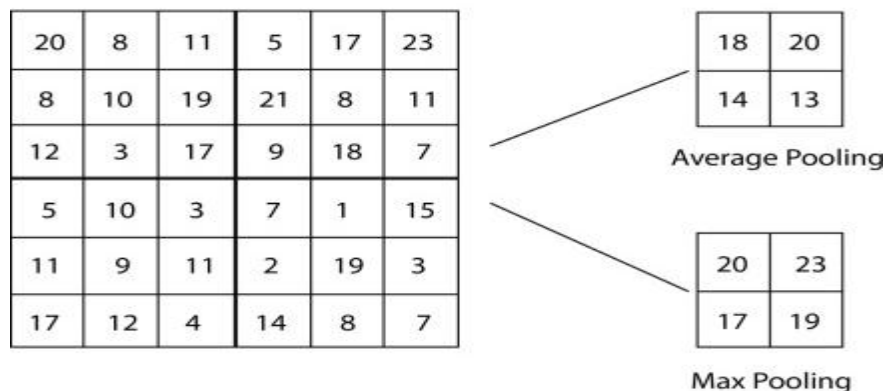


Figure 18: Pictorial representation of max pooling and average pooling.

An example of max pooling operation with a filter size of  $2 \times 2$ , no padding, and a stride of 2, which extracts  $2 \times 2$  patches from the input tensors, outputs the maximum value in each patch, and discards all the other values, resulting in downsampling the in-plane dimension of an input tensor by a factor of 2. An average pooling performs an extreme type of downsampling, where a feature map with a size of height  $\times$  width is downsampled into a  $1 \times 1$  array by simply taking the average of all the elements in each feature map, whereas the depth of feature maps is retained.

### **Fully connected**

The output feature maps of the final convolution or pooling layer are typically flattened or compressed, which means changed into a one-dimensional (1D) array of numbers (or vector in which every input is connected to every output by a learnable weight.

Once the features extracted by the convolution layers and downsampled by the pooling layers are created, they are mapped by a subset of fully connected layers to the final outputs of the network, such as the probabilities for each class in classification tasks.

The final fully connected layer usually has a similar number of output nodes as the number of classes. In an end-to-end CNN model fully connected applying before a classifier SoftMax classifier layer.

### **Last layer activation function**

The activation function applied to the after fully connected layer is usually different from the others. An activation function applied to the multiclass classification task is a softmax function that regularizes output real values from the after fully connected layer to mark class likelihoods, where each value ranges between 0 and 1 and all values sum to 1. The formula of softmax is:

$$\text{Softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad \text{for } j = 1, 2, \dots, K \quad 2.7$$

Where K is the number of classes.

The values can now represent a probability distribution over K different results.

The usual choices of the last layer activation function for various types of tasks are summarized in the table 4.

| Task                            | Last layer activation function |
|---------------------------------|--------------------------------|
| Binary classification           | Sigmoid                        |
| Multiclass classification       | Softmax                        |
| Regression to continuous values | Identity                       |

Table 4 commonly applied last layer activation functions

## Training a network and its challenge

Training a network is a process of discovering kernels (fruits) in convolution layers and weights in fully connected layers which reduce variances between output estimates and given ground-truth labels on a training dataset.

The backpropagation algorithm is the system normally used for training neural networks where loss function and gradient descent optimization algorithm play essential roles.

A model performance under particular kernels and weights is calculated by a loss function through forwarding propagation on a training dataset, and learnable parameters, namely kernels and weights are updated according to the loss value through an optimization algorithm which is called backpropagation and gradient descent, among others.

### Gradient descent

Gradient descent is normally used as an optimization algorithm that repeatedly updates the learnable parameters which are the kernels and weights, of the network to reduce the loss.

The gradient of the loss function offers us the direction in which the function has the sharpest rate of increase, and each learnable parameter is updated in the negative direction of the gradient with an arbitrary step size determined based on a hyperparameter called the learning rate.

The gradient is, mathematically, a partial derivative of the loss for each learnable parameter, and a single update of a parameter is formulated as follows:

$$W := w - \alpha \frac{\partial L}{\partial w} \quad 2.8$$

Where:  $w$  stands for each learnable parameter,  $\alpha$  stands for a learning rate, and  $L$  stands for a loss function. In practice, a learning rate is the most important hyperparameters to be set before the training starts.

In practice, for motives such as memory limitations, the gradients of the loss function about the parameters are computed by using a subset of the training dataset called mini-batch and applied to the parameter updates. This method is called mini-batch gradient descent, also frequently denoted as stochastic gradient descent (SGD), and a mini-batch size is also a hyperparameter. Also, many improvements on the gradient descent algorithm have been proposed and widely used, such as SGD with momentum (Ruder 2016). But there are other optimization methods Adagrad, AdaDelta, RMSprop, and Adam are the most commonly used optimization algorithm.

The main role of Adagrad is to calculate the adaptive learning rate during training. For this method, the summation of the magnitude of the gradient is considered to calculate the adaptive learning rate.

When the number of epochs increases the summation of the magnitude of the gradient becomes large. This makes the learning rate decrease radically and leads the gradient to approach zero quickly. This causes problems during training.

Later, RMSprop was proposed considering only the magnitude of the gradient of the immediately previous iteration, which prevents the problems with Adagrad and offers improved performance in some cases. Then the Adam optimization approach is proposed based on the momentum and the magnitude of the gradient to calculate the adaptive learning rate as similar to RMSprop. However, Adam enhances the overall accuracy and helps for efficient training with the better convergence of deep learning algorithms (Vina Ayumi 2016).

### **Data and ground truth labels**

Data and its preparation are the most important components in research applying deep learning or other machine learning methods.

To obtain better results, careful collection of data and ground truth labels with which to train and test a model is required for successful deep learning development, but obtaining high-quality labeled data can be costly and time-consuming (Clark K 2017).

Available data are typically split into three sets: training, validation, and a test set.

A training set is used to train a network, where loss values are calculated by forwarding propagation and learnable parameters are updated by backpropagation.

A validation set is used to assess the model during the training process, perfect hyperparameters, and perform model selection.

A test set is ideally used only once at the very end of the thesis to evaluate the performance of the final model that was perfected and selected on the training process with training and validation sets.

Separate validation and test sets are needed because training a model always involves adjusting its hyperparameters and performing the model selection.

This process is done based on the performance on the validation set, some information about this validation set losses into the model itself that is overfitting to the validation set, even though the model is never directly trained on it for the learnable parameters. For that reason, it is guaranteed that the model with perfected hyperparameters on the validation set will perform well on the same validation set.

Therefore, a completely unseen dataset, that is a separate test set, is necessary for the appropriate evaluation of the model performance, as what we care about is the model performance on never-before-seen data which is generalizability.

It is worthy to mention that the term “validation” is used differently in the machine learning field (Park SH 2018).

## **Overfitting**

Overfitting refers to a situation where a model learns statistical consistencies specific to the training set, which ends up learning the unrelated noise instead of learning the signal, and, therefore, performs less well on a successive new dataset.

This is one of the main challenges in machine learning, as an overfitted model is not generalizable to never-seen-before data. In that sense, a test set plays an essential role in the proper performance evaluation of machine learning models, as discussed in the previous section. A tedious check for recognizing overfitting to the training data is to monitor the loss and accuracy of the training and validation sets. If the model performs well on the training set compared to the validation set, then the model has likely been overfitting to the training data.

There have been several approaches proposed to minimize overfitting. The best solution for reducing overfitting is to obtain large training data. A model trained on a larger dataset typically generalizes better, however, that is not always possible in the audio data.

The other solutions include regularization with dropout or weight decay, batch normalization, and data augmentation, as well as reducing architectural complexity.

Dropout is a recently introduced regularization technique where randomly selected activations are set to 0 during the training so that the model becomes less sensitive to specific weights in the network (Hinton GE 2012).

Weight decay also referred to as L2 regularization, reduces overfitting by punishing the model's weights so that the weights take only small values.

Batch normalization is a type of additional layer that adaptively normalizes the input values of the following layer, justifying the risk of overfitting, as well as improving gradient flow through the network, allowing higher learning rates, and reducing the dependence on initialization (Ioffe S 2015).

Data augmentation is also effective for the reduction of overfitting, which is a process of modifying the training data through random transformations, such as reversing, translation, cropping, rotating, and random erasing so that the model will not see the same inputs during the training iterations (Zhong Z 2017).

Despite these efforts, there is still a concern of overfitting to the validation set rather than to the training set because of information leakage during the hyperparameter adjustment and model selection process. Therefore, reporting the performance of the final model on a separate test set, and ideally on external validation datasets is vital for verifying the model's generalizability.

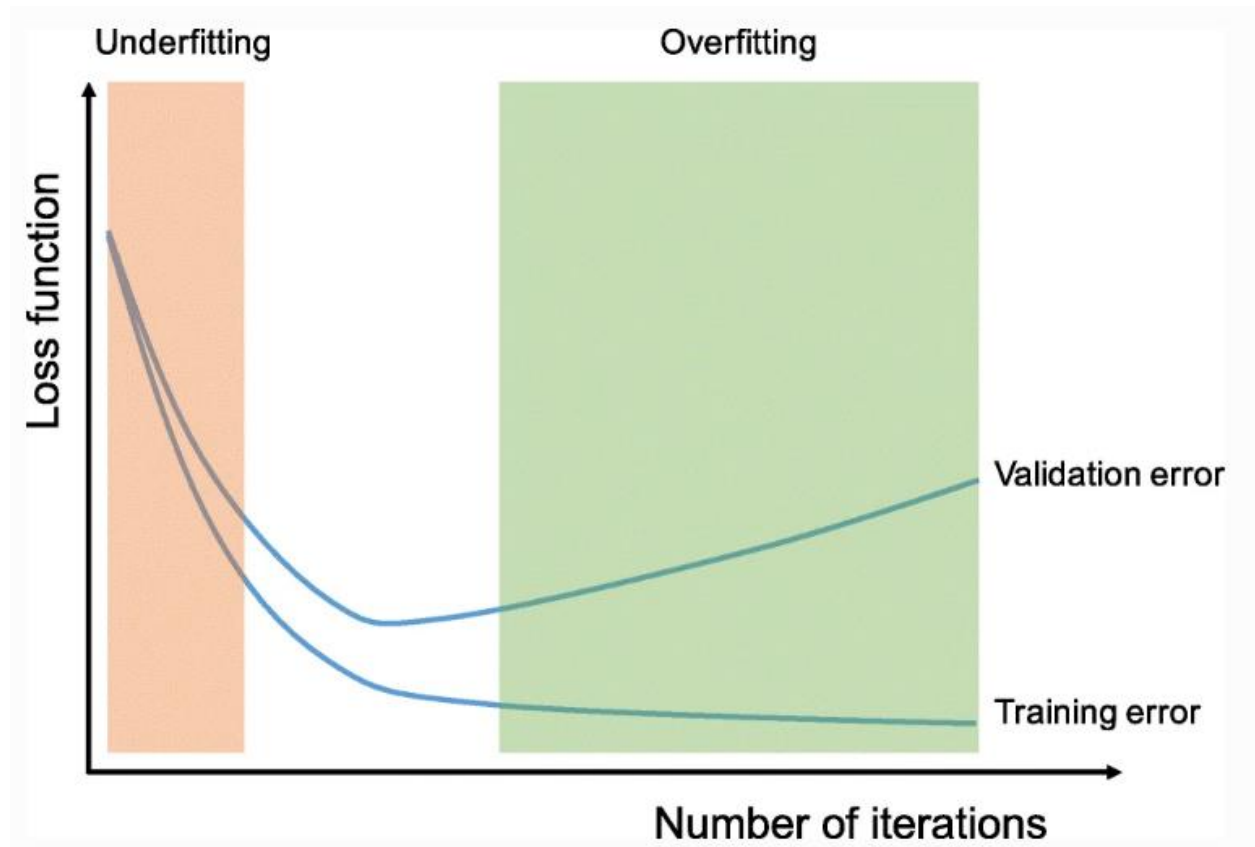


Figure 19 A routine checks for recognizing overfitting

### Dropout

Dropout is the simplest and efficient regularization technique to reduce overfitting to some extent. If the training accuracy is significantly higher than validation accuracy or the training loss is significantly less than validation losses, then that is overfitting. If training accuracy is significantly less than validation accuracy or the training loss is significantly greater than validation loss, then that is underfitting. Overfitting gives an insight that the network has memorized training data very well but is not guaranteed for unseen data. So, to minimize this problem dropout turns off a randomly selected fraction of neurons within a layer during the training process. In this case, all the neurons are not active at the same time. Hence, the inactive neurons will not be able to learn anything. The number of fractions of neurons you want to turn off is decided by a hyperparameter. In most cases, the dropout turns off parameter value is between the ranges

### 2.6 Evaluation technique

Different performance metrics are used to evaluate the performance of the proposed model. Among these, accuracy, precision, recall, and f1-score are used extensively for measuring the performance of the proposed model (Anderson 2020).

Accuracy: is the ratio of true positives (true negatives and true positives) against the whole population. However, accuracy may mislead the value of the model if the class is not balanced. Mathematically it is expressed as:  $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{P} + \text{N})$

Precision: is the ratio of true positives against the whole positives. Mathematically, it is expressed as  $\text{Precision} = \text{TP} / \text{P}$

Recall or sensitivity: is the ratio of true positives against the whole true or correct prediction. It computes how well the model avoids false negatives. It is also known as true positive rate or hit rate.  $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$

F1-score is the weighted average precision and recall. The relative contribution of precision and recall to the F1-score are equal.

$\text{F1-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

## 2.7 Related works

In speaker recognition, several studies and researches have been conducted by different researchers on different languages. In this topic, related works that are done in speaker recognition for different foreign and local languages are reviewed as follows.

Supaporn B.T.I.(2019) Presented “Text-independent speaker identification using deep learning model of convolution neural network”. Their motivational intention was to solve the two difficult challenges in speaker recognition. These challenges were using a text-independent utterance and using short utterances of speech to develop a speech recognition model. As a pre-processing phase, first, they were trying to transform each sample short utterances of the speech signal to a gray-scale spectrogram image. To collect their dataset, they were used a dataset of 5 Thai language speakers from social media (YouTube). Each speaker sample utterances were segmented into 100 samples of 2 seconds through MATLAB. Then, to conduct their experiments they used 80 samples for training and 20 samples for testing from each speaker. They try to conduct three experiments. The first one is by using MFCC as feature extraction and by using SVM as a speaker modeling technique. The second one is by using the raw waveform signal image as input to CNN. The third one is by using a spectrogram as input to end-to-end CNN. Then, they got an average accuracy of 91.26 %, 49.77%, and 95.83% is obtained for experiment numbers 1, 2, and 3 respectively. Finally, they concluded that CNN with spectrogram image input is very capable for the text-independent approach. However, the used dataset is limited and their dataset is not public since they didn’t describe the detailed information about the used audio samples. On the other hand, SVM has good generalization ability and can solve the small sample problem. Therefore, we proposed a CNN-SVM combined model to make use of their advantages.

Ankur Mauryaa (2017) in their work entitled “Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach” presented text-dependent and text-independent speaker identification for the Hindi language. They were used Mel frequency cepstral coefficient (MFCC) as a feature extraction technique for both text-dependent and text-independent applications. As speaker modeling technique they were used vector quantization (VQ) and Gaussian mixture model (GMM) for text-dependent and text-independent tasks respectively. To train their model they used a dataset of 15 speakers (10 male and 5 female). From each speaker, 17 speech samples were collected. Finally, they found an accuracy of 77.64% and 86.27% for text-independent MFCC-VQ and MFCC- GMM respectively. However, for text-dependent, they found an accuracy of 85.49 % and 94.12 % for MFCC-VQ and MFCC-GMM models respectively. However, GMM fails to work if the dimensionality of the problem is too high.

Radoslaw Weychan (2010) in their work carried out experiments base on a database of short polish language utterances. The scope of their study was to develop a fast speaker recognition using a 1seconds record of polish language utterances. However, characteristically automated speaker recognition systems may need up to 7seconds. To achieve the scope of their study they were used Mel Frequency Cepstral Coefficients (MFCC) as features, Gaussian mixture model (GMM), and vector quantization (VQ) as feature matching techniques. To train their model they used a dataset of 25 speakers. Finally, they found an accuracy of 90% and 84% results for GMM and VQ respectively. However, GMM and VQ are Template matching algorithms. This is ineffective in terms of both required storage and processing power needed to perform the matching.

Aykefam (2015) presents the implementation of a text-independent speaker identification system for the Amharic language. As a feature extraction method, they have used MFCC,  $\Delta$ MFCC, and  $\Delta\Delta$ MFCC and as a speaker modeling, they have used VQ and GMM. To implement his work, they have used a dataset of 50 speakers (25 male and 25 female). And each speech signal was taken with a sampling frequency of 16 kHz and 16 bits. They were trying to conduct the first experiment on a total of 50 speakers using MFCC features and they got 74.2% accuracy for VQ and 84.3% for GMM. However, they were got improved results when they were used an increased number of features  $\Delta$ MFCC and  $\Delta\Delta$ MFCC. After the first experiment, they have conducted the second experiment based on gender groups using 25 males and 25 females. Then, the accuracy will be achieved at 86.2% and 85.9% for males and females respectively. Finally, they have conducted the last experiment based on age groups by considering 20 speakers whose age is less than or equal to 40 years and 20 speakers whose age is greater than 40 years. And they were got 86.8% accuracy for the age group that is less than or equal to 40 years and 86.3% accuracy for the age group is greater than or equal to 40 years. Finally, it was concluded that speaker identification is age and gender-dependent. However, MFCC is

disposed to background noise and it considers only the power spectrum of a signal. To handle this problem we used a convolutional neural network since CNN can handle the noise by itself.

Abraham(2017), has implemented text-independent speaker identification for the Amharic language in a noisy environment.

As a feature extraction technique, they used hybrid MFCC and GFCC feature vectors and for speaker modeling, they were used VQ, BPNN, and GMM models. The study aimed to analyze the performance of different algorithms for speaker identification. First, they were used Principal component Analysis for feature reduction from 39 feature vector dimensions of MFCC and GFCC to 24 feature vectors. Then, they used a Genetic algorithm for feature reduction from 24 feature vector dimensions to 14 feature vectors. For them, this enables to reduce the training time of the system.

To implement their work, they used a dataset of 90 speakers. Three utterance samples were gathered from each speaker with a time of 10 seconds long. 270 speech records were recorded at each sampling rate of 16 kHz and 16-bit using an audio recorder. Based on their experiment, as a result, they were got an accuracy of 59.2%, 70.9%, and 84.7% for VQ, GMM, and BPNN respectively. Using hybrid features and a sufficient amount of dataset for speaker recognition enhances the accuracy of the system as compared with the previously reviewed literature. However, these classical pattern recognition models use handcrafted features and a comparatively simple trainable classifier. In addition to this, BPNN may not converge quickly during the training and it is too slow for real applications. Handcrafted features require domain knowledge for researchers to be designed and take a long time to be tested and selected for high classification performance.

| Authors and year                        | Features technique                          | Recognition technique | Dataset                                          | Best result in %                                                                                         | Limitations                                                                                                                            |
|-----------------------------------------|---------------------------------------------|-----------------------|--------------------------------------------------|----------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| (Aykefam Azene, 2015)                   | MFCC, $\Delta$ MFCC and $\Delta\Delta$ MFCC | VQ, GMM               | 50 speakers                                      | 74.2 and 84.3 for VQ, GMM respectively for a total of 50 speakers 86.2 for 25 males, 85.9 for 25 females | In background noise, MFCC doesn't perform precise results and its efficiency could be influenced by the number of filters              |
| (Abrham Debasu, Dagnachew Melese, 2017) | MFCC, GFCC                                  | VQ, BPNN, &GMM        | 90 speakers<br>10-second duration<br>270 samples | 59.2 for VQ, 70.9 for GMM, 84.7 for BPNN                                                                 | BPNN may not converge quickly during the training and it is too slow for real applications b/c it can be quite sensitive to noisy data |

Table 5 comparisons of various feature extraction and recognition techniques for speaker identification in Amharic languages

## CHAPTER THREE

### 3 DESIGN OF AUTOMATIC SPEAKER IDENTIFICATION

#### 3.1 Introduction

A speaker recognition system is molded by various massive systems each of which response to different tasks. As already mentioned, a speaker recognition system can perform in training and testing models. The speech input signal is initially digitalized, pre-emphasized, and segmented into short quasi-stationary frames, windowing, silence removals, change into a spectrogram, resize, and change RGB to grayscale. Because, RGB has a 24-pixel value in the three colors but in grayscale, it is 8-pixel values. After that, a feature vector is extracted from each frame.

This feature vector represents a speech properly just by taking the speech parameters which are significant for the speaker recognition duty.

**Training model:** - The speaker's identity is recognized as a priori. The speaker produces a given length voice signal that is similar to all the speakers. The feature vectors are got in a similar way as in the earlier model and also, each frame is classified using the voice classification algorithm.

The speaker models are initialized so that they get some values that are supposed to be a good start. Each speaker is harmonized with N models, being N the number of classes into which voice is classified. Then, the pattern matching algorithm re-estimates the values using the order of feature vectors and the voice class category information.

**Testing mode:** - In this case, is a single random word or phrase speech input that is produced by a priority unknown speaker. In the feature extraction procedure, an arrangement of feature vectors is found. This sequence is compared with a given set of speaker models using a pattern-matching algorithm (SVM). This algorithm produces a probability score for each frame and speaker. With this information, a decision-making algorithm gives an estimation of which was the speaker that produced the speech signal.

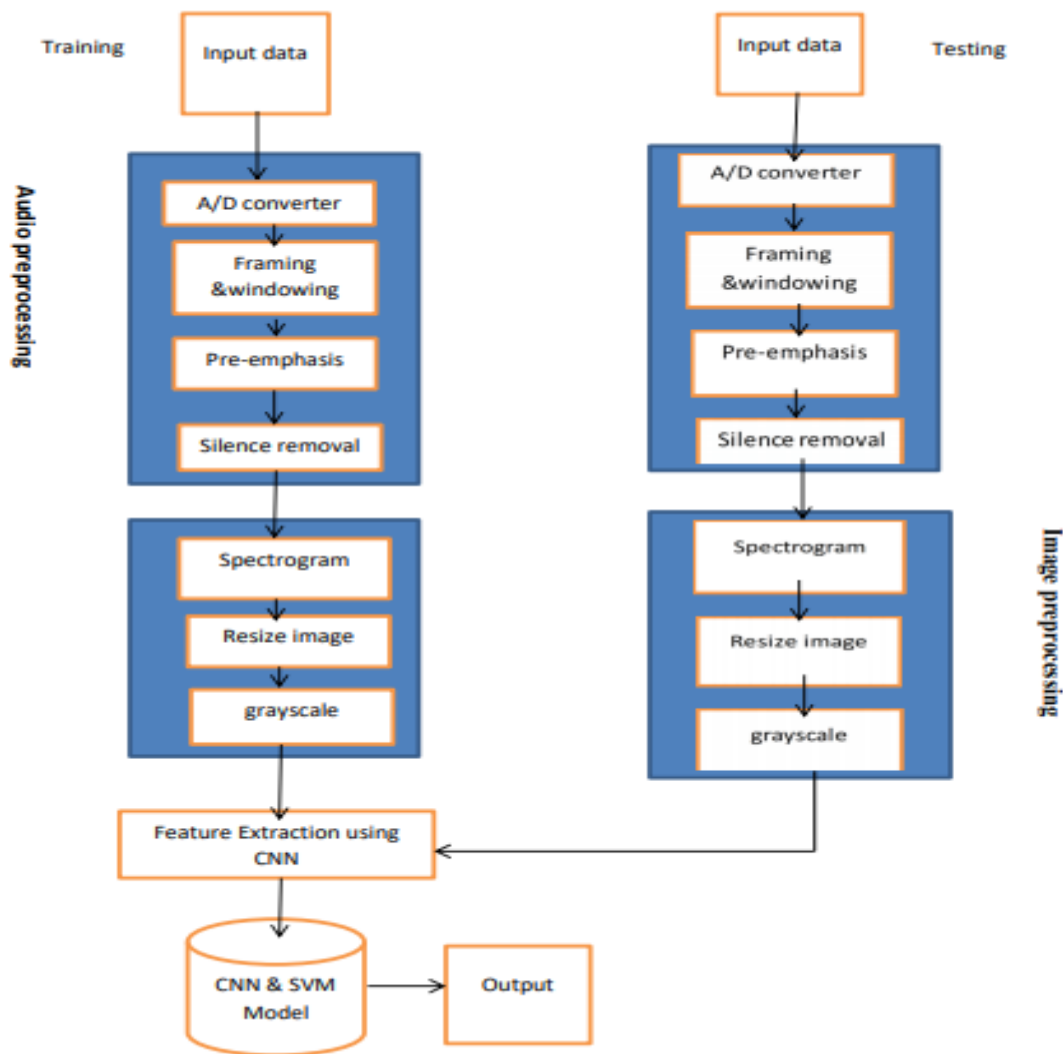


Figure 20 detail a diagrammatic representation of the Speaker identification system

The detail of the diagram is to be discussed as follows. Generally, all these procedures can be divided into four main subsystems.

1. **Signal pre-processing**: separating the voiced and silence/unvoiced portion of the captured signal is usually supported as a crucial step in the advance of a reliable speech or speaker recognition system. Because most of the speaker-specific characters are present in the voiced part of the speech signals. Similarly, the extraction of the voiced part of the speech signal by marking and removing the silence and unvoiced part leads to a large reduction in computational complexity at future stages. Mainly, it includes the A/D conversion, basic filtering, and segmentation into frames, etc. (Kasymbek. A 2017).
2. **Feature extraction**: The human voice contains numerous discriminative features that can be used to identify speakers. Feature extraction is one of the most important aspects of speaker recognition and generates a vector that represents the speech signal (M. M. Thiye Modipa 2019).

3. **Modeling:** this subsystem has the voice classification algorithm and the right pattern matching algorithm. It also holds a database of speaker models and an algorithm to initialize a model prototype when a new user has to be added to the system. Its functions are to initialize and re-estimate speaker models in the training phase and to output a likelihood score for each speaker in the testing phase (Ruili Wang 2017).
4. **Decision:** this subsystem gives the final result taken about the estimated speaker who produced the test utterance in the testing phase. This decision is based on the likelihood scores given by the pattern matching system.

### 3.2 Signal Preprocessing

Before the feature extraction phase, the speech signal that is used either as training or as test input data to the speaker identification system has to be preprocessed.

The purpose of preprocessing is to derive a set of parameters to represent speech signals in a form that is suitable for successive processing and to process the sampled speech signal and produce a representation that is independent of amplitude variations, speaker stress, and noise introduced from the transmission medium (B. Singh 2012).

Speaker identification is worked using the two different modes called training, and testing. For all training and testing modes, the following common procedures are performed. The gathered speech signal is sent to the preprocessing step. After receiving the speech signal the preprocessing stage of the speech is started. The preprocessing includes file format conversion, Pre-emphasis, framing, and windowing operations then remove silence. After preprocessing the signal is converted to spectrogram images by applying a fast Fourier transform digital signal processing technique. Then the generated RGB spectrogram images are changed to grayscale then it is given to the feature extraction block. Then, feature vectors are extracted by a convolutional neural network from each spectrogram image of a signal and are used for model training and testing. Finally, the trained model prints out the speaker's identity in the last block of the model.

#### 3.2.1 A/D Conversion

The voice signal is by default a continuous air-pressure signal that can be taken by a microphone. The recorder converts this pressure-signal into a continuous electrical signal. The A/D converter goal is to convert this continuous representation into the discrete domain so that it can be processed in the digital domain (I. Y. Kelbesa 2014). To do this operation, the A/D module takes a sample of the original signal with frequency ( $f_s$ ) called sampling frequency.

According to the Nyquist theorem, this sampling frequency must be at least double the input signal's bandwidth to avoid spectral aliasing. In speech signal this bandwidth can be approached to 8 kHz

since above this boundary there is no relevant information, so a sampling frequency is 16 kHz. Also, most of the information in the speech is located below 4 kHz. This is abused by the telephone lines which apply a 4 kHz low pass filtering to the voice signals. Thus telephone lines decrease significantly the speech signal bandwidth without an important loss in the quality of speech.

The A/D converter also quantizes the amplitudes of the samples obtained into symbols or sequences of bits. This is called PCM (pulse code modulation). Normally 8 to 16 bits provide sufficient resolution in the voice signal quantization then coding it.

### 3.2.2 Pre-emphasis

Usually, speech is pre-emphasized before any further processing. Pre-emphasis is a process of passing the signal through a filter, which emphasizes higher frequencies. In the speech signal, most of the energy is carried by the low frequencies. When the frequency increases pre-emphasis also increases the energy of the signal. It also serves to emphasize the formant peaks, to make them more “visible” in the spectrum. Pre-emphasis makes the spectrum flatter by raising the energy in the high-frequency region.

The pre-emphasis filter is designed to compensate for the effect of the glottal-source and energy radiation from the lips (Kushagra 2016). When the spectrum of the recorded speech signal is 6 dB/octave lower than the original, i.e. vocal tract and spectrum. In a digital signal processing system, pre-emphasis is a digital high pass filter that processes the digitized speech signal. The high-pass filtering action may be achieved digitally using the first-order difference equation. Such filter is implemented by the formula:

$$Y(n) = s(n) - a s(n-1) \quad (3.1)$$

Where  $y(n)$  denotes the current output sample ( $n$  is the sample index) of the pre-emphasis filter,  $s(n)$  is the current input sample,  $s(n-1)$  is the previous input sample and ‘ $a$ ’ is a pre-emphasis constant and its transfer function is given by:

$$Y(z) = 1 - az^{-1} \quad (3.2)$$

Where  $a > 0$  controls the slope of the filter, in speech recognition experiments, ‘ $a$ ’ varies from 0.94 to 0.97 with 0.95 as the most commonly used value.

The magnitude response and phase response of the first order FIR filter with 0.97 pre-emphasis factors are shown in the figure 21. Consequently, the output is formed as follow:

The goal of pre-emphasis is to reward for the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants. The next example demonstrates the effect of pre-emphasis.

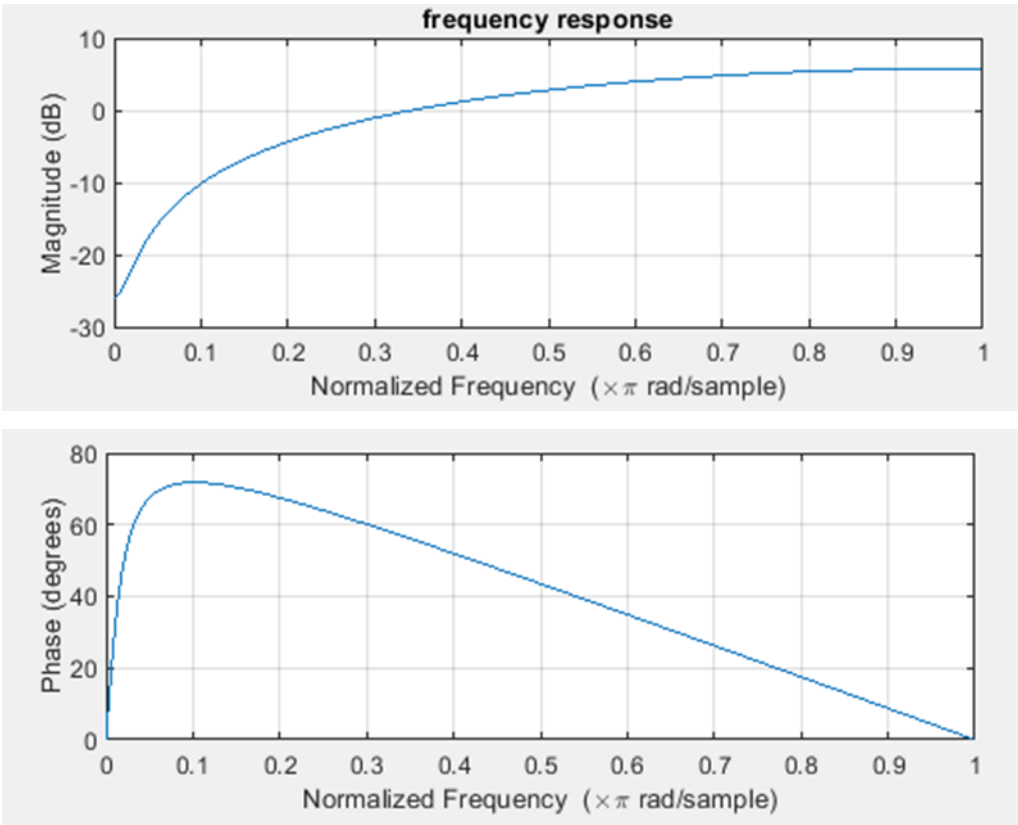


Figure 21: Magnitude responses and Phase response of a first order

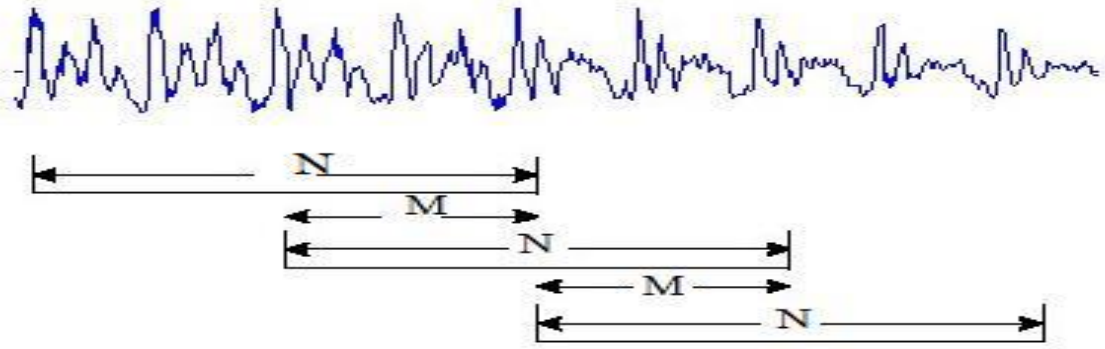
FIR filter the first-order FIR filter with a 0.97 pre-emphasis factor works as a high pass filter.

### 3.2.3 Framing

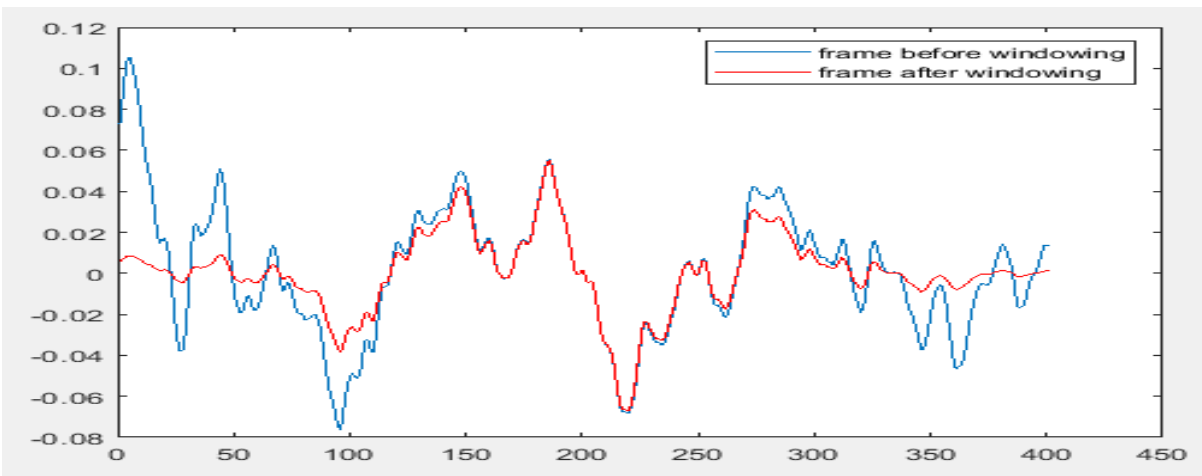
Speech is a non-stationary signal and its characteristics may change in a very short time instances. To capture the variability in the waveform, every state-of-the-art system firstly segments the signal into frames. At a given, very short time frame (10-30ms), the speech segment is close to stationary. In this interval, the speech signal remains unchanged. The features are extracted from these frames. The original Fourier transform operates on a signal of theoretically infinite length, so the Short-time Fourier transform (STFT) requires that each frame be somehow explained to infinite length. For extracting the spectral features of a speech signal a short-time analysis is applied (A. Ahmad 2015). When the signal is blocked into frames of  $N$  samples, between two adjacent frames; there is an overlap of  $M$  samples (hop size). The usefulness of overlapping is given by the fact that a single frame in uniform segmentation can contain a non-stationary transition between two frames. Usually, the overlapping region ranges from 0 to 75% of the frame size. But, an overlap of half the window size (or less) is representative.

In this thesis, we used 0.025-sec frame duration and 16000 sampling frequency with 50 % overlap.

Figure 9(a) illustrates the blocking into frames for the case in which  $M = (1/2) N$ .



(a) Blocking of speech into overlapping frames ( $N = 2M$ )



(b). Windowing speech into without overlapping frames

Figure 22 Speech frame blocking and windowing.

### 3.2.4 Windowing

Once the frame blocking procedure is completed, to every frame a windowing function is applied to suppress the effect of discontinuities at frame edges. The purpose of the windowing is to reduce the effect of the spectral artifacts that result from the framing process. The DFT computation assumes that the input signal repeats over and over. If there is a gap between the first point and the last point of the signal, artifacts occur in the DFT spectrum. By multiplying a window function to easily reduce both ends of the signal near zero; these unwanted artifacts can be avoided. Windowing in the time domain is a point-wise multiplication of the frame and the window function.

According to the convolution theorem (Blackledge 2005), this corresponds to the convolution of the short-term spectrum with the window function magnitude response. Hence each frame is multiplied with a window function  $w(n)$  with length  $N+1$ , where  $N$  is the length of the frame. The most popular window is the Hamming window (A. Ahmad 2015).

The hamming window, as shown in figure 23, is usually used in speech signal spectral analysis, because its spectrum falls off rather rapidly so the resulting frequency resolution is improved, which is suitable for detecting formants. The mathematical expression of the Hamming window is shown in Equation 3.3.

$$W(n) = 0.54 - 0.46\cos(2\pi n/N-1) \quad n=0, 1, 2, \dots, N-1 \quad (3.3)$$

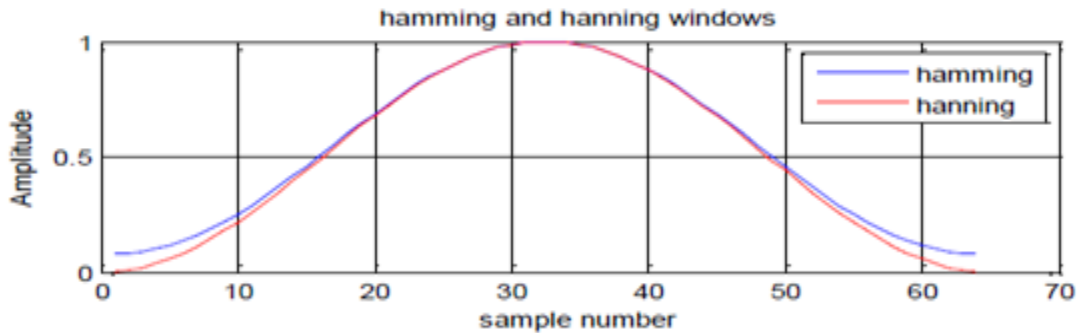


Figure 23: Comparison of Hanning and Hamming window

### 3.3 Silence Removal of Speech Signal

Silence removal and filtering is the process of extracting out the silence part from the speech signal; otherwise, the training might be seriously biased. Speech is not produced in the silent region, but in

unvoiced regions, the vocal cords do not vibrate, which makes the resulting speech waveform periodic or random.

In the voiced region, the vocal cords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic.

It should be clear that the segmentation of the waveform into the defined regions of silence unvoiced and voiced is not exact; it is often difficult to distinguish a weak, unvoiced sound from silence or weak voiced sound from unvoiced sounds or even silence. But, it is not always serious to segment the signal to that much less than several milliseconds; therefore, minor errors in boundary locations usually have no consequence for most applications. Meanwhile, for many of the sensible cases, the unvoiced parts have low energy content and thus silence (background noise) and unvoiced part is assessed together as silence/unvoiced and are categorized from the voiced part (D. A. Reynolds 1997). To remove those silence parts we used frame by frame methods.

### **3.3.1 speech silence removal using frame by frame**

The method used in this work is simple and may overcome the problem of removing silence from the speech.

To remove silence from the speech using frame by frame methods first, the audio is loaded and the original signal is plotted. To remove silences, the signal is first split into frames of 10s. The speech signal is then divided into intervals within each frame without any overlapping. The timing used for splitting the framing is energetic as it determines the frame length and also the number of frames.

Through the use of looping, each frame is next extracted and the non-silent frames with the maximum amplitude of less than 0.03 are identified.

This task is repeated until it reached the end of the signal. Once no more frames can be extracted in the process, the new signal is then plotted without the silence signal.

Initially, K is assigned as 1 to start the looping and its value will increase after each loop as we can see in the flow chart.

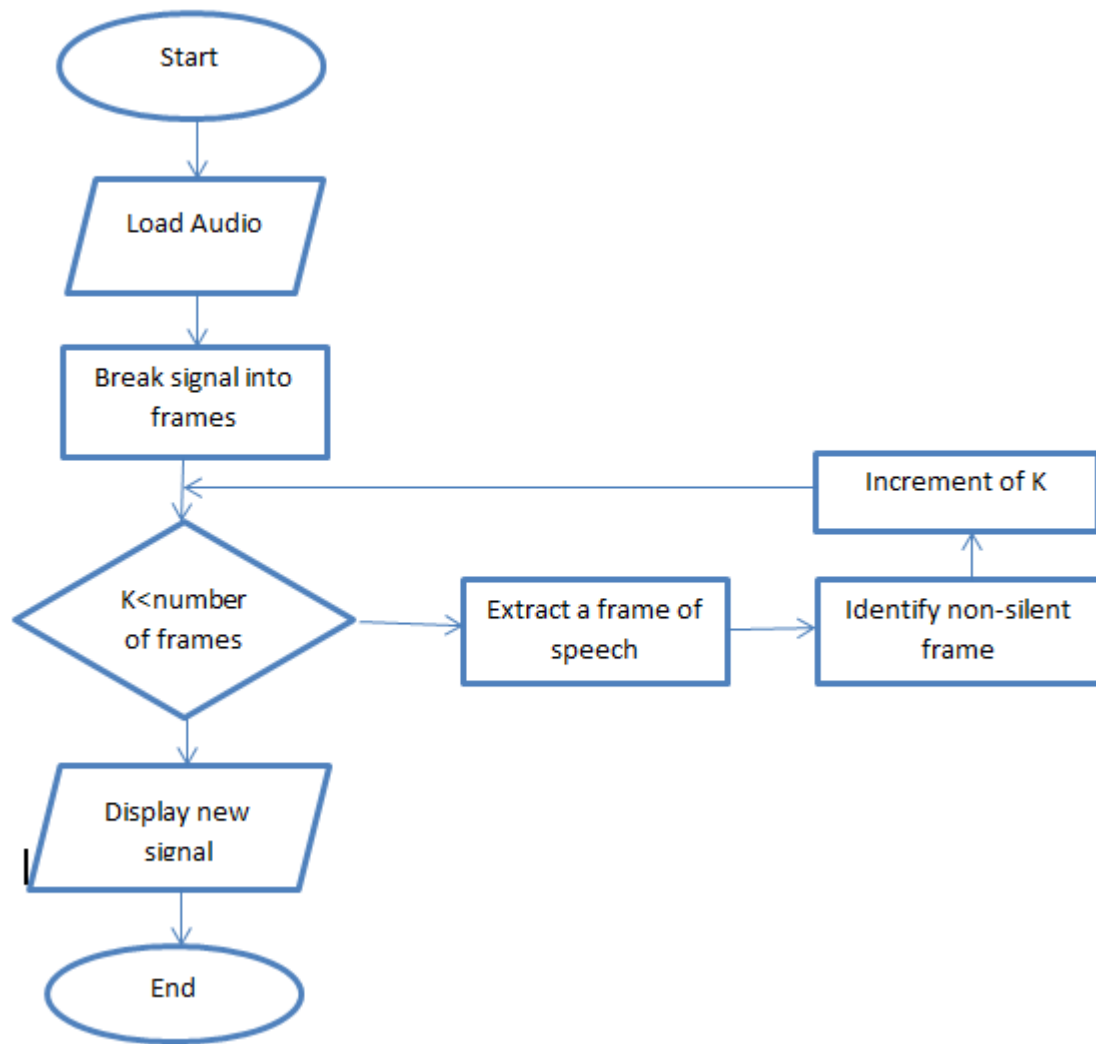


Figure 24 frame by frame silence removal technique flow chart.

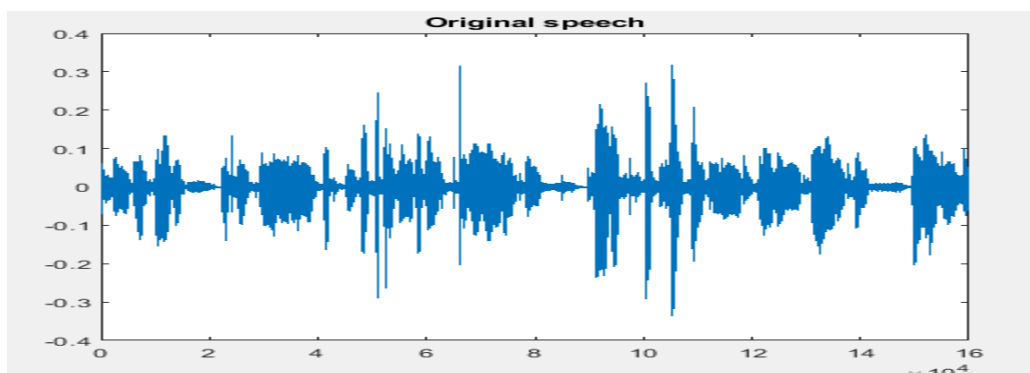


Figure 25: Original speech

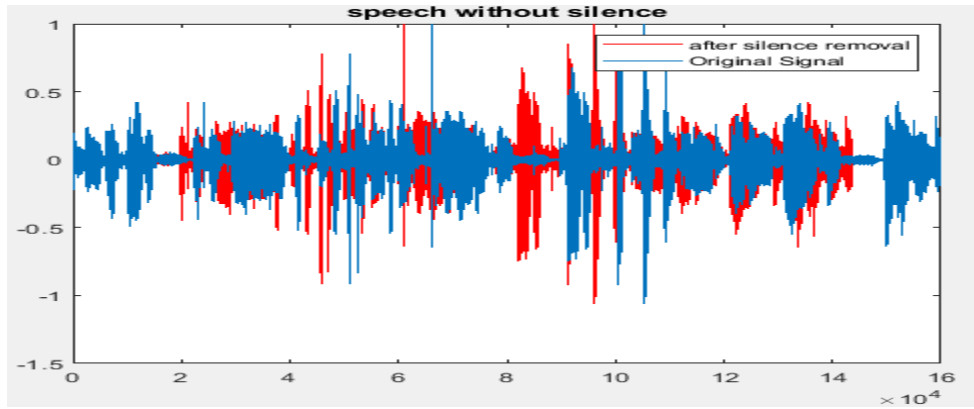


Figure 26: speech without silence

### 3.4 Spectrogram

The speech signal is one-dimensional and non-stationary. This means that the frequency domain of the signals does not remain constant to the time variable. In terms of data exist speech signal exists only in one dimension while image exists in 2D space. To implement deep learning convolutional neural networks for the speaker recognition task, first our wave file audio dataset signals are converted from one -dimension signal to a two-dimensional spectrogram image and save as an image file in PNG format. A spectrogram is a visual representation of the spectrum of frequencies in a sound.

The Time-frequency (spectrogram) representation of a signal describes how a spectral component in a signal evolved in the function of time. This approach enhances the patterns that may not be visible in the original signal and directly show relevant information in the audio signal. Where the horizontal dimension represents time (reading from left to right), and the vertical dimension represents frequency.

To convert the speech signals into a spectrogram image the Short-Time Fourier Transform (STFT) is used. The Short-Time Fourier Transform (STFT) is preferred as it represents the spectrum changing over time. STFT uses a sliding FFT window to obtain spectra for each segment in time of the original signal.

To generate the spectrogram image, we used the spectrogram function with its input parameters like speech signal, sampling frequency, windowing, noverlap, and nfft as shown below:

$S = \text{spectrogram}(\text{Signal}, w, \text{noverlap}, \text{nfft}, fs);$

Where Signal is the input speech signal,  $w$  is the frame size or window length which divides the speech signal into several segments and performs the windowing function. It is calculated by multiplying the  $f_s$  with its frame duration.

Noverlap or hop size is the number of overlap samples between each adjacent segment. Its value is found by multiplying the frame size with 50 % overlap. NFFT is the number of discrete Fourier transform (DFT) points. Its value is taken as the maximum value between the range (256,  $2^{\text{nextpow2}}(\text{Frame size})$ ). Finally,  $f_s$  are the sampling frequency of a signal which is 16 kHz. The number of frames is calculated by dividing the length of a signal ( $f_s \cdot 10\text{sec}$ ) by frame size.

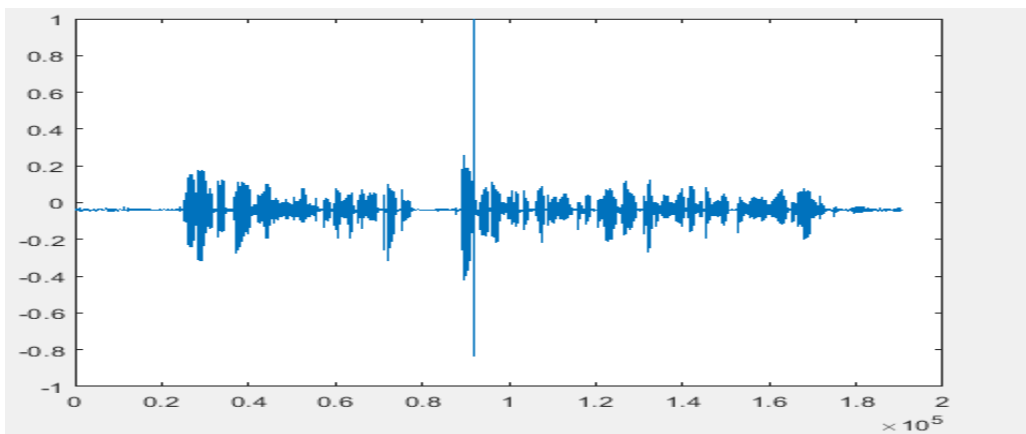


Figure 27 original signal

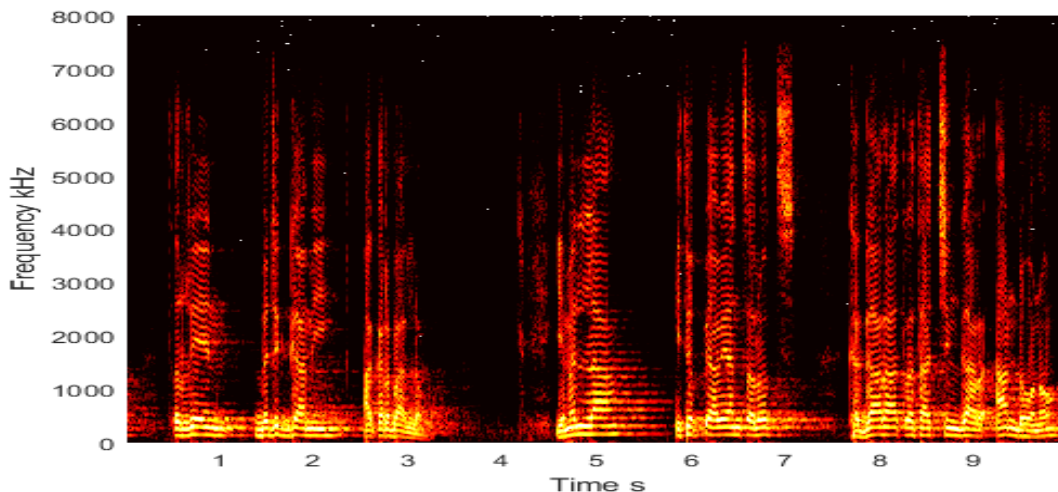


Figure 28 spectrograms of the speech signal

### 3.5 Resizing

The original generated spectrogram images have 560 by 420 dimensions and scale. Training our model CNN using this large image size using a local machine is impossible. So, to simplify this challenge we used the image resize MATLAB function.

To reduce the visual distortion or the information loss caused by the image resizing process we used the bicubic interpolation technique (Han 2013). The original RGB image is resized to 224 by 224-pixel images. To do this we used the following pseudo-code algorithm: a distortion or rotating an image.

|                                                             |
|-------------------------------------------------------------|
| Step1: Input: Accept the input image.img                    |
| Step2: Output: an array of resized image (to [output Size]) |
| Step3: im = imread(filename.png)                            |
| Step4: Image = imresize (im, [output Size], 'cubic')        |
| Step5: Image = image_to_array (Image)                       |
| Step6: Return Image                                         |
| Step7: End                                                  |

Table 6 Spectrogram Image resizing algorithm

### 3.6 Grayscale

A grayscale level is a range of monochromatic (gray) shades from black to white therefore a grayscale image contains only shades of gray and no color. Grayscale images are those images where color information is missing and all color information is converted into the grayscale format. Gray-scaling is the process of converting a continuous tone image to an image that a computer can manipulate while gray scaling is an improvement that requires larger amounts of memory because each dot is represented by from 4 to 8 bits.

A digital image is composed of groups of three pixels with colors (RGB) called channels in digital imaging each channel also contains a luminance value to determine the color to get a grayscale image, the color information from each channel is removed leaving only the luminance values and that is why the image becomes a pattern of light and dark areas empty of color essentially a black and white image (Kriti 2019).

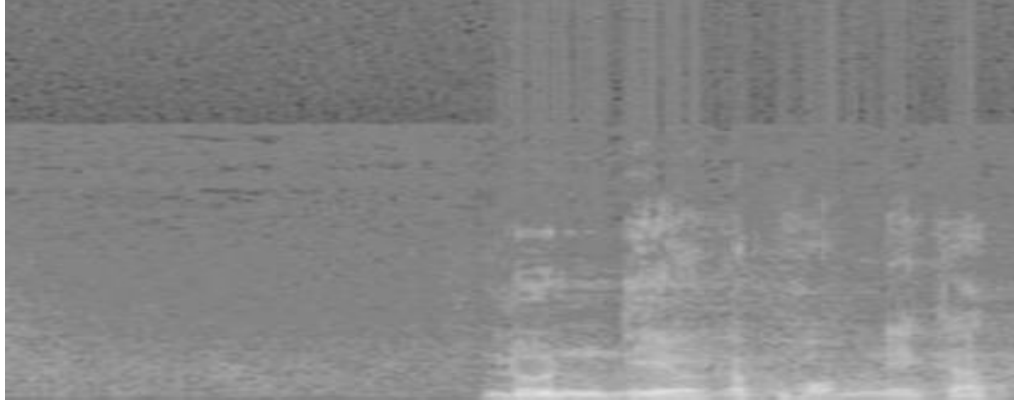


Figure 29 grayscale image.

### 3.7 Feature extraction

#### 3.7.1 Introduction

In the literature review, we discussed how the human speech was produced to find the speaker-specific characteristics of the speaker recognition task. To make the human speaking production processable from the signal processing point of view, discrete-time modeling was discussed to model the procedure as a source-filter model, where the vocal tract is viewed as a numerical filter to shape the sound sources from vocal cords. The speaker-specific characteristics include two main sources: **physical (low-level cues)** and **learned (high-level cues)**. Although high-level features are recently exploited successfully in speaker recognition, especially in noise environments and channel mismatched cases, our attention is on the low-level spectral features because they are widely spread, easy to compute and model, and are plentiful more related to the speaking production mechanism and source-filter modeling. With an overview of the mechanism of speech production, the aim of the front-end processing becomes clear, which is to extract the speaker's discriminative features.

#### 3.7.2 Review of Feature Representations

The function of the measurement phase of a speaker recognition system is to perform several characterizing measurements on the voice pattern under test. The speaker's exact characteristics of speech can be considered physical and learned. The physical characteristics are the inherent part shapes and sizes of the speech production organs, like vocal folds and vocal tract. Since the resonances of the vocal tract and the characteristics of the sound energy sources depend upon just these anatomical factors, physical/organic differences lead to differences in fundamental frequency,

laryngeal source spectrum, and formant frequencies and bandwidths. The learned characteristic rhythm, intonation style, accent, choice of vocabulary, and so on.

They are the result of differences in the patterns of coordinated neural commands to the separate articulators learned by each individual. Such differences give rise to variations in the dynamics of the vocal tract such as the rate of formant transitions and co-articulation effects. Naturally, many speaker-dependent characteristics are affected by both of these factors. Ideal speaker-discriminative feature representation is expected to be (P. Rose 2002.):

- Require large inter-speaker variability and small intra-speaker variability,
- It is reasonably robust to background noise and distortions,
- It Occurs naturally and frequently in normal speech,
- Be easily measurable,
- Be stable over time or not be affected by the speaker's health/mood,
- Be difficult to mimic.

Also, the dimension of features should have to be low because otherwise, the computation cost would be high, and discriminative models such as the support vector machine (SVM) cannot handle high-dimensional data (Sunil Kumar Singla 2017).

The features for speaker recognition can be classified into:

#### **I. Short-term spectral features**

- ❖ Describing the short-term spectral envelope that is an acoustic related to timbre, as well as the resonance properties of the vocal tract.
- ❖ Auditory frequency warping, bank-of-filters model, dynamic coefficients appending, etc.

#### **II. Voice source features**

- ❖ Characterizing the glottal flow.
- ❖ Pitch determination, pitch-synchronous analysis, pitch-epoch localization,

#### **III. Spectral-temporal features**

- ❖ Interpreting speaker properties in flexible time-frequency resolutions.
- ❖ Sub-band energy separation, multiple frequency band demodulation, etc.

#### **IV. Prosodic features**

- ❖ Including pitch, intonation, duration, and rhythm, usually span over tens or hundreds of milliseconds.

- ❖ F0 tracking, dynamic coefficients appending, etc.

## **V. High-level features**

- ❖ Attempting to capture the conversation-level characteristics of speakers.
- ❖ Speech recognizer, statistical language modeling, etc.

Generally speaking, short-term spectral and voice source features are relatively easy to extract, and there is no need for a huge amount of data. Up until now, the short-term spectral features have always dominated the front-end of the leading speech, speaker, even language recognition systems. Besides the stable performance provided, their low demand on computational cost makes the real-time application feasible. However, the biggest challenge they are faced with is parameter degradation in the presence of background noise and channel mismatch circumstances.

Prosodic and high-level features are supposed to be more robust, but less discriminative and easier to mimic. High-level features, since connecting with the personalized lexicon and recording the dialectal pattern of individual speakers, are less affected by the variation in noise or channel conditions. The high-level speaker-related characteristics are whereas difficult to extract, and there will be a lot of training data needed in the feature extraction process (C. D. Acken 2016).

Thus, it is hard to apply this type of feature to real-time recognition tasks considering their delay in making the decisions. In general, there are not existing globally “best” features yet but the choice is a trade-off between speaker discrimination, robustness, and practicality.

The state-of-the-art speaker recognition system often combines these features, attempting to achieve more accurate recognition results (Nakasone 2003).

In this thesis, we used the short-term spectral features since it is relatively easy to extract, having low demand on computational cost, and having the stable performance it provided. Specifically using CNN as feature extraction has become increasingly popular.

### **3.9.1 Convolution Neural Network (CNN) As Feature Extraction**

In this thesis, we use CNN as a feature extraction method for deep feature extraction. Because CNN has great power in image processing as mentioned in the literature parts.

Images with size  $M \times N$  are the inputs in CNN-based feature extraction. This image passed through many layers the called hidden layers and by applying operations like convolution, pooling on the image we get useful information for the recognition stage.

We use the Relu activation function because it is powerful. And also use maximum pooling because it reduces the dimension of the feature vector and it is also suitable for images that have noise.

As shown in the table 7 the first stage is providing the input image by specifying its height and width. The height and width of the image are set to be 224 x 224. Bicubic interpolation image resampling techniques are selected for image resizing because it is effective in all applications of image processing. This size is selected randomly and it is within the range that CNN performs best in the literature which is 64 up to 360.

The next stage in the network is convolution and pooling layers. The network contains three convolution and two pooling layers.

The kernel size of the convolution and pooling layer is set to be three and two respectively. Finally, the feature obtained from the last pooling layer is feed to flatten layer and this layer changes the feature from multi-dimensional to one-dimensional vector.

Finally, for the layer before output, the softmax function is applied to transform the output values of the network in terms of probability. The softmax function is defined by.

$$\text{Softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, j = 1, 2, \dots, K \quad (3.17)$$

Where  $z_j$  is an output of each  $j$ ,  $e^{z_j}$  is the exponential value of, and  $k$  is the component of vector  $Z$ .

| Layer No. | Layer Name            | Detail                             |
|-----------|-----------------------|------------------------------------|
| 1         | Input                 | 224x224x1                          |
| 2         | Convolution           | 3x3,8 filters, 'Padding', 'same'   |
| 3         | Batch Normalization   |                                    |
| 4         | ReLU                  |                                    |
| 5         | Max Pooling           | 2x2, Stride 2                      |
| 6         | convolution           | 3x3,16 filters, 'Padding', 'same'  |
| 7         | Batch Normalization   |                                    |
| 8         | ReLU                  |                                    |
| 9         | Max Pooling           | 2x2, Stride 2                      |
| 10        | convolution           | 3x3, 32 filters, 'padding', 'same' |
| 11        | Batch Normalization   |                                    |
| 12        | ReLU                  |                                    |
| 13        | dropout layer         | 0.5                                |
| 14        | Fully Connected       | 50 neurons                         |
| 15        | SoftMax               |                                    |
| 16        | Classification Output | 50 classes                         |

Table 7 the structures of proposed CNN model

### 3.10 Classification

Pattern classification involves computing a match score in the speaker recognition system. The term match score refers to the similarity of the input feature vectors to some model. Speaker models are built from the features extracted from the speech signal. Based on the feature extraction a model of the voice is generated and stored in the speaker recognition system. To validate a user the matching algorithm compares the input voice signal with the model of the claimed user. In this paper, two

techniques in pattern classification have been compared. Those two major techniques are CNN and SVM.

### **3.10.1 Support Vector Machine (SVM)**

A Support vector machine (SVM) was developed by Vapnik (1998). Its model is closely related to neural networks and the form of a supervised machine learning model of learning by providing examples of the training data, the model finds a function that couples input data to the correct output. The output for novel data can then be predicted by applying the saved function. SVM is often used for classification problems for which the correct output is the class the data belongs to. The model works by creating a hyperplane that separates data points from one class to the other class, with a margin as high as possible.

The margin is the maximal width of the chunk parallel to the hyperplane that has no inner data points. The support vectors which give the model its name are the data points closest to the hyperplane and therefore determine the margin (Krishna Samdani 2019).

The SVM also main advantages are prediction accuracy is high, it is less sensitive and flexible even if training example contains errors like neural networks the computational complexity of SVMs does not depend on the dimensionality of the input space.

However, this classifier involves a long training time. It is also difficult to understand the learned function (weights).

The large number of support vectors used from the training set to perform classification task which can cause unbalanced result (El-Naqa and Wernick 2003).

## CHAPTER FOUR

### 4 EXPERIMENTAL RESULT AND DISCUSSION

#### 4.1 Introduction

In this chapter, an experimental evaluation of the proposed models for text-independent Amharic language speaker identification is described in detail.

Experimental evaluation approves the understanding of the proposed model or architecture. The proposed end-to-end CNN and CNN+SVM feature extraction and classification models are assessed.

We conducted an experiment based on the different categories depending on the problems.

The dataset used and the implementation of the proposed model are described carefully. In this method, we have found an accuracy of 82 % and 95 % for the end-to-end CNN and the CNN+SVM model respectively.

#### 4.2 Data Collection and Preparation

We collected spontaneous data set to collect the data set we used primary sources of data gathering techniques using an audio recorder.

In our dataset collection, we considered the Bahir dar city area, which also consists of both males and females. Having such a data set helps us to decide the potential use of speaker identification on varieties of speech samples.

Speakers name are put it as the file name of the voice sample. The format that we follow is speaker name\_number of audio sample like Metadel\_01. All the voice samples are collected according to this format.

In this thesis, we considered a dataset of 50 speakers including 25 male and 25 female speakers. From each 50 speaker class ten (10), 10 seconds duration speech samples are prepared. Totally 500 speech samples are prepared. Then, we perform a sequence of preprocessing steps.

To build our model we used 75% of the total dataset (375 samples) for training data and 25 percent of the total dataset (125 samples) is used for testing data. Out of 75 percent of the training dataset, 25 percent of datasets are used for validation dataset during training.

Each sample is taken at a sampling rate of 16 kHz and 16 bit. After that, all these data are properly preprocessed and the necessary features are extracted.

### **4.3 Development Environment**

Experiments are done using the MATLAB 2019a 64 bit version for prototype development because it is a powerful tool for signal processing.

Intel Core <sup>™</sup> i3-2350M CPU and 8 GB of RAM laptop computer also used to do it. The model is trained based on different parameters for SVM and CNN classifiers; network search is implemented to select the optimal parameter automatically for each classifier.

The overall end-to-end CNN model training is taking 1 hr: 39 min: 25 seconds for 25 epochs with a batch size of 64 and a learning rate of (0.0001) for Adam optimizer.

### **4.4 Evaluation Techniques**

To evaluate our proposed model, we used accuracy, confusion matrix, precision, and recall performance metrics. Since confusion matrix demonstrates how much the model is confused when it makes predictions. In the confusion matrix plot in figures 32, a row represents the predicted (output) class and a column represents the true class. Here, the diagonal cells in the confusion plot represent the observations that are correctly classified whereas the off-diagonal cells represent the observations that are incorrectly classified. In each cell of the confusion matrix, both the number of observations and the percentage of the entire number of observations are shown. The column at the far right of the plot displays the precision that shows the false positive rate and the row at the bottom of the plot displays the recall that shows the false-negative rate. The cell in the bottom right of the plot displays the overall accuracy.

Below is the graph showing the Mini-Batch Accuracy, Mini-Batch Loss, Validation Accuracy, Validation Loss, and confusion matrix result of the trained model.

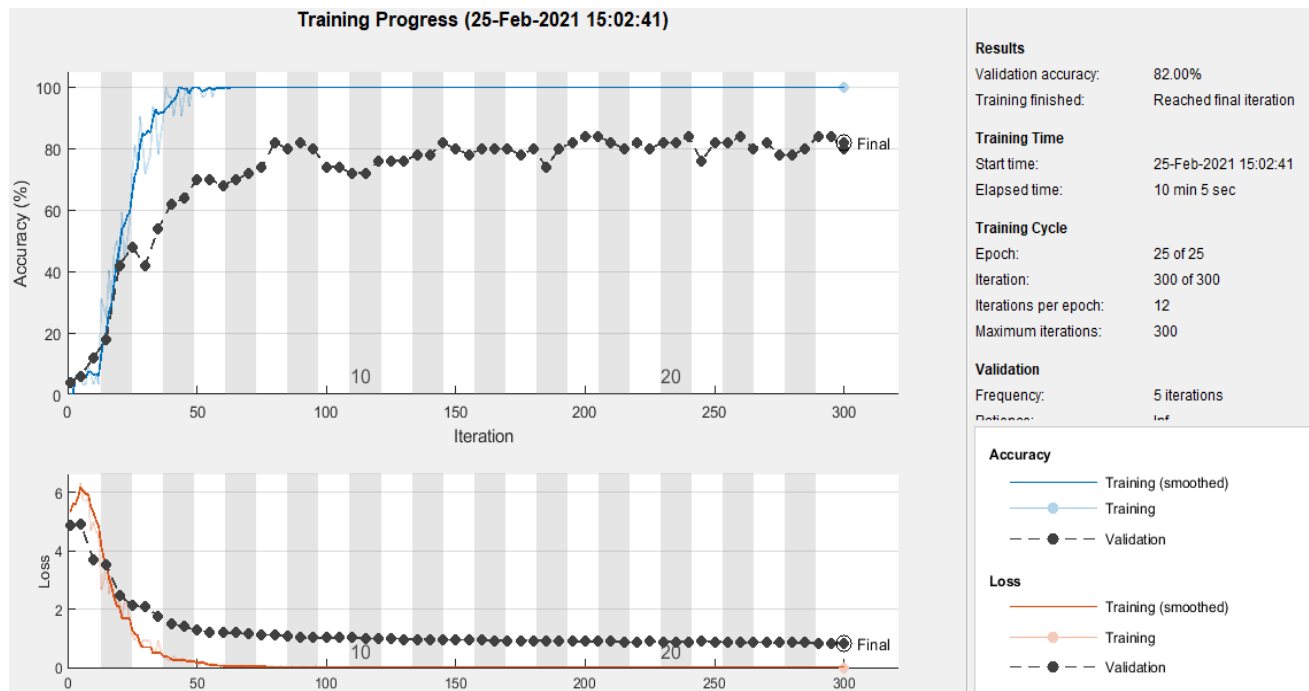


Figure 30 Training Progress of our end-to-end CNN model.

As we can see from the training progress graph in figure 30 above, the Mini-Batch Accuracy is higher than Validation Accuracy in the arc. Associated with this, the gap between the Mini-Batch Accuracy curve and the Validation Accuracy curve is slight. The Mini-Batch Accuracy after epoch number 12 becomes insistent until it stops the training progress. As we can see from the graph both the Mini-Batch loss and Validation Loss primarily look higher starting from epoch number 1 to epoch number three but starting from epoch number 3 to epoch number 10 both the Mini-Batch loss and Validation Loss is decreased gradually. Then, particularly starting from epoch number 11 to epoch number 25 both the Mini-Batch loss and Validation Loss become very close to zero. Also, the gap between the Mini-Batch loss and the Validation Loss curve is slight. So, the graph shows that our model is free from overfitting and underfitting.

| ANALYSIS RESULT |                                                                              |                       |             |                                  |
|-----------------|------------------------------------------------------------------------------|-----------------------|-------------|----------------------------------|
|                 | Name                                                                         | Type                  | Activations | Learnables                       |
| 1               | <b>imageinput</b><br>224x224x1 images with 'zerocenter' normalization        | Image Input           | 224x224x1   | -                                |
| 2               | <b>conv_1</b><br>8 3x3x1 convolutions with stride [1 1] and padding 'same'   | Convolution           | 224x224x8   | Weights 3x3x1x8<br>Bias 1x1x8    |
| 3               | <b>batchnorm_1</b><br>Batch normalization with 8 channels                    | Batch Normalization   | 224x224x8   | Offset 1x1x8<br>Scale 1x1x8      |
| 4               | <b>maxpool_1</b><br>2x2 max pooling with stride [2 2] and padding [0 0 0 0]  | Max Pooling           | 112x112x8   | -                                |
| 5               | <b>relu_1</b><br>ReLU                                                        | ReLU                  | 112x112x8   | -                                |
| 6               | <b>conv_2</b><br>16 3x3x8 convolutions with stride [1 1] and padding 'same'  | Convolution           | 112x112x16  | Weights 3x3x8x16<br>Bias 1x1x16  |
| 7               | <b>batchnorm_2</b><br>Batch normalization with 16 channels                   | Batch Normalization   | 112x112x16  | Offset 1x1x16<br>Scale 1x1x16    |
| 8               | <b>maxpool_2</b><br>2x2 max pooling with stride [2 2] and padding [0 0 0 0]  | Max Pooling           | 56x56x16    | -                                |
| 9               | <b>relu_2</b><br>ReLU                                                        | ReLU                  | 56x56x16    | -                                |
| 10              | <b>conv_3</b><br>32 3x3x16 convolutions with stride [1 1] and padding 'same' | Convolution           | 56x56x32    | Weights 3x3x16x32<br>Bias 1x1x32 |
| 11              | <b>batchnorm_3</b><br>Batch normalization with 32 channels                   | Batch Normalization   | 56x56x32    | Offset 1x1x32<br>Scale 1x1x32    |
| 12              | <b>relu_3</b><br>ReLU                                                        | ReLU                  | 56x56x32    | -                                |
| 13              | <b>fc</b><br>50 fully connected layer                                        | Fully Connected       | 1x1x50      | Weights 50x100352<br>Bias 50x1   |
| 14              | <b>softmax</b><br>softmax                                                    | Softmax               | 1x1x50      | -                                |
| 15              | <b>classoutput</b><br>crossentropyx with 'speaker_01' and 49 other classes   | Classification Output | -           | -                                |

Figure 31 Proposed end-to-end CNN model Summary

## 4.5 Experiment on the proposed end to end CNN model

In this section, we answered research questions, conducted an experiment based on gender group, learning rate, and activation functions using the proposed end-to-end CNN model.

### 4.5.1 Experiment based on Activation function using proposed end-to-end CNN

One of our research questions is “which activation function is suitable for an automatic text-independent Amharic Language speaker identification”, to answer this research question, researchers experimented with Relu and Tanh activation functions. Up to now, in all the experiments researchers used the Relu activation function. In this experiment, the comparison of Tanh and Relu activation functions shows in the table 8.

| Activation function                       | Relu | Tanh  |
|-------------------------------------------|------|-------|
| Accuracy (%)                              | 82   | 63    |
| Training Elapsed time (minutes & seconds) | 9:14 | 10:34 |

Table 8 CNN model on different activation functions

As shown in the table 8, the experiment was conducted for Tanh and Relu activation functions. As the result shows that Tanh activation functions are relatively slow training speed and the least test accuracy relative to Relu.

Based on experimental results, we can conclude that a Relu activation function provides improved test accuracy and training time when compared with Tanh.

Because Tanh activation function standardizes the output of the neuron to the range between a positive one and a negative one.

While the Relu activation function returns the output of the neuron to zero if the input value is less than zero otherwise it returns the input value itself if it is greater than zero. Relu reduces vanishing gradient problem relative to sigmoid and Tanh activation functions (Chigozie Enyinna 2018). The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time. This means that the neurons will only be deactivated if the output of the linear transformation is less than 0.

#### **4.5.2 Experiment based on distinct gender group using proposed end-to-end CNN**

To examine the effect of gender in speaker identification for the Amharic language, researchers conducted an experiment based on both gender groups. A can decide that speaker identification on gender is discriminative or not.

For this experiment, we considered distinct 25 male and 25 female speakers. Up to now, in all experiments, we experimented using one single model for both gender groups of speakers.

In this experiment, researchers compared the speaker identification rate for only male speakers and female speakers. The table 9 shows the comparison result of male and female speakers.

| Gender                                      | Male | Female |
|---------------------------------------------|------|--------|
| Accuracy                                    | 76   | 92     |
| Training Elapsed time (minutes and seconds) | 5:29 | 5:22   |

Table 9 Gender based CNN model

Table 9 shows that the percentage of identification rate of both male and female speakers is far away, which is 76 % for male speakers and 92 % for female speakers. But, the training time of the model for each gender is almost similar.

This is happened due to the different pitch lengths between female speakers' speech (120 - 200 Hz) and male speakers' speech (60 -120 Hz). Pitch is the loudness or lowness of a sound. And again, the intra-subject variability is so large (100–200 Hz for males and 120–350 Hz for females that gender categorization cannot rely on the pitch alone (Belin 2012).

In addition to this, it occurred due to the short-term acoustic features (numerous ms) which define the spectral components of the speech signal.

These speech signal spectral components have a great variability for the male and female speech during extraction via Fast Fourier Transform. Because Fourier transform can capture the discriminating phoneme-like features.

The phonemes are the fundamental building block of the human speech production process. In line with this, researching speaker identification using speakers of the same sex would be a bigger challenge for the model than doing the speaker identification using speakers of mixed female and male speakers. So, we can conclude that text-independent Amharic language speaker identification is gender-dependent and highly discriminative towards gender.

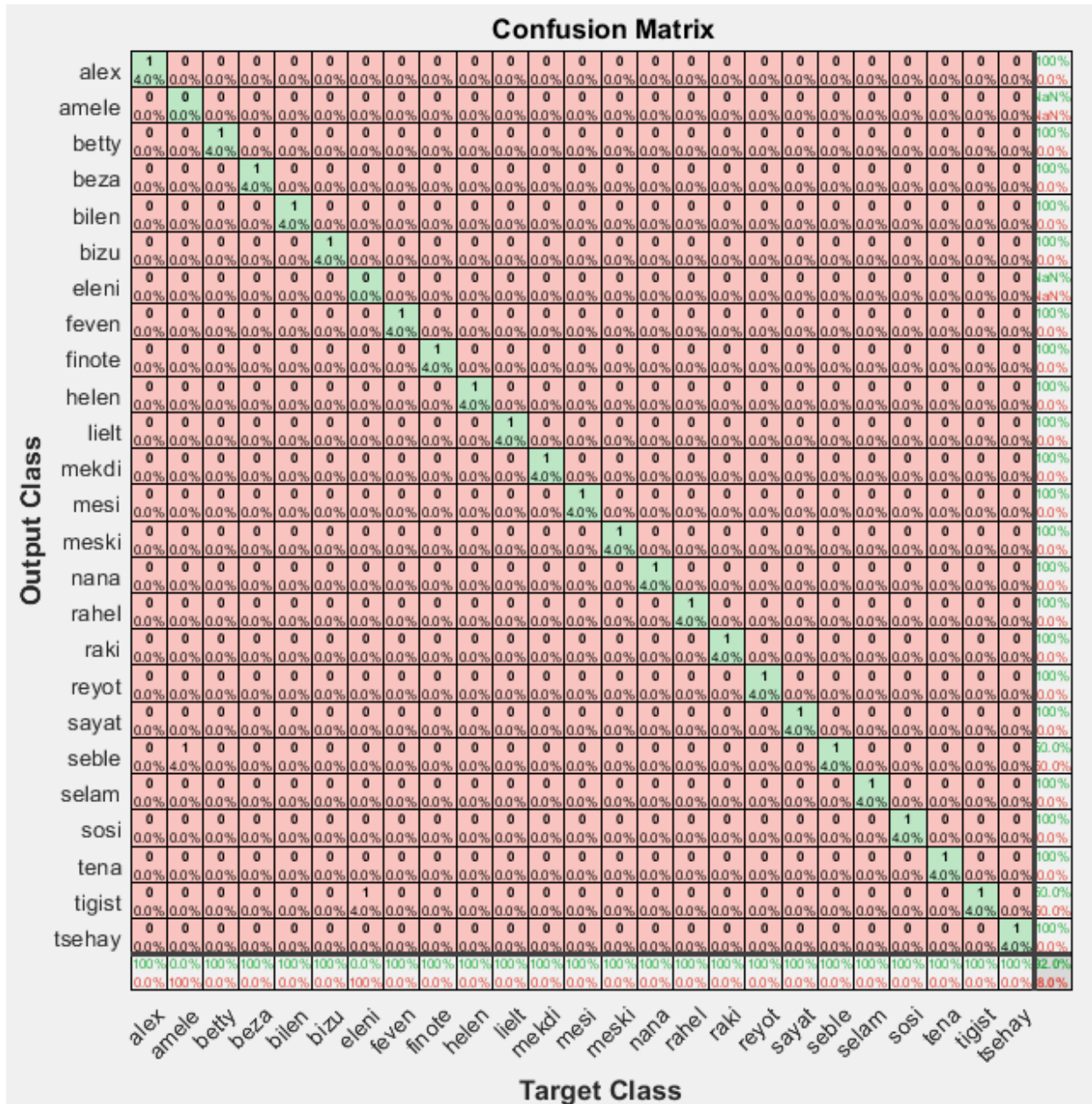


Figure 32 confusion matrix for female modeling

As shown in the figure 32, we have achieved an overall accuracy of 92 % as true positive (correctly classified) and 8 % as true negative (incorrectly classified) for the end-to-end CNN model. Where the labels from Alex to tsehay represent the name of the speaker and each speaker in confusion matrix has 3 samples for testing. From the confusion matrix, all 3 samples of the speaker are correctly predicted without any confusion except amele and eleni.

We used precision and recall for measuring the matrix (true positive, true negative, false positive, and false negative) rate of the proposed model. If a speaker is Alex and the model predicts Alex it is named as true positive. If a speaker is not Alex and the model predicts as not Alex this is named as true negative. If a speaker is not Alex and the model predicts as Alex this is named as false positive. If a speaker is Alex and the model predicts as not Alex this is named as a false negative. Precision displays a false positive rate while recall displays the false-negative rate.

#### 4.5.3 Experiment based on learning rate using proposed end-to-end CNN

To answer the research question “which learning rate is suitable for Amharic Language speaker identification”, we experimented using different learning rate functions. Up to now, in all the experiments we used the 0.0001 learning rate function.

In this experiment, researchers compared with 0.001, 0.0001, and 0.00001 learning rate functions. The table 10 shows a comparison result for the listed learning rates.

|                                              |       |        |         |
|----------------------------------------------|-------|--------|---------|
| Learning rate                                | 0.001 | 0.0001 | 0.00001 |
| Accuracy (%)                                 | 59    | 82     | 61      |
| Training elapsed time ( minutes and seconds) | 9:24  | 9:14   | 8:49    |

Table 10 compares the learning rate

Learning rate determines the weights of the network and updates the weights during the training process. It regulates the speed with which the model learns the weights in the network.

Large learning rate results in a fast learning model that requires fewer epochs but results in sub-optimal weights. On the other hand, a small learning rate makes the model learn slower but requires more epochs to train and results in optimal weights.

Learning rate is the most important parameter to be analyzed. Learning rate also impacts the training speed because larger learning rates result in faster convergence. On the other hand, too large or too small learning rates result in huge training time.

Finding a better value of learning rate for a particular model and a particular dataset is important for the best accuracy of a Neural-Network. That is why we tried to compare it based on its range 0- 1.so the result shows that 0.0001 obtained a better result. The reason it is not very large and also very small is

based on its rage. Chamarty Anusha and P S Avadhani paper also proved that the 0.0001 learning rate is better than others (Avadhani 2019).

#### 4.6 Experiment on proposed CNN-SVM model

To see the effect of training time using CNN and SVM on text-independent Amharic language speaker identification we proposed a combined model of CNN and a multiclass SVM classifier. Here, in this thesis, 50 classes are equivalent to the number of speakers. But, most commonly SVM is state-of-the-art for binary classification or linearly separable problems. In this experiment, we used CNN for feature extraction and multiclass SVM via Error-Correcting Output Codes (ECOC) approach for classification. As an experimental result, we got an accuracy of 82 % and 95 % for the proposed end-to-end CNN model and multiclass classification SVM via ECOC approach respectively. As we can see from the confusion matrix result of end-to-end CNN and CNN-SVM models is efficient SVM enhances the performance of the end-to-end CNN model speaker identification rate by increasing 13 % accuracy. So, using a combined model of CNN and multiclass SVM as feature extraction and classifier reduces the training time of the end-to-end CNN model with a greater number of layers and parameters. SVM with ECOC works well in its default linear learner kernel function.

#### 4.7 Comparison of the proposed model with the previous work

| Author          | Feature extraction algorithm               | Classifier                                 | Data set Amharic speakers | Accuracy |
|-----------------|--------------------------------------------|--------------------------------------------|---------------------------|----------|
| (Abrham 2017)   | Hybrid Approaches of LPCC, MFCC, and GFCC  | VQ, GMM, and a combination of GMM and BPNN | 90                        | 93.7%    |
| (Aykefam 2015)  | MFCC $\Delta$ MFCC and $\Delta\Delta$ MFCC | VQ& GMM                                    | 50                        | 86.8%    |
| Proposed system | CNN                                        | SVM                                        | 50                        | 95%      |

Table 11 Comparison of the proposed model with previous works

## CHAPTER FIVE

### 5 CONCLUSION AND RECOMMENDATION

#### 5.1 Conclusion

Many recent improvements and successes have been done with speech researchers, the challenges of providing actual robust speaker identification on short utterances remain the main considerations when installing automatic speaker recognition, as several real-world applications frequently have access to only limited duration speech data recorded under uncontrolled conditions.

This paper has introduced and evaluated the small data set using an SVM with CNN system-based feature vectors robust text-independent speaker identification.

The proposed system was precisely measured for speaker identification purposes using short-duration utterances for both enrollment and testing tasks obtained from random unrestricted speeches taken over the noise condition.

This proposed technique has focused on the design of a new approach looking for new information able to simplify the identification of speakers with much-reduced speech information.

We prove that this method is appropriate for a realistic speaker recognition application. We do not need to use a huge amount of training dataset as in out-of-date algorithms.

Besides, we don't involve long test utterances to identify the speaker. Also, there is no need to integrate long and complex calculations to handle the conditions having small amounts of speech data.

This is an interesting benefit, especially for realistic applications that need to decrease the computational and time complexity of the system and so the memory size of the system.

Compared with the previous study which implemented the classification using a conventional technique classifier with hand-crafted features, the CNN-SVM combined model could not only automatically extract features using the CNN, but also better improved the generalization ability of CNN and the classification accuracy utilizing combining the SVM.

#### 5.2 Recommendation

This thesis is relevant for different application areas such as for voice-based criminal investigations, Forensics and surveillance, video conferencing, Authentication systems, and for any application which requires the response to the question who said this.

So, the future work that needs to perform is to design a speaker identification model that capable of identifying the speaker with the speaker's mood, high noisy utterances, mimics, health condition, and speaker's session variability.

In the Amharic language, there is no prepared audio data set; collecting pure data set is the main challenge in speaker recognition areas. So, the task that the researcher needs to perform in the future is the preparation of large speech databases that are to be functional in this research area. So, that those who have an interest in the field can do this to save the huge amount of time required to collect those speech utterances.

In recent times, the interest in speech and speaker recognition applications over the mobile phone, and handheld devices has been increased. These devices are almost used in adverse environments such as city streets, airports, offices, and cars. The use of these different means is constantly increasing among private users and business customers.

In addition to the environmental noise in which the speech was produced, telephone communication channels introduce additional distortions to the speech. These different noise sources alter speech production so that most speaker recognition systems are exposed to failure in noise corrupted environments. Therefore in future work performing realistic applications is more valuable for mobile phones.

This thesis will be used as a start point for researchers who are interested to do their work in the area of speaker identification.

## REFERENCE

- Mustaf, Wan Azani, and Mohamed Mydin M Abdul Kader. "A Review of Histogram Equalization Techniques in Image Enhancement Application." *1st International Conference on Green and Sustainable Computing (ICoGeS)*. Perlis, Malaysia: IOP Publishing, 2017. 1-8.
- A. Ahmad, S.H.S. Rizvi, et.al. "Hindi Stop Consonants in Syllable-Final Position." *IETE Journal of Research*, 2015: 162-170 .
- A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks." *NIPS 2012*, 2012: 1106-1114.
- Abebe, Lily. *Example Based Amharic Text to Ethiopian Sign Language Machine Translation*. Addis Ababa, Ethiopia: academia, June 2018.
- Abebe, Lily. "Example Based Amharic Text to Ethiopian Sign Language Machine Translation." *academia*, 2018: 68-73.
- Abrham, D. "Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC, MFCC and GFCC." *IJASCSE*, 2017.
- Acken, C D. Shaver and John M. "The Development of Text-Independent Speaker recognition technology." *Journal of Electrical Engineering*, 2014: 1-8.
- Acken, C D. Shaver and John M. "The Development of Text-Independent Speaker Recognition Technology." *Journal of Electrical Engineering*, 2014: 1-8.
- Acken, C D. Shaver and John M. "The Development of Text-Independent Speaker recognition technology." *Journal of Electrical Engineering*, 2014: 1-8.
- Acken, Clark D. Shaver and John M. "A Brief Review of Speaker Recognition Technology." *Electrical and Computer Engineering Commons*, 2016: 109-116.
- Acken, D. Clark Shaver and J. M. "A Brief Review of Speaker Recognition Technology." *Electrical and Computer Engineering Faculty Publications and Presentations.*, 2016: 1-7.
- AL-Allaf, Omaima N. A. "REVIEW OF FACE DETECTION SYSTEMS BASED ARTIFICIAL NEURAL NETWORKS ALGORITHMS." *The International Journal of Multimedia & Its Applications (IJMA) Vol.6, No.1*, February 2014: 1-16.
- Alku P, Saeidi R. "The linear predictive modeling of speech from higher-lag autocorrelation coefficients applied to noise-robust speaker recognition." *IEEE/ACM Trans Audio Speech Lang process*. 2017. 1606–1617.

- Almabdy, Soad, and Lamiaa Elrefaei. "Deep Convolutional Neural Network-Based Approches for Face Recognition." *MDPI*, October 2019.
- Anderson, Catherine. "Signal processing evaluation techniques." *researchgate*, 2020: 339-345.
- Ankur Mauryaa, Divya Kumara, R.K.Agarwalb. "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach." *ScienceDirect*, 2017.
- Archer, Britta. "synecticsmedia." June 22, 2017. <https://synecticsmedia.com/5-unique-framing-techniques-every-filmmaker-needs-to-know>.
- Asker, Lars. "Experiences with developing language processing tools and corpora for Amharic." *SICS, Swedish Institute of Computer Science* , 2010: 47-54.
- Atal, B.S. "Automatic speaker recognition based on pitch contours." *Journal of Acoustical Society of America*, 1972: 1687-97.
- Avadhani, Chamarty Anusha and P S. "Optimal Accuracy Zone Identification in Object Detection Technique - A Learning Rate Methodology." *International Journal of Engineering and Advanced Technology (IJEAT)*, 2019: 6470-6476.
- Aweem A., Muhammad S,et.al. "Speaker Identification Using a Hybrid CNN-MFCC Approach." *IEEE Explore*, 2020.
- Aweem Ashar, Muhammad Shahid Bhatti,Usama Mushtaq. "Speaker Identification Using a Hybrid CNN-MFCC." *IEEE Xplore*, 2020.
- Aweem Ashar, Muhammad Shahid Bhatti,Usama Mushtaq. "Speaker Identification Using a Hybrid CNN-MFCC Approach." *IEEE*, 2020.
- Aweem Ashar, Muhammad Shahid Bhatti,Usama Mushtaq. "Speaker Identification Using a Hybrid CNN-MFCC Approaches." *IEEE*, 2020.
- Aykefam, Azene. "TEXT-INDEPENDENT SPEAKER IDENTIFICATION FOR THE AMHARIC LANGUAGE." *researchgate* 8, no. 08 (2015).
- Azene, Aykefam. "TEXT-INDEPENDENT SPEAKER IDENTIFICATION FOR THE AMHARIC LANGUAGE." *researchgate*, 2015.
- B. Fan, F. C. Wu, and Z. Y. Hu,. "Rotationally invariant descriptors using intensity order pooling." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, 2012: 2031-2045.

- B. Gambäck, L. Asker. "Experiences with Developing Language Processing Tools and Corpora for Amharic." *SICS, Swedish Institute of Computer Science*, 2018: 47-54.
- B. Singh, V. Rani, et al. "Preprocessing In ASR for Computer Machine Interaction with Humans: A Review." *International Journal of Advanced Research in Computer Science and Software Engineering Research Paper*, 2012: 34-44.
- B. Z. Iko, W. Kozłowski, M. Z. Iko, R. I. Samborski, D. Sierra, J. G. Ika. "Hybrid Wavelet-Fourier-HMM Speaker Recognition." *Department of Electronics, AGH University of Science and Technology Krakow*, 2017.
- B. Z. Iko, W. Kozłowski, M. Z. Iko, R. I. Samborski, D. Sierra, J. G. Ika. "Hybrid Wavelet-Fourier-HMM Speaker Recognition." *Department of Electronics, AGH University of Science and Technology Krakow*, 2016.
- Badura, Stefan. "Feature extraction for automatic lips reading system for isolated vowels Summary." *The 4th International Virtual Scientific Conference on Informatics and Management Sciences*. Itall, Slovakia, 2015.
- Belete, Befkadu. "Audio-Visual Speech Recognition Using Lip Movement for Amharic Language." Addis Ababa, Ethiopia, 2017.
- Belin, Cyril R. Pernet and Pascal. "The role of pitch and timbre in voice gender categorization." *Researchgate*, 2012: 1-11.
- Berhanu, Solomon. *Isolated Amharic Consonant-Vowel (CV) Syllable Recognition. An experiment using the Hidden Markov Model*. Addis Ababa, Ethiopia: unpublished, 2001.
- Birara, Muleken, and Gebeyehu Belay. "Automatic word recognition based on lip motion for amharic speech: a computer vision approach(Unpublished thesis)." *Bahir Dar university*, 2017.
- Bishop, C. "Pattern Recognition and Machine Learning." *s.l Springer Science Business Media*, 2006: 435-441.
- Bishop, C.M. "Neural Networks for pattern recognition." *Oxford University press*, 1995.
- Björn, G. "Experiences with developing language processing tools and corpora for Amharic." *SICS, Swedish Institute of Computer Science*, 2010: 47-54.
- Blackledge, Jonathan M. "The Convolution Theorem." *science direct*, 2005: 108-112.

- Boser, B.E., I.M. Guyon, and V.N Vapnik. "A training algorithm for optimal margin classifiers." *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh, PA, USA: pp, July 1992. 27–29.
- Boureau, Y.-L., J. Ponce, and Y. A LeCun. "Theoretical analysis of feature pooling in visual recognition." *In Proceedings of the 27th International Conference on Machine Learning*. Haifa, Israel: pp, 21–24 June. 111–118.
- Breiman, Leo. "RANDOM FORESTS." *University of California*, 2001.
- Bricker, P.D. "Statistical techniques for talker identification." *Bell systems technical*, 1971: 1427-1454.
- Britannica, The Editors of Encyclopaedia. "whiper speech." <https://www.britannica.com/topic/pitch-speech>, 2020.
- C. Wang, D. Xu, C.P. Jose. "Speaker verification and identification using gamma neural networks." *In international conference on neural networks*, 1997.
- C.C.T. Chen, C.T. Chen, P.W. Cheng. "Hybrid KLT/GMM approach for robust speaker identification." *Electronics Letters IEE*, 2003: 1552-1553.
- Çetingül, H. Ertan. "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading." *IEEE*, 2006: 2879-2891.
- Chakroun R, Frikha M, Zouari LB. "New approach for short utterance speaker identification." *IET Signal Process*, 2018.
- Chakroun R, Frikha M, Zouari LB. "New approach for short utterance speaker identification." *IET Signal Process*, 2018.
- Chakroun R, Frikha M, Zouari LB. "New approach for short uttrance speaker identification." *IET Signal Process*, 2018: 873–880.
- Chang J, Wang D. "Robust speaker recognition based on DNN/i-vectors and speech separation." *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017.
- Chang J, Wang D. "Robust speaker recognition based on DNN/i-vectors and speech separation." *IEEE*, 2017: 5-9.
- Chang J, Wang D. "Robust speaker recognition based on DNN/i-vectors and speech separation." *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017: 5-9.
- Chen MC, Ball RL, Yang L. "Deep learning to classify radiology free-text reports." *RSNA*, 2018: 820-829.

- Chen, Z., and S. K. Sun. "A Zernike moment phase-based descriptor for local image representation and matching." *IEEE Transactions on Image Processing*, vol. 19, no. 1, Jan. 2010: 205-219.
- Cheung, R.S. "Feature selection via dynamic programming for text independent." *IEEE Transactions on Acoustics*, 1978: 397-403.
- Cheung, R.S. "Feature selection via dynamic programming for text independent speaker identification." *IEEE Transactions on Acoustics, Speech and Signal*, 1978: 397-403.
- Chigozie Enyinna, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning." *arXiv:1811.03378v1*, 2018: 1411-1421.
- Choras, Ryszard S. "Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems." *INTERNATIONAL JOURNAL OF BIOLOGY AND BIOMEDICAL ENGINEERING*, 2007: 6-16.
- Clark K, Vendt B, Smith K. "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *IEEE*, 2017: 2097-2106.
- Columbus, Louis. *The State of Physical Access Control in EMEA Businesses – 2020 Report*. Worcester England: IFSEC GLOBAL, 2020.
- Comp.speech*. n.d. <http://svr-www.eng.cam.ac.uk/comp.speech>.
- COWELL, JOHN, and FIAZ HUSSAIN. "Amharic Character Recognition using a Fast Signature Based Algorithm." *Proceedings of the Seventh International Conference on Information Visualization (IV'03)*. IEEE COMPUTER SOCIETY, 2003.
- D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny and E. Lleida. "Unscented Transform for i-vector-based Noisy Speaker Recognition." *IEE International Conference on Acoustics, Speech and Signal Processing*. 2014. 4042–4046.
- D. Ververidis, C. Kotropoulos. "Gaussian mixture modeling by exploiting the mahalanobis distance." *IEEE transactions on signal processing*, 2008: 2797-2811.
- D.A. Reynolds, R.C. Rose. "Robust text-independent speaker identification using Gaussian Mixture speaker models." *IEEE Trans. on Speech and Audio Processing*, 1995: 72-83.
- davidp.b. 2019. <http://www.davidpbrown.co.uk/help/top-100-languages-by-population.html>.

- Debasu, Abrham. "Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC, MFCC and GFCC." *IJASCSE*, 2017.
- Degu, Yeshiwas. "The Quest for identity and autonomy in ethiopia." *addis ababa*, 2014: 1-7.
- Deshpande, Raghunath S. HolambeMangesh S. "There has long been interested to be able to identify a person based on his or her voice. For numerous years, attorneys, judges, investigators, and law execution agencies have wanted to use forensic voice verification to investigate a suspect or to check ." *SpringerBriefs in Electrical and Computer Engineering*, 2012.
- Doddington, G.R. "A new method of speaker verification." *Journal of acoustical*, 1971: 139.
- Dragos Burileanu1, Lucian Pascalini1, Corneliu Burileanu1 and Mihai Puchiu. "An Adaptive and Fast Speech Detection Algorithm." *Proceedings of the Third International Work-shop on Text, Speech and Dialogue*, 2000: 177 - 182.
- El-Naqa, Y.Y., and et.al Wernick. *A Practical Guide to Support Vector Classification*. Taipei: Department of Computer Science and Information Engineering, National Taiwan University, 2003.
- EthiopianTreasures.c.uk. *Ethiopian Treasures*. january 19, 2002.  
<http://www.ethiopiantreasures.co.uk/pages/language.htm> (accessed january 21, 2002-2021).
- F. Bimbot, J. Bonastre, et.al. "A Tutorial on Text-Independent speaker verification ." *EURASIP Journal on Applied Signal Processing*, 2004: 431-450.
- F.K. Soong, et al. "A vector quantization approach to speaker recognition." *AT &T Technical Journal*, 1987: 14-26.
- Fantaye, Tessfu Geteye. "CHARACTERIZATION OF AMHARIC VOWEL SOUND UNITS." *AAU*, 2011: 36–40.
- Feng, Ling. "Speaker Recognition." *Kgs. Lyngby*, 2004.
- Fenglei, WANG Bingxi. "Text-independent speaker recognition using support vector machine." *IEEE*, 2017: 1-4.
- Fiaz Hussain, Cowell John. "Amharic Character Recognition using a Fast Signature Based Algorithm." *IEEE*, 2013: 18-27.
- Finch, G. "PHONETICS AND PHONOLOGY." *Springer*, 2017: 33-76.
- Flanagan, James L. "The Mechanism of Speech Production." *springer*, 2016: 9-23.

- Frew, Befkadu Belete. "Audio-Visual Speech Recognition using LIP Movement for Amharic Language." *International Journal of Engineering Research & Technology (IJERT)* 8, no. 08 (2019): 594-604.
- Furui, S. "50 years of Progress in Speech and Speaker Recognition." *Department of Computer Science, Tokyo Institute of Technology*, 2004: 1-9.
- Furui, S. "Vector quantization based speech recognition and speaker recognition techniques." *25th Asilmar conference signals, systems and computers*, 1991.
- Furui, S.I. "Cepstral analysis technique for automatic speaker verification." *IEEE Transactions on Acoustics Speech and Signal Processing*, 1981: 254-72.
- G. E. Hinton, S. Osindero, and Y.-W. Teh. "A fast learning algorithm for deep belief nets." *Neural Computation*, 2006: 1527–1554.
- G. Kolano, P. Regel-Brietzmann. "Combination of vector quantization and Gaussian mixture models for speaker verification with sparse training data." *In EUROSPEECH*, 1999: 1203-1206.
- G. Saha, Sandipan Chakroborty, Suman Senapat. "A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications." *Proceedings of the NCC*, 2005.
- G. Saha, Sandipan, et.al. "Silence Removal and Endpoint detection algorithms for speech and speaker recognition." *Indian Institute of Technology, Kharagpur, Kharagpur*, 2014.
- G. Ye, B. Mak and M. Mak. "Fast GMM Computation for Speaker Verification using Scalar Quantization and Discrete Densities." *In Proc. Int. Speech Commun. Association (Interspeech)*, 2009: 2327–2330.
- Ganapathiraju, Aravind. "Support Vector Machines for Speech Recognition." *Mississippi*, 2002.
- Ganapathy S, Mallidi SH, Hermansky H. "Robust feature extraction using modulation filtering of autoregressive models." *IEEE Trans Audio Speech Lang Process*, 2014.
- Ghitu, Alin, and Leon J.M. "Visual Speech Recognition automatic Sysytem for Lip Reading of Dutch." *Information technologies and control*, 2009.
- Gökay Dişken, Zekeriya Tüfekçi, Lütfü Saribulut & Ulus Çevik. "A Review on Feature Extraction for Speaker Recognition under Degraded Conditions." *IETE Technical Review*, 2016.
- Goodfellow, Ian. "Deep learning." In *Convolutional Networks*, by Jeff Heaton, 1-27. Montréal: The MIT Press, 2016.

- H s jyanna, S R MAHADEVA PRASANNA. "Limited data speaker identification." *Indian Academy of Sciences*, 2010.
- H. Jialong, L. Liu, P. Gunther. "A new codebook training algorithm For VQ-based speaker recognition." *IEEE international conference on acoustics, speech and signal processing*, 1997: 1091-1094.
- Han, Dianyuan. "Comparison of Commonly Used Image Interpolation Methods." *ICCSEE*, 2013: 1556-1559.
- Hansen, Keith W. Godin & John H. L. "Physical task stress and speaker variability in voice quality." *springer*, 2015: 1-5.
- Hasan T, Saeidi R, Hansen JHL, van Leeuwen DA. "Duration mismatch compensation for i-vector based speaker system." *Proceedings of IEEE ICASSP*, 2013.
- Hassanat, Ahmad B. A. "Visual Speech Recognition." *INTECH*, June 2011: 280-302.
- Hassanat, Ahmad. "Visual Passwords Using Automatic Lip Reading." *IJSBAR*, 2014: 218-231.
- Haykin, Simon. *Neural Networks a Comprehensive Foundation, 2nd edition*. Hamilton, Ontario, Canada: McMaster University, n.d.
- He, K., X. Zhang, S. Ren, and J Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition ." *IEEE Trans. Pattern Anal. Mach. Intell*, 2015: 1904–1916.
- Hedayati, Ali. "An analysis of identity theft: Motives, related frauds, techniques and prevention." 2011.
- Heigold G, Moreno I, Bengio S, Shazeer N. "End-to-end text-dependent speaker verification." *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2016. 5115–5119.
- Hinton GE, Srivastava N. "Improving neural networks by preventing co-adaptation of feature detectors." 2012.
- Hinton GE, Srivastava N. "Improving neural networks by preventing co-adaptation of feature detectors." *Computer Science.ArXiv*, 2012: 237-242.
- Hugo, Daniel, and Campora P ´ erez. "A practical approach to Convolutional Neural network." *Inverted CERN School of Computing*. Sevilla, 2019.
- Hui-Ling Lu, Julius O. Smith. "Glottal source modeling for singing voice synthesis." *Center for Computer Research in Music and Acoustics*, 2012.

- Hurmalainen A, Saeidi R, Virtanen T. "Noise robust speaker recognition with convolutive sparse coding." *sigtheenth annual conference of the international speech communication association*. 2015.
- Ioffe S, Szegedy C. "Batch normalization: accelerating deep network training by reducing internal covariate shift." 2015.
- J. G. Proakis, D. G. Manolakis. "Digital signal processing. Principles, Algorithms and application." *Third ed. Macmillan, New York*, 1996.
- J. Harrington, S. Cassidy. "Techniques in Speech Acoustics." *Kluwer Academic*, 1999.
- J. Meribah Jasmine, S. Sandhya , Dr. K. Ravichandran , Dr. D. Balasubramaniam. "Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameter." *IJIRCCE*, 2016.
- J. P. Campbell Jr. "Speaker recognition: a tutorial." *Proceedings of the IEEE*, 2016: 1437-1462.
- J.M, Alin G and Leon. "Visual speech recognition: automatic system for Lip reading of Dutch." *collaborative information system ICIS project supported by the Dutch Ministry of Economic Affairs*, Mar 2009.
- Jakobson. "Natural classes and distinctive." *The Emergence of Distinctive Features*, 2014: 87-102.
- Jin, Q. "Robust Speaker Recognition." *PhD Thesis, Language Technologies Institute School of Computer Science Carnegie Mellon University, Pittsburgh*, 2007: 23-177.
- Joe Wolfe, Maëva Garnier, and John Smith. "Vocal tract resonances in speech, singing, and playing musical instruments." *Researchgate*, 2008: 6-23.
- Joel Trussell, Anuj Srivastava, Amit K. "Speech Production Modeling and Analysis." *science direct*, 2014: 985-1018.
- JOHN COWELL, FIAZ HUSSAIN. "Amharic Character Recognition using a Fast Signature Based Algorithm." *IEEE COMPUTER SOCIETY*, 2003.
- John M, S. Furui, Q. Jin, D. A. Reynolds. "The Development of Text-Independent Speaker Recognition Technology." *Journal of Electrical Engineering*, 2014: 1-8.
- Jones, M. J. "Why a lowered larynx is good for human speech production." *University of Cambridge & University of York*, 2017.
- Joshua. *joshua project*. June 23, 2019. <https://joshuaproject.net/languages/amh> (accessed january 19, 2020).

- Ju, A.M. "Could Biometrics Make Skies Safer?" *In PC World*, 2002.
- Juang, L. R. Rabiner and B.-H. "Fundamentals of speech recognition." *PTR Prentice Hall Englewood Cliffs*, 2017: 9-40.
- Jyoti B. Ramgire, Prof. Sumati M.Jagdale. "A Survey on Speaker Recognition With Various Feature Extraction And Classification Techniques." *International Research Journal of Engineering and Technology (IRJET)*, 2016: 139-151.
- K. Delac, M. Grgic, et.al. "A Survey Of Biometric Recognition Methods." *International Symposium on Electronics in Marine*, 2004.
- K. Prasad, P. Lotia and M. R. Khan. "A Review on Text- Independent Speaker Identification Using Gaussian Super vector SVM." *International Journal of u-and e- Service, Science and Technology*, 2012.
- Kandagal, Amaresh P, and V. Udayashankara. "Visual Speech Recognition Based on Lip Movement for Indian Languages." *International Journal of Computational Intelligence Research*, 2017: 1-14.
- Karatzoglou A, Gentile C, Li S. "Collaborative filtering bandits." *In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval.* ACM, 2016. 539–548.
- Karpov, E. *remember the poorest community*. 2018. <http://www.rpcafrica.com/>.
- Kasymbek. A, Ozhikenov, et.al. "Pre-processing Voice Signals for Voice Recognition Systems." *18th INTERNATIONAL CONFERENCE ON MICRO/NANOTECHNOLOGIES AND ELECTRON DEVICES EDM*. Erlagol: researchgate, 2017. 2325-2334.
- Kelbesa, I. Y. "An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System." *Universit à degli Studi di Brescia*, 2014: 1-146.
- Kelbesa, Ittansa Yonas. "An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System." 2014.
- Kelbesa, Ittansa Yonas. "An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System." *ACCADAMIA*, 2014: 1-146.
- Kheder WB, Matrouf D, Bousquet PM, Bonastre JF, Ajili M. "Fast i-vector denoising using MAP estimation and a noisedistributions database for robust speaker recognition." *comput Speech Lang*, 2017.

- Khosravani A, Homayounpour MM. "Nonparametrically trained PLDA for short duration i-vector speaker verification." *Comput Speech Lang*, 2018: 105-122.
- Kinnunen, T. "Spectral Features for Automatic Text-independent Speaker." *University of Joensuu, Department of Computer Science*, 2003.
- Kinnunen, T. "Spectral Features for Automatic Text-Independent Speaker Recognition." *Licentiate's Thesis, University of Joensuu, Department of Computer Science Joensuu Finland*, 2003: 1-151.
- Král, P. "Discrete Wavelet Transform for automatic speaker recognition, in Image and Signal Processing." *3rd International Congress on*, 2010: 3514-3518.
- Krishna Samdani, Vats, Prakhar. "Study on Machine Learning Techniques In Financial Markets." *IEEE*, 2019: 983-995.
- Krishnamoorthy P, Jayanna HS, Prasanna SM. "Speaker recognition under limited data condition by noise addition." *Expert Syst Appl*, 2011.
- Krishnamoorthy P, Jayanna HS, Prasanna SM. "Speaker recognition under limited data condition by noise addition." *Expert Syst Appl* 38(10), 2015: 525–546.
- Kriti, et.al. "PNN-based classification of retinal diseases using fundus images." *sciencedirect*, 2019: 215-242.
- Krizhevsky, A, and G.E Hinton. "Classification in pattern Recognition." *Advances in neural information processing systems*, 2012: 1-9.
- Krizhevsky, A., I. Sutskever, and G.E Hinton. "Imagenet classification with deep convolutional neural networks." *In Proceedings of the Advances in Neural Information Processing Systems*. Lake Tahoe, CA, USA: pp, 3–6 December 2012. 1097–1105.
- Kushagra Singh, Nachiketa Tiwari. "Identification of Hindi Stop Consonants in Syllable-Final Position." *IETE Journal of Research*, 2016: 162-170.
- Kushagra, Singh. "Feature selection via dynamic programming for text independent speaker identification." *sciencedirect*, 2016: 397 - 403.
- L. Baum, T. Petrie, G. Soules, N. Weiss. "A Maximization Technique in the Statistical Analysis of Probabilistic Functions of Markov Chains." *Annals of Mathematical Statistics*, 1970.
- L. P. Cordella, P. Foggia, C. Sansone, M. Vento. "A Real-Time Text-Independent Speaker Identification System." *Proceedings of the ICIAP*, 2003: 632.

- L. Siegel, A. Bessey. "Voiced/Unvoiced/Mixed excitation classification of speech." *IEEE Transactions on Signal Processing*, 2009: 451 - 460.
- L. Zhang, F. Yang, Y. Daniel Zhang, and Y. J. Zhu. "Road crack detection using deep convolutional neural network." in *Proceedings of the IEEE International Conference on Image Processing (ICIP '16)*. USA: Phoenix, Ariz, 2016. 3708–3712.
- Larson, C. R. "How the Larynx (Voice Box) Works." *Departments of Communication Science Disorders and Neurobiology and Physiology Northwestern University.*, 2010.
- LeCun Y, Bengio Y, Hinton G. "Deep learning nature." *nature*, 2015: 436–444.
- Lei Y, Burget L, Scheffer N. "A noise robust i-vector extractor using vector Taylor series for speaker recognition." *Proceedings of the 38th ICASSP, Vancouver, BC, Canada*, 2013. 6788–6791.
- Lieberman, Blumstein. *speech production* . 1988. <http://www.haskins.yale.edu>.
- Lienhart, R., and J. Maydt. "An extended set of haar-like features for rapid object detection." *International Conference on Image Processing (ICIP2002)*. Rochester, New York, USA, 2002. 900-903.
- LinM, Chen Q, Yan S. "Network in network." *Researchgate*, 2018: 324-336.
- Liu JC, Leu FY, Lin GL, Susanto H. "An MFCC-based text independent speaker identification system for access control." *Concur Comput Pract Exp* 30(2):e4255, 2018: 32187 - 32202.
- Liu X, Sadeghian R, Zahorian SA. "A modulation feature set for robust automatic speech recognition in additive noise and reverberation." *IEEE international conference on acoustics speech and signal processing (ICASSP)*. IEEE. 2017. 5230–5234.
- Liu Z, Wu Z, Li T, Li J, Shen C. "GMM and CNN hybrid method for short utterance speaker recognition." *IEEE Trans Ind Inf*, 2018.
- Liu, L., L. Shao, and K. Lu, X. L. Li. "Learning spatio-temporal representations for action recognition: A genetic programming approach." *IEEE Transactions on Cybernetics*, vol. 46, no. 1, 2016: 158-170.
- Longkumer, Nungsanginla, Mukesh Kuma, and Rohini Saxena. "Contrast Enhancement Techniques using Histogram Equalization: A Survey." *International Journal of Current Engineering and Technology*, june 2014.
- Lorentzon, M., and M. Wallenberg. "Feature extraction for image selection using machine learning Feature extraction for image selection using machine learning." *IEEE*, 2017.

- Lowe, D. G. "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, vol. 60, no. 2,, 2004: 91-110.
- Lu, Yuanyao, and Hongbo Li. "Automatic Lip-Reading System Based on Deep Convolutional Natural Network and Attention Based Long Short-Term Memory." *MDPI applied science* , 2019: 1-12.
- Lu, Yuanyao, and Hongbo Li. "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory ." *applied sciences*, 2019.
- Luo, Shi-Huang Chen and Yu-Ren. "Speaker Verification Using MFCC and Support Vector Machine." *International MultiConference of Engineers and Computer Scientists*. Hong Kong, 2009.
- Lutz, Richard J. DavidsonAntoine. "Forensic Automatic Speaker Recognition Systems." *IEEE Signal Processing Magazine*, 2015.
- M. Breen, L C. Dilley, J. Kraemer, and E. Gibson. "'Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)", *Corpus linguist. ling.*" 277-312: 2012.
- M. Ito, R. Donaldson. "Zero-crossing measurements for analysis and recognition of speech sounds." *IEEE Transactions on Audio and Electroacoustics*, 1971: 235 - 242.
- M.M. Homayounpour, G. Chollet. "A comparison of some relevant parametric representation for speaker verification." *In ESCA Workshop on automatic speaker recognition,identification and verification* , n.d.
- M.S. Zilovic, R.P. Ramachandran, R.J. Mammone. "The use of robust cepstral features obtained from pole-zero transfer functions for speaker identification." *In Canadian conference on electrical and computer engineering*, 1995.
- Madhuri Yadav, Alok Kumar. "Character Recognition using Convolutional Neural Network." *International Journal for Research in Applied Science & Engineering Technology(IJRASET)*, 2018: 2403-2409.
- Markowitz, J. "Using speech recognition in customer relationship management to be more effective." *In DCI customer relationship management conference*, 1999.
- Markowitz, Judith. "Voice Biometrics." *researchgate* (researchgate), 2019: 1-17.
- Md. Atiqul Islam, Wissam A. Jassim, Ng Siew Cheok, Muhammad Shamsul. "A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery." 2016.

- Md. Atiqul Islam, Wissam A. Jassim, et al. "A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery." *journal.pone.0158520*, 2016.
- Mesbah, Abderrahim, Hicham Hammouchi, Aissam Berrahou, and Hassan Berbia. "Lip Reading with Hahn Convolutional Neural Networks moments." *HAL archives-ouvertes.fr*, 24 Apr 2019: 1-28.
- Meuwly, D. "Forensic Individualization from Biometric Data." *Science & Justice*, 2006: 205 – 213.
- Mirco Ravanelli, Yoshua Bengio. "SPEAKER RECOGNITION FROM RAW WAVEFORM WITH SINCNET." *EESS*, 2019.
- Mohana, Manjunath Jogin, Madhulika M S. "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning." *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology*. Bermingham: IEEE, 2018. 2319-2323.
- Mohanaiah, P, P Sathyanarayana, and L GuruKumar. "Image Texture Feature Extraction Using GLCM Approach." *International Journal of Scientific and Research Publications, Volume 3, Issue 5*, May 2013 : 1-5.
- Mohans, W.S. "Statistical feature evaluation in speaker identification." *department of electrical engineering*, 1969.
- Mokgonyane, Tshephisho Joseph Sefara, et al. "Automatic Speaker Recognition System based on Machine Learning Algorithms." *SAUPEC/RobMech/PRASA Conference*. Bloemfontein, South Africa: ResearchGate, 2019. 7.
- Mondher F., Rania C. "Robust features for text-independent speaker recognition with short." *Springer-Verlag London Ltd., part of Springer Nature 2020*, 2020.
- Mukherjee, Rishiraj. *Speaker Recognition Using Shifted MFCC*. Florida : scholar commons, 2012.
- Nakasone, H. "Automated speaker recognition in real world conditions Controlling the uncontrollable." *8th European Conference on Speech Communication and technology*, 2003: 697-700.
- Nakasone, H. "Automated speaker recognition in real world conditions: Controlling the uncontrollable." *8th European Conference on Speech Communication and Technology*, 2003: 697-700.
- Narayanan, S. "Acoustic Theory of Speech Production." 2015.
- Nilu Singh, Alka Agrawal and R. A. Khan. "THE DEVELOPMENT OF SPEAKER RECOGNITION TECHNOLOGY." *researchgate* , 2018: 8–16.

- Nishani, Eralda, and Bitm Cico. "Computer vision approaches based on Deep Learning and Neural Networks." *MEDITERRANEAN CONFERENCE ON EMBEDDED COMPUTING*. BAR,MONTENEGRO, June 2017. 1-4.
- O. P. Prabhakar, N. K. Sahu. "Performance Improvement of Human Voice Recognition System using Gaussian Mixture Model." *International Journal of Advanced Research in Computer and Communication Engineering*, 2014.
- Osuna, E., Freund, R., Girosi, F. "Improved Training Algorithm for Support Vector Machines." *Proc. IEEE*, 1997.
- P.B. Denes, E. N. Pinson. "The speech chain: The physics and biology of spoken language, 2nd ed. New York." *W.H. Freeman and Company*, 1998.
- Pan, W., K. Qin, and and Y. Chen. "An adaptable-multilayer fractional fourier transform approach for image registration." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, 2009: 400-414.
- Parashar Dhakal, et.al. "A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface." *MDPI*, 2019.
- Park SH, Han K. "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology formedical diagnosis and prediction." *NLOM*, 2018: 800-809.
- Pengqi Li, Guanyu Li. "Channel Mismatch Speaker Verification Based on Deep Learning and PLDA ." 2020.
- Philomina S., Uma V. "Deep Learning based Feature Extraction for Texture Classification." *sciencedirect*, 2020.
- Philomina Simon, Uma V. "Deep Learning based Feature Extraction for Texture Classification." *ScienceDirect*, 2020: 1680-1687.
- Pillay, S. "Voice Biometrics under Mismatched Noise Conditions." *University of Hertfordshire*, 2010: 19-26.
- Pinson, P.B. Denes and E. N. *The speech chain: The physics and biology of spoken language 2nd edition*. New York: W.H. Freeman and Company, 1998.
- Pinson, P.B. Denes and E. N. "The speech chain: The physics and biology of spoken language, 2nd ed." *New York: W.H. Freeman and Company*, 1998.

- Prasanna, H. S. Jayanna and S. R. Mahadeva. "Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition." *IETE TECHNICAL REVIEW*, 2014: 37-41.
- Prasanna, H. S. Jayanna and S. R. Mahadeva. "Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition." *IETE TECHNICAL REVIEW*, 2014: 37-41.
- Prateek Sangwan, Shamsher Malik, Deepak Sharma. "Speaker Recognition System: A Review." (*IJARESM*), 2015: 7.
- Praveen P S, Thimmaraja Yadava G, et.al. "Text Independent Speaker Identification: A Review." *Signal Processing Magazine, IEEE*, 2017: 1-6.
- Pullella, R. Togneri and D. "An overview of speaker identification: Accuracy and robustness issues." *Circuits and Systems Magazine, IEEE*, 2011: 23-61.
- Quatieri, T.F. "Discrete-time speech signal processing: principles and practice ser. Prentice Hall Signal Processing Series." *Prentice Hall*, 2002.
- Quatieri, T.F. "Discrete-time speech signal processing: principles and practice, ser. Prentice Hall Signal Processing Series, A.V. Oppenheim, Ed. Upper Saddle River." *Prentice Hall*, 2002.
- R. Brunelli, D. Falavigna. "Person identification using multiple cues." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995: 955-966.
- R. Price, J. Willmore, W. Roberts. "Genetically Optimized Feed forward Neural Networks for Speaker Identification." *Information Technology Division, Electronics and Surveillance Research Laboratory*, 2003.
- R. Sarikaya, B. Pellom, and J.H.L Hansen. "Wavelet Packet Transform Features with Application to Speaker Identification." *IEEE Nordic Signal Processing Symposium*, 81–84: 1998.
- Radoslaw Weychan, Tomasz Marciniak, et.al. "Speaker recognition based on short polish sequences." 2010.
- Ramachandran P, Zoph B, Le QV. "Searching for activation function." *Researchgate*, 2018: 60-77.
- Rania Chakroun, Mondher Frikha. "Robust features for text-independent speaker recognition with short utterances." *springer*, 2019: 1-21.
- Rashmi .H\*, Suganya S†. "Survey on Image Scaling With Edge Detection and Interpolation methods ." *IJISSET - International Journal of Innovative Science, Engineering & Technology*, April 2015: 345-347.

- Reynolds D, Quatieri T, Dunn R. "Speaker verification using adapted Gaussian mixture models." *Digit Signal Process*, 2014.
- Reynolds, D. A. "Comparison of background normalization methods for text-independent speaker verification." in *Proc. Euro speech*, 1997: 963–966.
- Reynolds, D. A. "Comparison of background normalization methods for text-independent speaker verification." in *Proc. Euro speech*, 1997: 963–966.
- Reynolds, D. A. "Comparison of background normalization methods for text-independent speaker verification." in *Proc. Euro speech*, 1997: 963–966.
- Reynolds, D. "An overview of automatic speaker recognition." in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002: 4072-4075.
- Reynolds, D.A. "A Gaussian Mixture Modelling Approach to Text-Independent Speaker Identification." *Georgia Institute of technology, Atlanta, GA*, 1992.
- Reynolds, D.A. "A Gaussian Mixture Modelling Approach to Text-Independent Speaker Identification." *PhD Thesis, Georgia Institute of technology, Atlanta, GA*, 1992.
- Reynolds, D.A. "Speaker identification and verification using Gaussian mixture speaker models." *Speech Communication*, 1995: 91-108.
- Ripley, B.D. "Pattern recognition and neural networks." *New York: Cambridge University Press*, 1996.
- Roach, P. "English Phonetics and Phonology." *A Practical Course," Fourth Edition, Cambridge University Press*, 2015.
- Roberto Togneri, Daniel Pullella. "An Overview of speaker identification:accuracy and robustness issue." *IEEE CIRCUITS AND SYSTEMS MAGAZINE*, 2011: 23-61.
- Rodman, V. Fromkin and R. "The Speech Chain." *researchgate*, 2013: 1-9.
- Rodríguez, D. S. "Text-Independent Speaker Identification." *AGH University of Science andTechnology*, 2008.
- Rodríguez, D. S. "Text-Independent Speaker Identification." *AGH University of Science andTechnology*, 2008.
- Rodríguez, D. S. "Text-Independent Speaker Identification." *AGH University of Science andTechnology*, 2019.
- Rodríguez, David Sierra. "Text-Independent Speaker Identification." *Kraków*, 2008.

- Rosca, Justinian. "Cepstrum-like ICA representations for text independent speaker recognition." *ICA*, 2003: 999-1004.
- Rose, P. "Forensic speaker identification." *ser. Taylor & Francis forensic science series*, 2002.: 376-384.
- Rose, Philip. "NIST Speaker Recognition Evaluation." [http://nist.gov/itl/iad/mig/upload/registration\\_sre12-v0.pdf](http://nist.gov/itl/iad/mig/upload/registration_sre12-v0.pdf). May 30, 2012. <http://nist.gov>.
- Rosenberg, F. Soong and A. "On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition." *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1998: 871–879.
- Ruder, S. "An overview of gradient descent optimization algorithms." *Researchgate*, 2016: 14-27.
- Ruili Wang, Seyed Reza Shahamiri, et.al. "Speaker identification features extraction methods: A systematic review." *ScienceDirect*, 2017: 250-271.
- S G Bagul, R.K.Shastr. "TEXT INDEPENDENT SPEAKER RECOGNITION SYSTEM USING GMM." *IEEE*, 2013: 245-254.
- S. Bengio, J. Mariethoz. "Learning the decision function for speaker verification." *In IEEE International conference on Acoustics, Speech and Signal Processing*, 2001.
- S. M. Kamruzzaman, A. N. M. Rezaul Karim, Md. Saiful Islam and Md. Emdadul Haque. "Speaker Identification using MFCC-Domain Support Vector Machine." n.d.
- S. Prabhakar, S. Pankanti, A. Jain. "Biometric recognition for security and privacy ." 2013.
- S. Prabhakar, S. Pankanti, A. Jain. "Biometric recognition: security and privacy concerns." *IEEE Security & Privacy Magazine*, 2003: 33–42.
- S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu. "Audio-Visual Speech Recognition Using Deep Bottleneck Features and HighPerformance Lipreading." *in Proceedings of APSIPA Annual Summit and Conference. Asia-Pacific*, December 2015.
- S.R.M. Prasanna, S.G. Cheedella, B. Yegnanarayana. "Extraction of speaker-specific excitation information from linear prediction residual of speech." *Speech Communication*, 2006: 1243-1261.
- Saghee, Alaa, Naoyuki Tsurut, and Rin-Ichiro Taniguchi. "Japanese Lip-Reading System." *TECHNICAL REPORT OF IEICE*, 2004.

- Saitoh, Takeshi. "Analysis of Efficient Lip Reading Method for Various Languages." *IEEE Xplore* (ResearchGate), 2009.
- Salahat, Ehab, and Murad Qasaimeh. "Recent Advances in Features Extraction and Description Algorithms: A Comprehensive Survey." *IEEE*, Mar 2017.
- Sambur, M.R. "Speaker recognition using orthogonal linear prediction." *IEEE Transactions on Acoustics Speech and Signal Processing*, 1976: 283-9.
- Sana Boujnah, Paul L. Rosin, et.al. "3-Step Speaker Identification Approach in Degraded Conditions." *researchgate*, 2018.
- Sara Saravi, Iffat Zafar, Eran A. Edirisinghe, Roy S Kalawsky. "Real-time speaker identification for video conferencing." 2020.
- Saravanan, Chandran. "Color Image to Grayscale Image Conversion." *IEEE Xplor*. Durgapur, West Bengal, India: ResearchGate, April 2010. 1-5.
- Sarkar A, Matrouf D, Bousquet P, Bonastre J. "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification." *In: Thirteenth annual conference of the international speech communication association, INTERSPEECH*,. 2012. 2662–2665.
- Scherer, D., A. Müller, and S Behnke. "Evaluation of pooling operations in convolutional architectures for object recognition ." *In Proceedings of the International Conference on Artificial Neural Networks*. Thessaloniki, Greece: pp, 15–18 September 2010. 92–101.
- Schroeder, M. R. "Speech Recognition and speaker recognition." *Lincoln Laboratory, Massachusetts Institute of Technology*, 2014: 41-65.
- Sholokhov A, Sahidullah M, Kinnunen T. "Semi-supervised speech activity detection with an application to automatic speaker verification." *Comput Speech Lang*, 2018.
- Shreya. R. Garipalli, B. V. Sathe-Pathak, A. R. Panat. "Analysis of Speech Signals Using Excitation Source Information." *IEEE*, 2016.
- SIGMUND, ING.MILAN. "SPEAKER RECOGNITION IDENTIFYING PEOPLE BY THEIR VOICE." 2000: 21.
- Sigmund, Milan. "SPEAKER RECOGNITION IDENTIFYING PEOPLE BY THEIR VOICE." 2000.
- Sigmund, Milan. "SPEAKER RECOGNITION IDENTIFYING PEOPLE BY THEIR VOICE." 2000.

- Singh, Archana, Sanjana Yadav, and Neeraj Singh. "Contrast Enhancement and Brightness Preservation using Global-Local Image Enhancement Techniques." *Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. India, 2016.
- Singla, Atul Sharma and Sunil Kumar. "State of art modeling techniques in speaker recognition." *International Journal of Electronics Engineering*, 2017: 186-195.
- Singla, Atul Sharma and Sunil Kumar. "State of the are modeling for speaker recognition ." *International Journal of Electronics Engineering*, 2017: 186-195.
- Singla, Atul Sharma and Sunil Kumar. "State of the art Modeling techniques in speaker recognition ." *International Journal of Electronics Engineering*, 2017: 186-195.
- Singla, Atul Sharma and Sunil Kumar. "State of the art modeling techniques in speaker recognition." *International Journal of Electronics Engineering*, 2017: 186-195.
- Slomka, S. "Multiple classifier structures for automatic speaker recognition under adverse conditions." *PhD thesis, QUT, Brisbane Australia*, 1999.
- Snyder D, Ghahremani P, Povey D, Garcia-Romero D, Carmiel. "Deep neural network-based speaker embeddings for end-to-end speaker verification." *IEEE spoken language technology workshop (SLT)*, 2016: 165–170.
- Solomon Tefera. "Automatic Speech Recognition for Amharic." Addis Ababa, 2005.
- Spectrogram online available*. n.d. <https://en.wikipedia.org/wiki/Spectrogram> (accessed June 23, 2020).
- "State of art modeling techniques in speaker recognition." *International Journal of Electronics Engineering*, 2017: 186-195.
- Suma Paulose, Dominic Mathew, et.al. "Performance Evaluation of Different Modeling Methods and Classifiers with MFCC and IHC Features for Speaker Recognition." *ScienceDirect*, 2017: 55-62.
- Sunil Kumar Singla, Atul Sharma. "stat of art modeling technique in speaker recognition." *International Journal of Electronics Engineering*, 2017: 186-195.
- Supaporn B., et.al. "Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network." *International Journal of Machine Learning and Computing*, 2019.

- Supaporn B., Thuttaphol Inkian, et.al. "Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network." *International Journal of Machine Learning and Computing*, 2019.
- Supaporn Bunrit, et.al. "Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network." *International Journal of Machine Learning and Computing*, 2019.
- Supaporn Bunrit, Nittaya Kerdprasop. "Text-independent speaker identification using deep learning model of convolution neural network." *International Journal of Machine Learning and Computing*, 2019.
- T. Matsui, and S. Furui. "Comparison of text-independent speaker recognition methods using VQ-distortion and Discrete/continuous HMMs." *IEEE Transactions on Speech Audio Processing*, 1994: 456-9.
- Tadess, Kinf. *Sub word based Amharic word recognition : an Experment using hidden markov model*. Addis Ababa, Ethiopia: unpublished, 2002.
- Tadesse, Kinf. *Sub word based Amharic word recognition : an Experment using hidden markov model*. Addis Ababa, Ethiopia: unpublished, 2002.
- Terissi, Lucas D., Marianela Parodi, and Juan C. G'omez. "Lip reading using wavelet-based features and Random Forests classification." *International Conference on Pattern Recognition*. IEEE, 2014.
- Thein, Thein, and Kalyar Myo San. "Lip Movements Recognition Towards An Automatic Lip Reading System for Myanmar Consonants." *IEEE Xplore*, 2018.
- Thipe Modipa, Mercy Mosibudi Mogale, et.al. "Automatic Speaker Recognition System based on Machine Learning Algorithms." (*SAUPEC/RobMech/PRASA*). Bloemfontein, South Africa: IEEE, 2019. 24-32.
- Thipe Modipa, Mercy Mosibudi, et.al. "Automatic Speaker Recognition System based on Machine Learning Algorithms." *SAUPEC/RobMech/PRASA Conference*. Bloemfontein, 2019. 28-30.
- Togneri R, Pullella D. "An overview of speaker identification: accuracy and robustness issues." *IEEE Circuits Syst Mag*, 2011.
- Toh, K.A. "Fingerprint and speaker verification decisions fusion." *In Proceedings of 12th International Conference on Image Analysis and Processing*, 2003: 626-631.

- Toh, K.A. "Fingerprint and speaker verification decisions fusion." *In Proceedings of 12th International Conference on Image Analysis and Processing*, 2003: 626-631.
- Tolba, Hesham. "A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach." *science direct*, 2018: 43-47.
- Tushar Ranjan Sahoo, Sabyasachi Patra. "Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification." 2014: 27-35.
- Vapnik, V. "Estimation of Dependences Based on empirical data." *Springer-Verlag*, 1982.
- Vats, Prakhar, and Prof. Krishna Samdani. "Study on Machine Learning Techniques In Financial Markets." *IEEE 978-1-7281-1524-5*, 2019.
- Venkatesan R, Ganesh AB. "Binaural classification-based speech segregation and robust speaker recognition system." *Circuits Syst Signal Process*. 2018. 3383–3411.
- Ville Vestman, Tomi Kinnunen, Rosa González, Hautamäki MdSahidullahb. "Voice Mimicry Attacks Assisted by Automatic Speaker Verification." *science direct*, 2019.
- Vina Ayumi, L.M. Rasdi Rere, et.al. "Optimization of Convolutional Neural Network using Microcanonical Annealing Algorithm." *researchgate*, 2016: 6424-6430.
- W. M. Campbell, J. P. Campbell, et.al. "Support vector machines for speaker and language recognition." *Computer Speech & Language*, 2006: 210-229.
- W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-. "Support vector machines for speaker and language recognition." *Computer Speech and Language*, 2006: 210-29.
- Wan, V. "Speaker verification using support vector machines." *Ph.D. dissertation, University of Sheffield*, 2003.
- Wan, V. "Speaker verification using support vector machines." *Ph.D. dissertation, University of Sheffield*, 2003.
- Wang D, Li L, Zhang C, Zheng TF. "Improving short utterance speaker recognition by modeling speech unit classes." *IEEE/ACM Trans Audio Speech Lang Process (TASLP)*, 2016.
- Wang, X. "Text dependent speaker verification using recurrent neural time delay neural networks for feature extraction." *In IEEE signal processing workshop*, 1993.
- X. Zhu, et al. "A comparative study of mixture-Gaussian VQ, ergodic HMMs and left-to-right HMMs for speaker recognition." *In international symposium on Speech Image processing and Neural Networks*, 1994.

- Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. "Deep learning for visual understanding: a review." *Neurocomputing*, 2016: 27–48.
- Y. Li, S. Q. Tian Wang, and X. Ding. "A survey of recent advances in visual feature detection." *Neurocomputing*, 2015: 736–751.
- Yadav, D.G. Bhalke and Shailaja S. "Speaker Identification System using Wavelet Transform and VQ modeling Technique." *International Journal of Computer Applications*, 2015: 267-282.
- Ye, J. Liu and Y. "An instantiable speech biometrics module with natural language interface: Implementation in the telephony environment." *Proc. of the ICASSP*, 2000.
- Yegnanarayana, K. Sri Rama Murty and B. "Combining evidence from residual phase and MFCC features for speaker recognition." *IEEE Signal Processing Letters*, 2006: 52-55.
- Yonas, Itansa. "An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System." *Università degli Studi di Brescia*, 2014: 15-27.
- Yu C, Liu G, Hahm S, Hansen JHL. "Uncertainty propagation in front end factor analysis for noise robust speaker recognition." *Proceedings of the 39th ICASSP, Florence*,. Italy, 2014. 4017–4021.
- Yuk, C.C.Q.L.D.-S. "An HMM approach to text independent speaker verification." *In IEEE International conference on Acoustics, Speech and Signal Processing*, 1996.
- Zahir H., Muntasir Al-Asfoor, Mohammed Hamzah Abed. "Architectural Heritage Images Classification Using Deep Learning With CNN." *Researchgate*, 2020: 1-11.
- Zhang C, Koishida K. "End-to-end text-independent speaker verification with triplet loss on short utterances." *Interspeech*, 2017.
- Zhang SX, Chen Z, Zhao Y, Li J, Gong Y. "end-to-end attention based text-dependent speaker verification." *arXiv preprint*, 2017.
- Zhong Z, Zheng L, Kang G, Li S, Yang Y. "Random erasing data augmentation." *Computer Vision and Pattern Recognition*, 2017: 12-23.
- Zhonghua, Fu, and Zhao Rongchun. "An overview of modeling technology of speaker recognition." *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*. Nanjing, China: IEEE, 2004. 887-891.
- Zia Saquib, Nirmala Salam\*, Rekha P Nair, Nipun Pandey, Akanksha Joshi. "A Survey on Automatic Speaker Recognition Systems." *science direct*, 2017.

