DESIGNING AN AUTOMATIC SPEECH BASED DIALECT IDENTIFICATION USING MACHINE LEARNING APPROACHES

Mezmur, Bekalu
BAHIRDAR UNIVERSITY
BAHIRDAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTING

DESIGNING AN AUTOMATIC SPEECH BASED DIALECT IDENTIFICATION USING MACHINE LEARNING APPROACHES

MSc. Thesis

By

Mezmur Bekalu

Program: MSc. Extension Program on IT

Advisor: Abraham Debasu (Assistant Professor)

Bahir Dar, Ethiopia
August, 2022
I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student: Mezmur Bekala Signature: Date: 23/03/2022

As members of the board of examiners, we examined this thesis entitled “Designing an automatic speech based dialect identification using machine learning approaches” by Mezmur Bekala. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in “Information Technology”.

Board of Examiners
Name of Advisor: Berhanu Debebe Signature: Date: 01/10/2022

Name of External examiner
Martha Yifru (PhD) Signature: Date: 28/09/2022

Name of Internal Examiner
Tesfu Tesfaye (PhD) Signature: Date: 30/09/2022

Name of Chairperson
Dagnechew Melkewen Signature: Date: 01/10/2022

Name of Chair Holder
Abulkarim M. (PhD) Signature: Date: 17/11/2022

Name of Faculty Dean: Megabregan E. Signature: Date: Nov 17, 2022

Faculty Stamp
Dedicated to:
Acknowledgements

First of all, I would like to thank my advisor Abrham Debasu for his guidance, suggestions, encouragement, and support throughout and giving me his time. Further I would like to thank my friend Zemenu Mekonnen, for his support and helpful comments in writing paper.
# Table of contents

Acknowledgements .................................................................................................................. i  
List of Figures .......................................................................................................................... v  
Abbreviations .......................................................................................................................... vii  
Abstract ................................................................................................................................... viii  

**CHAPTER ONE** ...................................................................................................................... 1

1. Introduction ........................................................................................................................... 1  
1.1. Background ......................................................................................................................... 1  
1.2. Statement of the Problem .................................................................................................... 3  
1.3. Motivation ............................................................................................................................ 5  
1.4. Objectives ............................................................................................................................ 6  
1.4.1. General objective ............................................................................................................. 6  
1.4.2. Specific Objectives ........................................................................................................... 6  
1.5. Scope and limitation of the study ..................................................................................... 6  
1.6. Methodology ....................................................................................................................... 7  
1.6.1. Research Design ............................................................................................................. 7  
1.7. Significance of the study .................................................................................................... 11  
1.8. Organization of the Thesis ................................................................................................ 11  

**CHAPTER TWO** .................................................................................................................... 12

2. Literature review .................................................................................................................. 12  
2.1. Amharic Language ............................................................................................................. 12  
2.1.1. Amharic Phonetics ......................................................................................................... 12  
2.2. Speech production in humans ............................................................................................ 14  
2.3. What is dialect? .................................................................................................................. 15
2.4. Amharic dialects .................................................................................................................................. 16
2.4.1. Phonological difference .................................................................................................................. 16
2.4.2. Approaches to dialect identification .............................................................................................. 19
2.5. Convolutional Neural Network ......................................................................................................... 21
2.5.1. Layers in CNN ................................................................................................................................ 22
2.5.1.1. Hyper parameters of Convolution ............................................................................................... 23
2.6. Support Vector Machine (SVM) ........................................................................................................ 27
2.7. Audio feature extraction ..................................................................................................................... 28
2.7.2. Mel Spectrogram ............................................................................................................................ 28
2.8.1. Dialect ............................................................................................................................................. 30
2.8.2. Applications of dialect recognition ................................................................................................. 30
2.9. Related Works ..................................................................................................................................... 31
2.10. Summary ........................................................................................................................................ 34

CHAPTER THREE ........................................................................................................................................ 36
3. Dialect classification Design Methodology ............................................................................................... 36
3.1. Introduction ......................................................................................................................................... 36
3.2. Proposed architecture ........................................................................................................................ 37
3.3. Audio data acquisition ........................................................................................................................ 38
3.3.1. Audio Preprocessing ....................................................................................................................... 39
3.3.2. Silence removal ............................................................................................................................... 39
3.3.3. Noise removal ................................................................................................................................. 41
3.3.4. Audio to Mel Spectrogram generation ........................................................................................... 43
3.3.5. Image size adjustment .................................................................................................................... 45
3.3.6. Feature extraction .......................................................................................................................... 45
List of Figures

Figure 2.1-1: Speech Production ................................................................. 14
Figure 2.4-1: Architecture of Convolutional Neural Network ................................. 21
Figure 2.4-2: CNN Convolution Operation .................................................. 24
Figure 2.4-3: Maximum pooling .................................................................... 26
Figure 2.4-4: Fully Connected Layer ................................................................ 26
Figure 3.2: Architecture of the model ............................................................. 37
Figure 3.2-1: Audio Wave in the presence of silence .............................................. 40
Figure 3.2-2: Silence and sound representation of Audio wave ............................ 40
Figure 3.2-3: Audio Wave after silence removal ................................................ 41
Figure 3.2-4: Plot of Audio signal after noise removal .......................................... 43
Figure 3.2-5: Spectrogram images .................................................................... 44
Figure 3.2-6: CNN for dialect recognition .......................................................... 49
Figure 4.5-1: Confusion Matrix for CNN-SVM ................................................ 56
Figure 4.5-2: Classification Report for CNN .................................................... 57
Figure 4.5-3:- The architecture of CNN model ................................................... 59
Figure 4.5-4: CNN-Dialect identification model training ....................................... 60
Figure 4.5-5: Accuracy to Loss Curve .............................................................. 61
Figure 4.5-6: Confusion Matrix of CNN model .................................................. 62
Figure 4.5-7: Classification Report of CNN model ............................................. 63
Figure 4.6.1-1: VGGNet Model Training ........................................................ 64
Figure 4.6.1-2: Training and Validation accuracy and loss .................................... 64
Figure 4.6.1-3: Training of ALexNet network .................................................. 65
Figure 4.6.1-4: Training and Validation accuracy and Loss of ALexNet ................ 66
Figure 4.6.1-6: training of LeNet Model ............................................................ 67
Figure 4.6.1-7: training and Validation accuracy and loss curve .......................... 68
Figure 4.6.1-8: chart for accuracy of different models ......................................... 72
Figure 4.6.1-9: Chart for precision, recall and F-measur ....................................... 73
List of Tables

Table 2.1-1: Categories of Amharic Vowels ................................................................. 13
Table 2.1-2: Categories of Amharic Consonants ..........................................................13
Table 2.1-3: Amharic consonant phonemes................................................................. 17
Table 2.1-4: Example for Amharic phonemes in Amharic regional dialects ............. 18
Table 2.1-5: Summary of Related work ........................................................................ 35
Table 4.3-1: Experimental Tool Specification ..............................................................53
Table 4.3-2: Audio data Distribution for each Dialect ................................................. 53
Table 4.3-3: Comparison of Amharic dialect identification Models ............................69
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCI</td>
<td>Human Computer Interaction</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantization</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstral Coefficient</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural network</td>
</tr>
<tr>
<td>ADI</td>
<td>Automatic Dialect Identification</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>DID</td>
<td>Dialect identification</td>
</tr>
<tr>
<td>MSA</td>
<td>Modern Standard Arabic</td>
</tr>
<tr>
<td>LDC</td>
<td>Linguistic Data Consortium</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
Abstract

In Ethiopia more than 200 different dialects are spoken in 83 languages, including the Afan Oromo, Amharic and Tigrigna which have the largest ethnic and linguistic groups (Wimsatt & Wynn, 2011). According to Wimsatt & Wynn (2011), Amharic has over 30 million native speakers living in Ethiopia and 62 million speakers globally. It has five known dialect categories that are spoken in different parts of Ethiopia such as Addis Ababa dialect, Gojjam, Gondar, Wollo and Shewa dialects.

But for the purpose of this study only the dialects that are spoken in Amhara region namely gojam, wollo, shewa and gonder are considered. There are few research attempts to develop dialect classification models to classify Amharic dialect categories. However Most of the studies focused on identifying Amharic dialects using audio data recorded in controlled environment which is relatively free from noise. On top of this the methods used on the existing research has its own drawbacks to perform classification. The purpose of this study therefore is to explore the possibility of developing dialect classification model by using audio data recorded in uncontrolled environment that involves background noises using other machine learning techniques.

In this study an attempt is made to develop Amharic dialect classification model using CNN and CNN-SVM techniques. A spontaneous data is collected from amhara media corporation archive system. Since the data is recorded from different parts of amhara region by a camera man, it is uncontrolled data which contains background noises. This data is then stored on AMECO archive system as raw data. For each Amharic dialect category 300 speech data is collected which contains a total of 1200 Utterances spoken by people who live in Amhara region. Since the data contains background noises and other irrelevant things preprocessing operations are performed to remove the different types of noise and silence in the audio signal. Silence is removed by applying thresholding technique and the background noise removed by applying moving average filter which is a low pass filter.

Audio features are extracted as a form of spectrogram and used for model development. Our experiments confirmed that using CNN model an Accuracy of 85% is achieved when RELUE activation function is used and 79% accuracy achieve when tan activation function used. The
accuracy obtained for both techniques (CNN and CNN-SVM) is compared and CNN alone achieved better classification accuracy. Our CNN Amharic dialect identification model is compared with state-of-the-art models and showed better recognition performance on the current data sets used.

In general, training deep learning algorithms with more data will increase the accuracy of the recognition model. Therefore, it is better to use more data and other speech preprocessing operations to further improve the accuracy of Dialect Identification model. On top of this it is recommended to consider a robust system which handles background noises collected from uncontrolled environments by it to enhance the performance.

**KEYWORDS:** Amharic Dialects identification, CNN, CNN-SVM, Mel spectrogram, spontaneous speech, Aquostic feature, confusion matrix, low pass filter, state of the art models, uncontrolled data
Chapter One

1. Introduction

1.1. Background

As Wimsatt & Wynn (2011) shown, as far as human exchange continued, there are many languages in the world with their own dialects to share opinions, ideas and feeling between individuals. These dialects have their own assortments based on its geographic region and ethnicity.

According to Webster dictionary dialect is defined as regional variety of a language that is different from other language varieties of the same language in grammar, pronunciation and vocabulary. It is the variety of a language that is used by members of a certain group. Cambridge dictionary defines dialect as system of language that is spoken in particular part of the country and includes certain number of grammar and words. It is the variety of language that tells us where the person comes from. According to Bowen (2011), dialect is a form of language that is different in grammar , lexis and speech phonology from other varieties of the same language. It is the act of expressing the same thing in different way and describes the social structure of an individual such as gender, social class and origin (SOLANO-FLORES, 2006). As stated by SOLANO-FLORE(2006), dialect expresses the linguistic and cultural traits of an individual whose origin is from the same broad linguistic group.

As Lei & Hansen (2011) described, dialect is a neighborhood or social variety of a language categorized by means of the technique they speak. It is an expression of indispensable speaker’s voice signature and it offers records about the speaker’s data such as origin, gender, fitness reputation and age conduct (A.Etman, 2015). Although dialects of the same language share many similarities, they are frequently differentiated at countless linguistic tiers such as phonological (i.e., how it sounds), grammatical, orthographic (e.g., “center” vs. “centre”) and very regularly extraordinary vocabularies (Mohamed G. Elfekya, 2015).

Dialect identification (DID) is a process of automatically identifying the dialect of a spoken or written utterance (Michon, 2018). DID is applicable in many areas such as improving some applications and services in areas of e-learning, e-health and remote access (A. Etman, 2015). In addition, it is also useful
Amharic dialect

Ethiopia is a country with more than 80 languages that are spoken in different parts of the country. Amharic is among one of these languages in Semitic language family and it is the official language of the country. It has around 30 million speakers that speak the language as mother tongue or as second language. It is one of the largest languages in Ethiopia next to Affan Oromo in terms of number of speakers. Amharic has five dialectal variations that are spoken in different parts of Ethiopia (Tadesse, 2018). According to Tadesse (2018), Amharic dialects are variations of language features among the speaking population of the language living in different geographical areas this is because geographical areas are among the necessary characteristics of dialects. The common dialects found in Ethiopia are Wollo, Gondar, Gojjam, Shewa and Addis Ababa. Wollo dialect is spoken in parts of wollo, Gondar dialect is spoken in areas around Gondar, Gojjam dialect is spoken in areas around Gojjam, and Addis Ababa dialect is spoken by peoples whose origin is in Addis Ababa.
1.2. Statement of the Problem

As stated in the background section of this study, there are more than 83 languages spoken in Ethiopia and these languages consist of more than 200 different dialects. Among these languages Amharic, Afan Oromo and Tigrigna are the largest linguistic and ethnic groups having many numbers of speakers. Public census (1994) conducted revealed that Amharic is the most widely spoken language of Ethiopia. Over 30 million peoples speak it as the first or second language and it is most spoken Semitic language in the world after Arabic. It is regarded as one of the largest languages in Africa. This language is originated from the ancient Geez language. Amharic in general has four dialectal variants such as Gojjam, Gondar, Wollo and Shewa. These dialects are slightly different in the way people speak them (Melesew, (2017)). It has its own writing system consisting of 33 primary characters with each character representing a consonant having seven variations with forms that represent vowels.

Dialect identification (DID) is an automatic identification of the corresponding dialect of utterance, either written or spoken (Michon, 2018).

Dialect Identification (DID) is a challenging task, and it becomes more complicated when it is about the identification of dialects that belong to the same country. Indeed, dialects of the same country are closely related and exhibit a significant overlapping at the phonetic and lexical levels (Mannouba, 2020).

One of the essential problems for processing real world spoken content material such as dialect identification is the availability distinct dialects. Amharic is one of the languages which possessed different dialects. Despite the fact that it has such varieties of dialects, only a few studies have been done so far. In the previous study machine learning techniques Gaussian mixture model, Vector quantization and artificial neural network were used to classify dialects. Even if the application of such techniques achieved good results, it has also limitations. VQ for example has limitations to recognize a similar sound due to VQ is a template matching algorithm.

When applying GMM, it is not capable of providing information on the temporal relationships due to lack of memory (to access information for future decisions) (A. Etman, 2015). Applying GMM for classification tasks has posed additional limitations such that when applied on problems having high dimensionality, it failed to work perfectly for computational reasons. Additionally, if the users need the GMM algorithm to fit to the training data, users need to set the number of mixture models. To set the optimal number of mixture models that worked better for the classification task, the user may be required to try different mixture models (Anifowose, 2010).
Recently neural networks are being applied in various tasks like speech recognition, image processing, computer vision and dialect identification and yields better results in terms of recognition accuracy (Derb et al., 2019). Even if applying neural network for Amharic dialect identification and achieved better recognition accuracy as done in (Mengistu & Alemayehu, 2017a), the dataset used for the recognition task is collected from controlled environment and it needs to be tested on datasets collected from uncontrolled environments that involves many types of noise. When speech data is recorded in uncontrolled environment it’s vulnerable to some types of noise such as background noise which limit its application in speech and dialect classification tasks. This is because uncontrolled environment contains disturbances such as sound of vehicle, birds, and people talking in the background and others which may degrade the quality of speech. Somehow, controlled environment is free from such types of noises which make it easier and relevant for speech recognition and dialect recognition tasks. All of the studies conducted on Amharic dialect identification have used speech data recorded in controlled environment. Studies conducted so far will not be robust to speech data recorded in uncontrolled environment, and therefore to make dialect recognition task as robust as possible, speech recorded in uncontrolled environment need to be considered.

Since speech data recorded in uncontrolled environment contains noise as said above it needs some techniques to remove noise from the speech signal. In addition, although the data set preparation included both sexes it excludes age as criteria. All the above holes have a huge impact on the performance of the model. Therefore, to handle such problems further researches has to be done on Amharic dialects using other supervised machine learning techniques such as support vector machine and convolutional neural network as a classifier for classification and using speech recorded in uncontrolled environment.

In this study, a spontaneous speech which contains Amharic dialects is collected in Amhara media corporation archive system. The archive contains programs, news, documentaries and entertaining events which contain the four known dialects. Since the data is recorded in different parts of Amhara region it is prone to background noises. To remove the noise filtering methods (moving average filter or low pass filter) is used.

Supervised Machine learning approaches convolutional neural network (CNN) and hybrid approaches of convolutional neural network and support vector machine (CNN-SVM) is used to identify these dialects.

MFCC is proved to be the most successful feature extraction technique that is applied for extracting relevant features from a given speech signal (Tiwari, 2010). But for the purpose of this study spectrograms
are used as an input to develop the dialect recognition system. Spectrograms are two dimensional representations of a given audio data that shows the time and frequency axis with color dimension to indicate the strength of the frequency component at a given time (Wyse, 2017). At the end of the study, the study will answer the following questions:

- Which technique is better in classifying Amharic dialects from speech recorded in uncontrolled environment?
- How to remove noise from speech recorded from uncontrolled environment?
- Which learning function is suitable for dialect identification?

1.3. Motivation

As it is stated in the background of the study, Amharic is one of the languages that are widely spoken by many peoples in Ethiopia. According to Woldeyohannis & Besacier (2017), Amharic language has more than 30 million speakers with four well known dialects such as Gojjam, Gondar, Wollo and Shewa. It is spoken by individuals from different locations of Amhara region differ slightly in their pronunciation pattern. Since Amharic is a huge language with such dialects researches has to be done to know the behavior of such dialects. Since Dialects are the "voice" of a region or a group, it is important to identify where we come from and who we are. Besides, doing a research on dialects help enrich the culture and identity of a particular group.

Generally speaking Dialect recognition has its application in different areas such as language identification, remote access, speech recognition and classification, e-learning and e-health (Etman & Beex, 2015). On top of this, it is important in identifying the residence of a person wither he is from gonder, gojam, wollo or shewa. Due to this it will have a great role in forensic speaker profiling and in identity authentication and verification.
1.4. Objectives

1.4.1. General objective

The main aim of this study is to develop a model which identifies Amharic dialects from a set of utterances which are spoken in Amhara region.

1.4.2. Specific Objectives

To achieve the general objective, the following activities will be performed:

- To explore different literature works to gain understanding of dialect recognition.
- To investigate the characteristics of the different Amharic dialects
- To prepare the necessary speech data that belong to the four Amharic dialects
- Preprocess the data to make it appropriate for the modeling task
- Design dialect identification model
- Conduct experiments by using different models
- Evaluate the performance of each model
- Fine tune the different models to improve their recognition accuracy
- Compare the performance of dialect identification models and select the best performing model.

1.5. Scope and limitation of the study

The scope of this study is specific to Amharic dialects which are spoken mainly in Amhara region (Gojjam, Shewa, Gondar and Wollo). Dialect identification passes a number of basic processes such as data collection, preprocessing, feature extraction and classification. It involves designing and developing a model for identifying Amharic dialects using CNN and CNN-SVM; make comparison between the two algorithms then select the technique with better accuracy. The data source for this study is limited to audio data collected from Amhara Media Corporation, no manual recording of speech data is made by the researcher. This is because the necessary audio data that belong to the four Amharic dialect categories is available in Amhara Media Corporation archive system. For the purpose of this study, a spontaneous speech having a duration ranging from 13 to 30 seconds is collected from Amhara Media Corporation and used as input for Amharic dialect classification tasks. The speech data includes Non-spontaneous speech that involves the speech made by the people during interview or summit or some type of discussion.
1.6. Methodology

In order to answer the research questions various methodological aspects such as research design, literature review, data collection, tools as well as evaluation techniques are applied. Design science research methodology is followed throughout the stage of the study.

1.6.1. Research Design

The aim of this particular study is to develop a model that can perform Amharic dialect identification. The output of the study is a model which is an artifact in design science, therefore the study follows design science research approach. According to Peffers et. al., (2007) and Esearch et al., (2004) design science researches in information technology (IT) are aimed at creating and evaluating information (IT) artifacts that are intended to solve an identified problem, to attain goals and serve human purpose. This study is also aimed at developing a model that can perform Amharic dialect identification.

Problem Identification and Motivation

In this phase of the study, deep analysis of different literatures is made to identify the common problems that are explored and left unresolved or totally unexplored by previous researchers and plan proper action and justifiable solution to solve them. Accordingly, deep literature review regarding dialect identification in general and Amharic dialect identification in particular is made to gain deep understanding of dialect classification process. Related study from various sources such as journal articles, conference papers, magazines, books are reviewed to explore different techniques used in dialect identification and the process to be followed in dialect recognition. The main problem identified here is that in previous research speech recorded in uncontrolled environment that includes some type of noise is not considered, such type of audio primarily includes background noise and need to be explored for dialect classification task. But dialect classification model needs to be robust to noise in that it should isolate the nosy part of the audio in order to identify noisy audio and classify it. Through different literatures, we can analyze the different audio preprocessing techniques that should be applied in removing silence and the different noises that are introduced during sound recording. Additionally, we explored in what form audio data should be feed to model for better recognition results.

**Objective of the Solution**
After properly stating the problems, the objectives of the study are clearly identified as stated above. The objective of the study is developing dialect classification model that can classify Amharic dialects and categorize them as Gojjam, Gondar, Shewa and Wollo, and then specific objectives are outlined in line with the general objective.

**Design and Development**

In this phase of the study, the researcher is expected to craft and design artificial solutions that demonstrate how the research problem is going to be solved. Developing CNN and CNN-SVM dialect identification model comparing the performance of the two models and adopt the model with high performance to dialect identification task. This is because CNN algorithm is very recent efficient more accurate technique even if it requires huge amount of data. CNN is used as feature extractor and classification algorism. SVM is also very accurate and efficient technique used mostly for classification and recreation problem when the amount of data is limited.

**Tools and Techniques**

Python is used for writing the code for developing the proposed model for Amharic dialect identification model. This is because, Python is a powerful tool and has many built-in functions and libraries which help to perform different speech processing tasks and provide different APIs for developing deep learning models. In addition to this, it has simple syntax, supports various operating systems and the code is shorter than other programing language making easier to use (Magimai-Doss, et al., 2015).

Pycharm is used as an editor to write the code for the implementation. This because pycharm provides hints while writing codes and makes the process easier. This tools provide us with a comprehensive framework for common packages such as Matplotlib, Librosa, OpenCv, NumPy, Pandas, SciPy, IPython and other packages that are used for scientific applications. In this study python libraries such as Keras and TensorFlow are used. TensorFlow is the most common and strongest deep learning framework and Keras is a python based high level neural network API that run on top of TensorFlow (Andargie, 2020).

Audacity, Sony Vegas and avid media composer software environment is used for editing audio data. Different deep learning and machine learning algorithms such as CNN and CNN-SVM are used to develop the model for dialect identification.
To make the sound wave more appropriate for the dialect recognition task, different preprocessing tasks such as silence removal, noise removal and converting sound wave into more abstract representations such as spectrogram is made as part of preprocessing stage. Keras library (using TensorFlow as backend) is used to design our CNN model. Edrawmax is used for drawing the architecture model and workflow of sub models.

Keras is a powerful library running on top of TensorFlow. TensorFlow provides a graph data structure that allows handling computational graphs. Keras on the other hand consists of two kinds of models (Sequential and functional). In sequential model has simple architecture in which layers are stacked together in a linear fashion. Functional model on the other hand provides more general models having diverse layer structure like multi-output models (Belsti, 2020).

TensorFlow, in general terms, is a software framework for numerical competitions based on dataflow graphs (Stanford&baidu, (n.d.). In this graph, nodes represent operations (such as addition or division) and edges represent data (tensors) flowing around the system. Tensors are the standard way of representing data in deep learning. Simply put, tensors are just multidimensional arrays, an extension of two-dimensional tables or matrices to data with high dimensionality.

Data collection
To design an Amharic dialect identification model, Amharic speech data consisting of the four dialect groups such as Gojjam, Gondar, Shewa and Wollo is needed. Since the aim of the study is to develop a model for Amharic dialect recognition for uncontrolled environment; data from uncontrolled environment is needed. Therefore, the data needed for developing the model is collected from Amhara Media Corporation archive system that contain recorded speech from different events such as news, programs, documentaries, entertaining and etc. The data collected consists of speech from all Amharic dialects such as Gojjam, Gondar, Shewa and Wollo where 300 utterances (speakers) of audio data for each dialect category are collected and organized to be used as data source.

Amhara Media Corporation archive system contains speech recorded in uncontrolled environment which consists of various types of noise introduced during recording such as sound of cars, birds, person talking in the background, sound of wind and others which may degrade the quality of speech recorded.

Preprocessing
The data collected from Amhara Media Corporation mainly consists of audio files collected from uncontrolled environment in different formats. Therefore, before the audio actually is used for model construction, it needs to be preprocessed in advance to make it suitable for the next process. So, some preprocessing operations need to be performed, these include the following.

- **Format conversion:** the audio collected for Amharic dialect recognition consists of audio having different formats such as .mp3, wma, web or any other formats. So, the audio data should be converted into similar data formats. For speech recognition and dialect identification tasks .wav audio formats are more preferable for python. So, all the audio files are converted to .wav file format by using audacity audio processing software.

- **Silence and noise removal:** Since the data is collected in uncontrolled environment, it includes silence and noise introduced. Thus, this silence and noise need to be removed to make the speech better for the identification task.

- **Audio to spectrogram conversion:** once the audio data is preprocessed and made appropriate for the next task, it is then converted into spectrograms images.

**Model Development**

The data collected is preprocessed and converted into appropriate format, then the next activity to be done is using preprocessed data to build dialect identification model. Deep learning algorithm like CNN and machine learning algorithm like SVM is applied to develop the model. This designed model is implemented by using python programming language and the code is written and executed on Pycharm notebook editors. Then comparison is made and the model that performed better is selected for dialect identification.

**Demonstration**

The important step in demonstration phase is validating the use of the artifact for solving the problem. Therefore, the developed model is demonstrated by showing to what extent the model identifies each audio data to the appropriate dialect category.

**Evaluation**
Evaluating the developed model is vital in design science researches as it measures how well the developed artifact supports a solution to the problem. Performance evaluation should be conducted using collected test data. The test data is given to the model and the performance of the system are evaluated using evaluation metrics like accuracy, precision, recall and F1-measure.

- Communication

The final phase in the process model is communicating the problem and its importance, the rigor of the design and novelty of the artifact, and its effectiveness to researchers and audiences who are concerned professionals. Communication is being done through this paper.

1.7. Significance of the study

Since dialects are the "voice" of a region or a group, dialect identification is important to identify where we come from and who we are. Besides, dialects help to enrich the culture and identity of a particular group. Dialect classification is very essential to indicate a speaker’s regional origin and ethnicity which means whether he is from Gondar, Gojjam, Wollo or Shewa and this will have a great role in forensic speaker profiling, in identity authentication and verification. Doing this research will have a great contribution especially in giving the language recognition in the areas of dialect recognition system.

In general accurate identification of dialects has a great role to advance certain areas of service like Automatic speaker identification, HCI, e-health, e-learning, etc. It has also a great role in Automatic Speech Recognition; it will enable the speech recognition system to adapt its pronunciation, acoustic, and language models appropriately (Chittaragi, 2017).

1.8. Organization of the Thesis

The first chapter of this study provides us the overall focus and purpose of the study. It clearly shows how the study is going to be performed in general and provides basic information such as dataset used and the techniques to be applied. The thesis consists of five chapters which includes the introductory chapter that provides an overview of the study the second is literature review and related works. The third chapter depicts the methodology of this thesis.
The fourth presents and discusses the result and the final chapter provides a conclusion, description and recommendation.

Chapter Two

2. Literature review

In this section, the different concepts related to Amharic dialect recognition and research works done related to dialect identification in general and Amharic dialect identification in particular are discussed. The different speech preprocessing and dialect recognition techniques that are widely applied in dialect identification are well discussed.

2.1. Amharic Language

Amharic is the official working language of Federal Government of Ethiopia. Within Semitic language family, it has the greatest number of speakers next to Arabic language. It is one of the languages with its own writing system (Tefera, n.d.). Many people use it as first language (mother tongue) for communication. It has five dialectical variations such as Gojjam, Gondar, Shewa, Wollo and Addis Ababa which are spoken in different parts of Ethiopia.

2.1.1. Amharic Phonetics

Like other languages, Amharic has its own phonology, phonetic and morphological features. It has speech sounds such as ወ Fence, ከ Fence, ኲ Fence and etc. that are not available in other languages. According to WOLDE (2011), Amharic language consists of its own inventory of speech sounds that include thirty (30) consonants and seven vowels. These consonants in general can be classified as stops, fricatives, nasals, liquids, and glides.
### Table 2.1-1: Categories of Amharic Vowels

<table>
<thead>
<tr>
<th>Fonts</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>እ</td>
<td>ኧ</td>
</tr>
<tr>
<td>Middle</td>
<td>ከ</td>
<td>ኩ</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>ኲ</td>
</tr>
</tbody>
</table>

### Table 2.1-2: Categories of Amharic Consonants

<table>
<thead>
<tr>
<th>Manner of Articulation</th>
<th>Voicing</th>
<th>Place of Articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Labials</td>
</tr>
<tr>
<td>Stops</td>
<td>Voiceless</td>
<td>ጡ</td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>ጫ</td>
</tr>
<tr>
<td></td>
<td>Glottal zed</td>
<td>ጲ</td>
</tr>
<tr>
<td>Fricatives</td>
<td>Voiceless</td>
<td>ጫ</td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>ቫ</td>
</tr>
<tr>
<td></td>
<td>Glottal zed</td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>Voiced</td>
<td>዗</td>
</tr>
<tr>
<td>Liquids</td>
<td>Voiced</td>
<td>ዧ</td>
</tr>
<tr>
<td>Glides</td>
<td>Voiced</td>
<td>ቦ</td>
</tr>
</tbody>
</table>
2.2. Speech production in humans

Speech is the common form of communication used by many people in the world. It allows people to share their ideas and thoughts to others. There are different ways of communication for people to convey and transfer information from person to person. These include natural speech, written text and different signs including sign languages used by impaired peoples. For many peoples Communication through spoken languages is the primary and most dominant mode of communication. People prefer to communicate through spoken languages by uttering words and phrases. When peoples, speak a word the vibration of air molecules in oral cavity will create sound.

![Speech Production Diagram]

Figure 2.2-1: Speech Production

According to Krishnamurth (2010), in humans, sound is created when air from the lungs excites the air cavity of the mouth. The diagram above shows the structure of human speech. When producing a voiced sound, the glottis periodically opens and closes, resulting in puffs of air exiting the oral cavity. When the air comes out of your lips, the waveform coming out of the lips has the qualities of both the excitation and the resonant cavity. Sound is produced when air flow from the lungs is modified by articulators, such as lips, jaw and tongue. Therefore, it is the position of the articulators that is responsible for producing sound. (Kareem et al., 2010).
2.3. What is dialect?

As ISIK (2020) stated, a dialect is a local form of speech that differs from the standard language at a certain rate. It is one of the topics that are widely studied in speech recognition and it helps to improve the performance of speech recognition systems. Dialect identification is the activity of detecting the dialect of specific fragment of speech spontaneously (Kareem et al., 2019). For speech recognition to achieve excellent results, the spoken dialect has to be identified in advance which calls to greatest applicability of dialect identification to speech recognition.

Most researchers presume dialect to be an accent, but accent is the pronunciation of a speaker, while dialect involves the lexical, grammatical, and phonological differences in pronunciation of the speaker. According to SOLANO-FLORES (2006), a dialect is a regional variation of a language, typically spoken by a specific group of people. There are a variety of speeches spoken in different societies and geographic areas. This voice characteristic can be recognized by pronunciation, phonemes, and other traits such as nasality, loudness, and tonality. It can indicate information about the speaker's origin, gender, age, and health status. (A. Etman, 2015).

Even though dialects of the identical language share many similarities, they are frequently differentiated at several linguistic tiers such as phonological (i.e., how it sounds), grammatical, orthographic (e.g., “center” vs. “centre”) and very often different vocabularies (Mohamed G. Elfekya, 2015).

According to Michon (2018), Dialect identification (DID) is an automated identification of the corresponding dialect of utterance, either written or spoken. It is a distinct occasion of language consciousness that requires a capability to favor between numerous individuals within the same language family, as opposed to throughout language families. In distinction to the language identification scenario, dialects commonly share an exclusive phonetic catalogue and written language. Thus, we also centered on linguistic aspects (i.e. words, characters and phonemes) that could be utilized by automatic speech recognition.
2.4. Amharic dialects

Dialects are local talking styles that are slightly exclusive from ordinary language having the same root language. People dwelling in similar vicinity have uniform dialect traits (ISIK, 2020). As ISIK (2020) stated, when it is contrasted with the language they belong to, dialects can also vary to each other in the concepts of morphological, lexical, syntactic aspects and phonetic characteristics. Amharic language is being spoken in different parts of Ethiopia both as first language and second language. The Speaking style of Amharic speakers who live in different parts of Ethiopia is different from one another. These makes Amharic to have different dialect categories based on the area speakers are grown-up such as Gondar, Gojjam, Shewa, Wollo and Addis Ababa. Being a language spoken in different areas of Ethiopia, Amharic has different local dialects. There is perceived difference between each local dialect in terms of vocabulary, pronunciation and grammar. But the standard for written and spoken Amharic is believed to be the dialect of Addis Ababa. Amharic has Gojjam dialect with a speaking style predominantly spoken by peoples who live around Gojjam area. Gondar dialect is a speaking style predominantly spoken by peoples who live in around Gondar area and Shewa and Wollo dialects are speaking styles exercised around Shewa and Wollo area. There is difference between these dialects such as Phonological Differences and Morphological variation are observed among the different dialects of Amharic.

2.4.1. Phonological difference

According to (TADESSE, 2018), phonological difference is the change in pronunciation that is observed in a specific language. It is the change in speech sounds between different dialects but similar in related dialects. As discussed in his study, (TADESSE, 2018) the phonological variation in Amharic language takes Addis Ababa dialect as a base dialect and classified the observed differences between dialects as devoicing, substitution, insertion, omission, labialization, lention, fortition, gemination, assimilation of consonant and vowel deletion, and metathesis of phonemes at initial, medial or final position of words. These aspects make some difference to appear between the different dialects especially between Addis Ababa dialect and regional dialects such as Gojjam, Gondar, Wollo and Shewa, but such differences do not impede speakers of different dialect from understanding to each other. Amharic has thirty (30) consonant phonemes as presented below.
As stated by Getahun and Baye, Amharic dialect that is regarded as the standard dialect (Addis Ababa dialect) contains greater number of consonant phenomes as compared to regional dialects due to the fact that regional dialects do not contain bilabial stop /p/, and the ejective bilabial stop /p’/.

<table>
<thead>
<tr>
<th></th>
<th>Labial</th>
<th>Alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Labio-velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stop</strong></td>
<td>VI</td>
<td>P</td>
<td>t</td>
<td>k</td>
<td>k’</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>Vd</td>
<td>b</td>
<td>d</td>
<td>g</td>
<td>g’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ej</td>
<td>p’</td>
<td>t’</td>
<td>k’</td>
<td>k’’</td>
<td></td>
</tr>
<tr>
<td><strong>Affricate</strong></td>
<td>VI</td>
<td></td>
<td>j’’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vd</td>
<td></td>
<td>d’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ej</td>
<td></td>
<td>j’’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td>VI</td>
<td>f</td>
<td>s</td>
<td>j’</td>
<td>h</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vd</td>
<td>z</td>
<td>ʒ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ej</td>
<td>s’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nasal</strong></td>
<td></td>
<td>m</td>
<td>n</td>
<td>j</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Liquid</strong></td>
<td></td>
<td>l</td>
<td>r</td>
<td>j</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Glide</strong></td>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.1-3: Amharic Consonant Phenomes [source: (TADESSE, 2018)]*
<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Possible phonetic realizations</th>
<th>Notes and Description</th>
<th>Examples</th>
<th>Place</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>bⁿ</td>
<td>Voiced, bilabial, stop, palatalized</td>
<td>bⁿ$^{t}_1$e</td>
<td>WO</td>
<td>‘equal’</td>
</tr>
<tr>
<td>bʷ</td>
<td>Voiced, bilabial, stop, labialized</td>
<td>abʷar’a</td>
<td>WO and NS</td>
<td>‘dust’</td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>Voiced, bilabial, fricative, spiranted</td>
<td>a. santiβ</td>
<td>WO, GR, NS and GM</td>
<td>‘coin’</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. noβs</td>
<td></td>
<td></td>
<td>‘soul’</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>tⁿ</td>
<td>Voiceless, alveolar, stop, palatalized</td>
<td>tⁿ$^{t}_1$f</td>
<td>GM</td>
<td>‘full to over flowing (measure, river in flood)’</td>
</tr>
<tr>
<td>tʷ</td>
<td>Voiceless, alveolar, stop, labialized</td>
<td>tʷəfa</td>
<td>GM</td>
<td>‘clay cooking pot’</td>
<td></td>
</tr>
<tr>
<td>tʰ</td>
<td>Voiceless, alveolar, stop, aspirated</td>
<td>a. tʰaza</td>
<td>WO</td>
<td>‘a veranda’</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.1-4: Example for Amharic phonemes in Amharic Reginal Dialects [source: (TADESSE, 2018)]

Amharic also contains other phonetic units such as vowels and syllables. Amharic has seven distinct vowel sequences as represented below.

<table>
<thead>
<tr>
<th>Front</th>
<th>central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>Mid</td>
<td>e</td>
<td>ò</td>
</tr>
<tr>
<td>Low</td>
<td>a</td>
<td></td>
</tr>
</tbody>
</table>

Syllables are also another phonetic unit that are regarded as the heart of phonological representation where it’s structure is attributed to sound change in a specific language. The
syllable structure of Amharic is studied by different scholars but according to Mengistu (2018), Amharic syllable structure can be generalized as V, CV, CVC, VC, CVCC, VCC. Several variations exist in different dialects of Amharic, one is conditioned consonant change that can be expressed as palatalization, labialization, gemination, assimilation, devoicing and deletion. For example in Wollo, Gojjam, South Gonder and North Shewa dialects bilabial /m/ is realized as alveolar [n] as a result of partial assimilation due to the presence of nearby alveolar ejective /t'/. Therefore, the word እምጡ (amt’u) in Addis Ababa and North Gondar dialect will be እንጡ ([ant’u]) in gojjam, wollo, north Shewa and south Gonder dialects due to assimilation effect.

In some cases, as a result of consonant deletion the letter w( воп) will be deleted when followed by d(ድ) in dialects of wollo and gojjam. Example the word awdimma (አውድማ) in addis Ababa, Gondar and Shewa dialects which is to mean threshold floor in English will be adwumma in gojjam and wollo dialect. In addition, the glottal fricative /h/ will be deleted when appear in word initial and medial positions with the appearance of a low central vowel /a/, for example the word hamus (ሐሙስ) in Addis Ababa and Gondar dialect which is to mean Thursday in English will become amus(አሙስ) in gojjam, wollo and north Shewa dialect.

As a result of morpheme change, the prefix ta(ታ) in gojjam, Wollo and Addis Ababa dialects is replaced with ka(ካ) in Gondar and North Shewa dialects. Example: the word teleyayu(ተለያዩ) which mean (Separated in English) will be keleyayu in Gondar and North Shewa dialects.

2.4.2. Approaches to dialect identification

2.4.2.1. Phonotactic Approaches

In dialect and accent recognition is based on the hypothesis that dialects or accents differ in their phone sequence distributions. In other words, texts of a same language can be recognized by these character distributions. (Behravan, 2012)

2.4.2.2. Spectral Approaches
It is based on the hypothesis that dialects or accents discriminate in terms of their spectral (acoustic) features. In this approach, speech utterances are represented by a set of spectral feature vectors and the recognition is based on maximum likelihood estimation (Ghule, 2015).

The approach that we used in this study is based on spectral modeling/aquostic modeling techniques. In spectral approaches the dialect identification task is based on the whole utterance including speech and non-speech frames, whereas in phonotactic based approaches the decisions are based on the single phonemes (Ghule, 2015).

2.4.2.3. Types of speech uttered

According to Ghule (2015), the speech recognition system is separated in different classes is based on what type of utterance they have ability to recognize. As stated in Ghule (2015), there are four types of speeches:

i. **Isolated speech:** It needs a single utterance at a time. It sets necessary condition that each utterance having little or no noise on both sides of sample window.

ii. **Connected word:** In this type of utterance the minimum pause between utterances is required to make speech flow smoothly. Which is almost similar to isolated words?

iii. **Continuous speech:** It is a normal human speech, excluding silent pauses between words. It is basically computer’s dictation. What makes more relevant is that it makes the machine understanding much more difficult.

iv. **Spontaneous speech:** It is thought as speech that is natural sounding and no tried out before. An ASR system with spontaneous speech ability should be able to handle a diversity of natural speech features such as words being run at the same time (Ghule, 2015).

2.4.2.4. The text-based approach

The text-based approach has become popular, as its results have been noted in the area of vocabulary acquisition. This approach works by learning target words through reading texts, such as acquiring words’ meaning and their typical language environment from texts. Texts include rich word information such as word family, word meaning, lexical chains, and word association (Mattisson, 2011).

Different techniques are being applied for dialect recognition. Among a variety of techniques deep learning is a newly emerging technique that is being applied in variety of tasks and achieved better results. It is applied in different tasks such as natural language processing, image processing, speech recognition and dialect classification and identification (K et al., n.d.). CNN is one of the special types of deep learning technique that is widely applied in image processing and computer vision. In addition to this, it is applied in speech identification
and dialect identification and achieved promising results. Machine learning algorithms such as SVM can also be applied for image classification and dialect identification of foreign languages.

As ISIK (2020) has shown, CNN and LSTM which are variants of deep learning neural network algorithm are applied for detecting Turkish dialects. In due course, these techniques achieved an accuracy of 85.1%. In the study both phonotactic and acoustic features are used for the recognition task.

According to (K et al., n.d.), B-LSTM and LSTM which are deep learning-based mechanism are applied for the task of dialectic identification in Arabic and German broadcast speech. In the study large vocabulary continuous speech is used and two unique B-LSTM models are created.

2.5. Convolutional Neural Network

CNN is an exceptional kind of neural network for processing facts that has grid-like topology such as time collection records that has 1-Dimension and image facts which is 2-Dimensional data. It is composed of neurons with learnable weights and biases and regarded to be similar to normal neural networks (Manganaro et al., 1999). Recent lookup works stated that Convolutional Neural Network is better in predicting phoneme classification of speech when compared to HMM-based models.
2.5.1. Layers in CNN

A simple CNN is a sequence of layers, with each layer associated with converting the information into a more complex representation and pass to the next layer for generalization. CNN is a result of two building blocks, convolution and fully connected block. The convolution block consists of pooling and convolution layers which are important for characteristic extraction. On the other hand, the utterly connected block consists of fully connected layer and performs classification primarily based on the inputs from the convolution layer. CNN in common has three layers, input layer, convolution layer and totally connected layer.

- **Input layer**

It is a layer where raw input data is loaded and stored for processing by the network. In the input layer, the dimensions of the input data are specified as combination of M by N and C.
where M represents the height of the input data, N represents the width, and C represents the Channel usually the color information included in the input signal.

- **Convolution Layer**

Convolutional Neural Networks are neural networks that automatically extract useful features without need of handcrafted methods from data-points like images to solve some given tasks. Therefore, prior to characteristic extraction, the data from the input layer is fed to Convolution layer where it is expanded with a filter to give convoluted output. Convolutional layers are the basic building blocks of CNN that transform the input data by using a patch of locally connecting neurons from the previous layer (Abdel-Hamid et al., 2014). The main function of the convolution layer is extracting features from an input data and considered as feature detector for CNN. To perform convolutional operations, it uses a filter function. Filters depict a particular feature of an image.

The convolution layer performs convolution operation on a specific matrix of input records the usage of a filter and it involves multiplying the values of a cell corresponding to rows and columns of an input matrix with the values of a cell in a corresponding filter matrix and sum up them to an output. At each time, the filter is convolved over the input matrix to constitute another matrix that consists of smaller facts than the input matrix. Convolution has two vital attributes kernel size and number of strides (Albawi et al., 2017).

**2.5.1.1. Hyper parameters of convolution**

Hyper parameters determine the spatial arrangement and size of the output volume from convolution layer. The commonly used hyper parameters are discussed below.

- **Kernel (Filter) size**

In convolution process of CNN, the convolution filter slides over all pixels of input matrix and take the dot product of the input matrix and the filter function. In this process, the combination of pixels convoluted by convolutional filter extracts important features from the
input data. Filters are functions with width and height smaller than the width and height of the input data.

**▪ Strides**

Strides specify how far the sliding window should move as per forward pass in filter function. At each time the filter function is applied, a new depth column is created in the output channel as per the number of strides. If the number of strides are lower, more columns in the output columns will be formed that creates more heavily overlapping receptive fields between columns that results in larger output columns in the output window. However, setting number of strides to maximum value will result in less overlap and smaller output volumes. Strides help CNN to decrease the number of learnable parameters. If we consider an input data of N-by-N dimension, filter size of F-by-F and strides of S then the output size will be as below.

\[ O = 1 + \frac{N - F}{S} \]

Where \( O \) is the output, \( N \) input size, \( F \) is filter size and \( S \) is the Number of strides.

Generally, strides tell us the number of pixels shifts over the input matrix. If the number of strides is one, the filter is moved by one pixel at a time. The figure below shows convolution with filter size of 3 by 3 and stride of one.

*Figure 2.4-2: CNN Convolution Operation*
### Padding

According to Albawi et al., (2017), in convolution stage of CNN there may be loss of information that found at the border due to the fact that this information does not have a chance to be seen when the filter slides over the input data. Then using a zero padding or filling a border element consisting of zeros will resolve the issue and manage output size. 

\[
O = 1 + \frac{N + 2F - F}{size, S \text{ strides } S}
\]

where \( O \) is the output, \( F \) is the Filter, \( N \) the input

Therefore, padding helps to prevent the network output size from shrinking in depth.

### ReLU Layer

ReLU (Rectified Linear Unit) layer applies an element wise activation function over the input data. When this function is run over the input data, it will change the pixel values of the input without changing the spatial dimension of the input data. ReLU helps to increase decision function and of the overall network’s nonlinear properties without affecting the receptive fields of the convolution layer thereby used to adjust or cut off the generated output (Albawi et al., 2017). In neural network, the weighted sum of every incoming connection for every node with in the layer is taken and the weighted sum is passed to an activation function. Therefore, the node output is denoted as:

\[
\text{Node output} = \text{Activation (Weighted Sum of Inputs)}
\]

ReLU is an activation function that performs some type of operation to transform the sum to some number between lower limit and the upper limit. Generally, \( \text{ReLU}(X) = \max (0, X) \) which implies that ReLU transforms an input to a maximum of either zero or the input itself. So, the more positive the input, the more activated it is. There are different activation functions used in addition to ReLU such as sigmoid and tanh. However, according to Albawi et al. (2017) recently ReLU is preferred because of the following reasons.

- When the planning of neural network is deeper the gradient signal begins to fade, which causes Vanishing gradient problem which causes problems in back propagation that went on due to the gradient of these functions is extremely on the brink of zero almost everywhere at the middle in
saturated function like sigmoid and tanh. But ReLU has constant gradient for positive inputs.

- The zero value within the gradient leads to getting an entire zero that leads ReLU to make a sparser representation. However, sigmoid and tanh always attains non-zero result from the gradient which may not favor training.

### Pooling Layer

This layer involves extracting a particular value from a set of values usually the maximum or average value to reduce the size of the output matrix. In this case, we will remove unwanted information from an input data and hence the basic function of pooling is down sampling to reduce complexity for further layers. It is inserted between successive convolutional layers to reduce the spatial size (Height, width) of the input data. Average pooling and maximum pooling are the two most commonly used pooling operations. Maximum pooling makes use of max () operation with 2 by 2 filter size and a stride of 2 which is the common pooling operation in neural network.

![Figure 2.4-3: Maximum pooling](image)

### Fully Connected Layer

This is the last layer of a neural network that performs classification functions. It consists of hidden layer and an output layer implemented using Softmax that perform classification among large number of categories. We use this layer to compute class scores that we will use as output of the network.
2.6. Support Vector Machine (SVM)

A support vector machine (SVM) is able to separate a group of input data into two categories with an optimal separating line. The algorithms aim is to separate the input patterns by maximizing the space between the closest vectors to the hyperplane and by separating them without error. It produces the pattern classifier by applying a spread of kernel functions like linear; polynomial functions because the possible sets of approximating functions (Belsti, 2020).

Even though, SVM was originally designed to handle two-class classification, it was later expanded into multiclass classification problems. Based on the sort of input patterns Different types of SVM classifiers are used: a linear maximal margin classifier is employed for linearly separable data, a linear soft margin classifier is employed for linearly non-separable, or overlapping, classes, and a nonlinear classifier is employed for classes that are overlapped also as separated by nonlinear hyper-planes (Tzotsos, 2005).

According to Tzotsos (2005), SVM for linear maximal margin classifier: used where the training data are often separated by a hyperplane, \( w \times x + b = 0 \). The goal of SVM is to seek out the optimal values for \( w \) and \( b \). After finding the optimal separating hyper plane, \( w_0x + b_0 = 0 \), an unseen pattern, \( xt \), are often classified by the choice rule for \( f(x) = \text{sign} (w_0x + b_0) \). Each unseen pattern, \( xi \), belonging because it does to at least one of two classes, features
a corresponding value $y_i$, where $y_i = \{-1, 1\}$. Because the hyperplane is $w \cdot x + b = 0$, the training data are often (doi.org) divided into two classes specified that:

$$W . x_i + b \geq 1, \text{if } y_i = 1$$

$$w . x_i + b \leq -1, \text{if } y_i = 1$$  \hspace{1cm} (2.7)$$

Linear soft margin classifier: handle input patterns that are overlapping or linearly none separable. Its objective is to separate the 2 classes of training data with a minimal number of errors. The linearly separable case within the above case (linear maximal margin classifier), are often rewritten as:

$$W . x_i + b \geq 1 - e_i, \text{if } y_i = 1$$

$$W . x_i + b \leq -1 - e, \text{if } y_i = -1$$  \hspace{1cm} (2.8)$$

Where, $e_i$, is non-negative slack variables.

**Nonlinear classifier:** kernel functions, like polynomial functions to transform the input space to a feature space of upper dimensionality, when the input vectors can't be linearly separated within the input space.

### 2.7. Audio feature extraction

Feature extraction is the method of removing unwanted and redundant information and retains only the useful information in automatic dialect recognition. But in practice this process may lose some important information. The goal of feature extraction is to seek out the set of properties called as parameter of utterances by processing of the signal waveform of the utterances. These parameters are the features. It produces a meaningful representation of speech signal. Feature extraction includes the method of converting speech signals to the digital form and measures important characteristics of signal i.e. energy or frequency and augment these measurements with meaningful derived
Measurements. Different feature extraction techniques are LPC, MFCC, LPCC, DWT, WPD, PLP, etc.

2.7.2. Mel Spectrogram

Feature extraction improves the accuracy of learned models by extracting features from input data and diminishing the dimensionality by removing redundant data.

As described by Wyse (2017), spectrogram is visual representation of the change in the frequencies of a signal such as audio as it varies in time. It is widely used in speech recognition, music genre classification, and linguistic analysis. It is a graph with two dimensions where the x-axis represents the time dimension and the y-axis represents the frequency dimension. It is produced from time domain signal either through band-pass filters or Fourier transforms. The following are the steps involved in creating the spectrogram of an audio signal.

- Separate to windows: - the input audio signal is first sampled at uniform window size and fixed hop length.
- FFT computation: - once the input signal is windowed to a fixed window size, each window is converted from time domain to frequency domain and this is achieved by FFT. Fast Fourier transform in general is a function that transforms an audio signal in time domain into frequency domain.
- Mel-scale generation: - The Mel scale in its mathematical term is the outcome of non-linear transformation of the frequency scale. It is constructed in such a way that sound or audio that are spaced on equal distance from each other on the Mel scale are also perceived equidistance from each other by humans. In this stage the whole frequency spectrum is taken and divided into series of frequencies on a Mel scale. The total frequency band that makes up an audio signal is taken and divided into evenly spaced frequencies consisting of fixed number of Mel. The frequencies are not divided evenly based on the distance on the frequency axis rather on the distance as heard by human ear.
• Spectrogram Generation: once the frequency spectrum at each window is separated by a Mel scale, the final stage will be decomposing the frequency at each window into its corresponding components. At this stage the magnitude of the audio signal found at each window is decomposed into its component frequencies in the Mel scale.

2.8.1. Dialect

Most researchers considered dialect to be accent but indeed it is different from accent. Dialect is defined as variations in some aspects of speaker’s pronunciation such as lexical, grammatical, phonological aspects. According to Maldonado (2015), dialect is described as a neighborhood or social range of a unique language unique by using the approach they speak. Dialect is a variety of speeches which is spoken through human beings in a particular society or geographic area. This dialect can be identified by using behaviors such as loudness, tonality and nasality in addition to pronunciation and phonemes of speakers. It is the representation of specific characteristics of speaker’s voice characteristics and gives necessary speaker information such as age, gender, fitness status and origin (A. Etman, 2015).

Even though dialects of the identical language share many similarities, they are frequently differentiated at several linguistic stages such as phonological (i.e., how it sounds), grammatical, orthographic (e.g., “center” vs. “centre”) and very frequently one of a kind vocabularies (Mohamed G. Elfekya, 2015).

Dialect identification (DID) is an automatic identification of the corresponding dialect of utterance, both written or spoken (Michon, 2018). It is identified as a different case of language recognition, which requires the capability to differentiate amongst extraordinary audio system in the equal language household but now not between specific language families. On the contrary side of language identification scheme, dialects share mutual phonetic listing and it also targeted on written language elements and linguistic elements such as words, characters and phonemes that could be utilized by means of computerized speech focus.

2.8.2. Applications of dialect recognition
Accurate identification of dialects has a top-notch position to enhance certain areas and services like Automatic speaker identification, HCI, E-health, E-learning, etc. It has also a notable function in Automatic Speech Recognition; it will allow the device to adapt its pronunciation, acoustic, and language fashions appropriately. It is also beneficial for figuring out a speaker’s regional origin and ethnicity and beneficial in speech-to-speech translation and forensic speaker profiling (Chittaragi, 2017).

2.9. Related Works

Two researches have been done in 2017 and 2018 on Amharic languages to design a model which identifies four Amharic dialects. In the study they used a Hybrid Approach of vector quantization (VQ) and Gaussian Mixture (GMM) and artificial neural network (ANN) Models to classify these dialects. Here Mel frequency cepstral coefficients (MFCC) feature vectors were used to recognize these dialects. Speech recording tools were used to collect the data sets in controlled environments. To have speech samples of different varieties, speakers are randomly chosen and included both sexes (male and female). MATLAB software was used for implementation. Based on the experiments they have done 92.7% and 95.7 accuracy achieved for the given dialects.

Even though, the authors have used VQ, GMM and a combination of two (VQ & GMM) to classify the dialects, VQ has limitations to recognize similar sounds due to VQ is template matching algorithm.

In 2017, a research conducted by Mengistu to model speaker identification system to identify Amharic language speakers in noisy environment was proposed. In the study Vector Quantization (VQ), Gaussian Mixture Model (GMM) and Combination of the back propagation neural network and Gaussian Mixture Models are applied for model implementation.

In addition to the above listed modeling techniques, feature extraction techniques such as Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Cepstral Coefficients (LPCC) and Gamma tone Frequency Cepstral Coefficients (GFCC) are used for extracting the useful features from the input speech signal.
In the study speech data used in the identification process consists of data collected from different speaker that includes a total of 300 speech samples each having duration of 10 seconds are collected. After the data is collected, the above-mentioned techniques are applied for modeling these data; as stated in the result section of the study, with the application of combination of GMM and Back propagation Artificial Neural Network an accuracy of 93.7% is achieved.

The other study conducted in the area of dialect recognition is Text Independent Amharic Language Speaker Identification (Mengistu & Alemayehu, 2017b). The aim of the study was to analyze performance of different methods for the development of Amharic speaker identification system. In this study the researcher applied diverse methods for recognition and feature extraction of Amharic speech data VQ, MFCC, GMM, GFCC and BPNN approaches are applied for the intended task.

The data used in the study consists of audio data having duration of ten seconds long collected from 90 speakers. The accuracy achieved in the study is 59.2%, 70.9% and 84.7% for VQ, GMM and BPNN and using the combined feature vector of MFCC and GFCC. But the total dataset that is used in the study is about 270 in total, which is small and the techniques used are template matching approaches. But recently deep learning and the hybrid of deep learning methods leads us to better performance.

Other related study also conducted by the same researcher (Mengistu & Alemayehu, 2017a), on Amharic language to identify the four known Amharic dialect. The authors in this study designed a model which identifies these dialects. Machine learning techniques such as Vector Quantization (VQ), Gaussian mixture models (GMMs), back propagation artificial neural network and the combination of GMM and back propagation neural network were used for classifying these dialects. Here Mel frequency cepstral coefficients (MFCC) feature vectors were used to extract important features to recognize dialects. The data sets were collected from speakers by directly recording speeches in controlled environments. This data includes both sexes (male and female). While recording speeches speakers are chosen randomly to have the speech samples of different variety. After collecting the data preprocessing takes place to avoid unimportant features such as silent the analog signal is collected to digital signal.
In this research a model that uses a tanh activation function have a better result instead of using the Logistic Sigmoid activation function in back propagation artificial neural network. 95.7% accuracy achieved when the hybrid approaches of GMM and back propagation artificial neural network with tanh activation function are used.

However, when we see VQ as a classifier to realize language dialects it has boundaries to apprehend a similar sound due to VQ is template matching algorithm. On the other hand, when used GMM is simple when applied for dialect classification tasks. Because of absence of memory to access information for next decisions, GMM do not provide information on the temporal relationship (A. Etman, 2015). Furthermore, GMM algorithm is not capable to work when the dimensionality of the problem is too high usually when it is greater than six and it becomes computationally inefficient. Additional downside of GMM is it requires customers to determine the quantity of combination fashions used by means of the algorithm while education. To discover the mixture mannequin that works higher for the classification task, users are predicted to retriever a number of numbers of mixture fashions seeing that they do not exactly understand how many mixture fashions are required. The data sets were recorded only in controlled environments. They did not handle the data sets which are recorded in uncontrolled environments and age is not considered as a criterion.

(Belayneh et al., 2021) conducted a study on artificial neural network based Amharic language speaker recognition. The aim of the study was to develop text independent Amharic speaker identification model that can identify the different Amharic speakers, Mel-Frequency Cepstral Coefficients are used as feature extraction techniques to extract features from audio signal.

In the study, 20 sampled speeches of 10 speakers’ which makeup a total of 200 speech samples are used for training and testing the speaker recognition model. In this particular study besides using MFCC as feature extraction technique, Artificial Neural Network is used to model the feature vector obtained from MFCC and to classify it into appropriate speaker class and test its performance. MATLAB which is regarded as 4th generation high level programming language is used for developing the overall implementation of the model. In general, three experiments with different setup are conducted and an accuracy of 96.0%, 96.7%, and 97.3% respectively is achieved for the three experiments. Depending on the
experimental result presented and conclusion given by the researcher, the study achieved promising performance.

Dialect recognition studies are also conducted in foreign languages, according to (Lei & Hansen, 2011); dialect classification via text-independent training and testing for Chinese, Spanish and Arabic is proposed. Diverse techniques such as different forms of Gaussian mixture model (GMM) such as Kullback-Leibler and FSD-GMM are applied in the study. In the study spontaneous speech recorded for different languages such as Chinese, Arabic and Spanish is used as a data source and classification model is developed using spontaneous speech obtained from such datasets. As shown in the result section of the study, both KLD-GMM and FSD-GMM techniques are applied on Spanish and Arabic speech data and achieved remarkable improvement in classification performance of dialect categories. As a concluding remark, the above listed methods (KLD-GMM and FSD-GMM) have improved the performance of dialect classification task for three-way dialect tasks.

2.10. Summary

In the above section, detailed discussion of related work in dialect recognition, speaker identification and language identification is provided with their corresponding accuracy level. As Mengistu (2017) stated, an accuracy of 93.7% is achieved for text independent speaker identification system for Amharic language in noisy environment by using different techniques. The data used in the study is 300 speech samples. But the data is not sufficient enough for the identification task and adding additional data will improve the accuracy further.

Text independent Amharic language identification in noisy environment is proposed and different techniques are applied for speech processing (Mengistu & Alemayehu, 2017b). The model has achieved an accuracy of 59.2%, 70.9% and 84.7%. In the study, the researcher collected the dataset used for the study from 90 speakers which is considered to be low. But adding extra dataset will advance the classification accuracy of the model, as revealed from the above research works, there is no speech dataset recorded in uncontrolled environment that may contain noise and pause between sounds and include different kinds of background
noise introduced either during recording or transmission. The amount of data used is too small and needs to be improved, on top of this the algorithm used has its own limitation to recognize speech dialects.

### Table 2.1-5: Summary of Related work

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Technique used</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abrham Debasu</td>
<td>Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC, MFCC and GFCC</td>
<td>LPCC, MFCC and GFCC</td>
<td>93.7% accuracy with combination of GMM and Artificial Neural Network</td>
</tr>
<tr>
<td>Abrham Debasu and, Dagnachew Melesew</td>
<td>Text Independent Amharic Language Speaker Identification In Noisy Environments using speech Processing Techniques</td>
<td>VQ, GMM, BPNN, MFCC, GFCC, and hybrid approach.</td>
<td>An accuracy of 59.2%, 70.9% and 84.7% is achieved</td>
</tr>
</tbody>
</table>
(Lei & Hansen, 2011)  | Dialect Classification via TextIndependent Training and Testing for Arabic, Spanish, and Chinese (doi.org) language | FSD-GMM and KLD-GMM are applied | It improved dialect classification performance for three-way dialect tasks

(Belayneh et al., 2021) | Developing text independent Amharic speaker identification model | Artificial Neural Network and MFCC | Accuracy of 96.0%, 96.7%, and 97.3% achieved.

Chapter Three

3. Dialect classification Design Methodology

3.1. Introduction

In this section, the proposed framework for text independent Amharic dialect identification model and the different tasks that constitute the entire task are discussed in detail. The different techniques that are employed to design dialect identification model are discussed and Tasks such as dataset preparation, preprocessing activities are included to transform the data into appropriate form for recognition, model development and evaluation procedures are presented in detail. The overall design of a better technique obtained after comparison between the different feature extraction and classification techniques are presented.
3.2. Proposed architecture

The proposed architecture shown in figure 3.2 has four components that include preprocessing, noise removal, feature extraction and classification. Each component is integrated and work together to discover recognized dialect category from a given Amharic dialect. The first component in the architecture is preprocessing that includes tasks such as audio format conversion, silence detection and removal and audio de-noising. The preprocessing component accepts continuous audio and convert format of the audio into wav format, and then removes silence present in it. The next component is noise removal that is intended for removing background noise found in the audio signal, where the noise present in the audio signal is removed by using cut off frequency of average moving filter which is low pass filter.

The other component in the architecture is audio feature extraction component that helps to capture the most distinctive part of an audio signal through feature extraction methods such as Mel spectrogram. It involves extracting Mel spectrogram images from the preprocessed audio and saving it as a form of image for the next recognition task. The main reason to use Mel-spectrogram over MFCC is that with lots of data and strong classifiers like CNN mel spectrogram performs better, MFCC performs better for linear models like Gaussian mixture model. On top of this Mel-Spectrograms are robust to additive noise as compared to MFCC(Demircan & Örnek, 2020). Since the audio data for this study is collected in uncontrolled environment which is full of additive noise, it is necessary to use techniques that are robust to additive noise.
In the course of audio feature extraction, spectrogram images are created and feed to the CNN model for image feature extraction and learning. But prior to image feature extraction, some image preprocessing operations such as size adjustment and normalization are performed on the spectrogram image. This involves adjusting the dimension of the spectrogram image into uniform dimension by resizing each image and then normalizing the
pixel values of an image between 0 and 1. After being normalized, spectrogram image is given to CNN architecture for feature extraction.

As shown in figure 3.2, the next step after feature extraction is classification of Amharic dialects into predefined class label. Classification involves two phases training phase and testing phase. Training phase consists of sequence of Convolution, Pooling, Activation and Dropout used for learning important local features from the given spectrogram.

Pooling layer will be added after the convolution one. A pooling layer is also arranged into feature maps of a number equal to those of the convolution layer, but with smaller maps. In the pooling layer, a sub-sampling will be done on the activations of the convolution layer to obtain new representations with a reduced resolution. This is because pooling layer leads to more efficient training by reducing the total number of trainable parameters.

Finally, SoftMax is used for classifying each Amharic sound into predefined dialect category such as Gojjam, Gondar, Wollo and Shewa. Since our task is multi-class problem, SoftMax that is differentiable allows to optimize a cost function to correctly classify into each predefined class is used for final classification. In this case a 4way SoftMax is used to classify Amharic dialects into predefined class labels.

3.3. Audio data acquisition

Audio data for each Amharic dialect category is acquired from Amhara Media Corporation. The corporation contains lot of diverse recorded audio and video data to be broadcasted to the public. People who speak Amharic language belonging to one of the four dialect categories may have an interview with the media journalists and their voice or speech recorded is stored in AMCO database or archive. When this study is planned, two options for data acquisition were considered (Recording Amharic speech by our-selves or using already recorded data from others). The first option is very expensive as it requires the physical involvement of several individuals who speak Amharic language belonging to one of the four dialect categories.
It is possible to get Amharic speech that belongs to the dialect categories from already recorded storage such as Amhara Media Corporation. The total data collected consists of 1200 utterance which has average duration of 20 second.

3.3.1. Audio Preprocessing

After collecting the input data, the next step is preprocessing since the speech signal contains many unnecessary things such as background noises, silent/unvoiced portions which reduce the recognition process so it has to be preprocessed first.

In dialects recognition, the first phase is preprocessing. First, the continuous dialect speech signal produced by the speaker and sensed by the microphone has to be converted to the discrete domain. Secondly, the audio signal is segmented into frames to get quasi stationary units of speech.

3.3.2. Silence Removal

In this process unvoiced portion removal along with endpoint detection is the fundamental step for dialects recognitions. These applications need efficient feature extraction techniques from speech signal where most of the voiced part contains speech or speaker specific attributes. Endpoint Detection, as well as silence removal is well known techniques adopted for many years for this and also for dimensionality reduction in speech that facilitates the system to be computationally more efficient. This type of classification of speech into voiced or silence/unvoiced sounds finds other applications mainly in Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, Stop Consonant Identification and End Point Detection for isolated utterances. As shown in the wave plot of the audio file given bellow, there exist a thin horizontal straight line at the beginning and the middle of the wave plot that represent the presence of silence in the audio signal. Therefore, pydub a python library is used to detect the presence of silence from this audio signal.
To do so the original audio signals is decomposed into chunks where the silence between audio tracks is 0.6 seconds and audio signals quieter than -26dBFS are treated as silence, therefore need to be removed from the signal. The figure below represents silence and signal part of a particular audio clip. The letter “S” represents the end points of a signal filled with silence and numbers such as 1 and 2 represent the actual audio signal where there is sound to be heard.

Audio signals having minimum silence duration above 0.6 seconds and silence threshold of 26dBFS are chucked into independent speech segments and exported to an independent directory. Audio data is split on the presence of silence with the help of pydub which is a python library to work with WAV files. Pydub is a python library that is used to split, edit or merge audio files, thereby in the presence of silence audio used for this study is split in the
presence of silence and merged again once silence is removed. The minimum silence duration to split audio file is set to 600ms, this is because most of the audio files have silence having duration of 600mili seconds. After splitting the entire audio file, each chunk is saved as a separate file.

Then later each chunk is read and merged based on their original sequence to makeup one independent full audio clip representing a particular dialect. The result after merging each chunked audio clip that is free from silence is presented in the figure 3.2-3.

![Audio Wave after silence removal](image)

**Figure 3.2-3: Audio Wave after silence removal**

### 3.3.3. Noise removal

In addition to silence there may exist different types of noise that are introduced into the audio signal during recording as a result of different factors such as temperature, noise from external environment and etc. these noise that make audio signal has to be removed in an appropriate way. This is because background noise introduced into an audio signal by any means will reduce the interpretability of audio signal and halt dialect identification process. Therefore, to remove the background noise moving average filter is applied.

Moving average filter is a low pass filter used for flattening an array of sampled data (Timbadia & Shah, 2021). It is considered as low pass filter used to remove unwanted noise
from the input data. It allows low frequency signals to pass and stop high frequency signals. Let the input data consists of N number of samples, then moving average filter takes N samples at a time and returns single output after calculating their average. To summarize, moving average filter performs the following functions. First the moving average filter takes N inputs then calculates the average of those inputs and returns a single output. As Timbadia & Shah (2021) stated, when the length of the filter increases the output smoothness increases but shrill change in data become rounded. From this it can be concluded that moving average filter has excellent time domain response but poor response in frequency domain. Given the input X and output Y, the general formula for calculating moving average filter is given below.

\[ Y(N) = \frac{1}{N} \sum_{k=0}^{N-1} x[n - k] \quad (1) \]

At each average filter, cut off frequency is used. The cut off frequency also called break frequency. It is a limit in the frequency response where the frequency of the signal to be processed begins to be reduced. It is the frequency where the rest above or below are permitted to stop or pass. This means if we consider an audio signal having a specified frequency and set a certain frequency as a cut off frequency, then if frequency of audio signal is above the cutoff frequency for moving average filter, then the signal will stop otherwise pass (Transform et al., n.d.).

In this study the cutoff frequency is set to 400, where audio signals below 400 are allowed to pass those above this border frequency is stopped from passing. Frequency ranges from 200 to 700 are tested as a cutoff frequency, but there is noise for higher cutoff frequencies ranging from 500 to 700. When the cutoff frequency is set lower than 400 some useful audio features are lost and the sound become unrecognizable.
As shown in the diagram above, the blue color represents audio wave before noise removal whereas the signal in the red color represents audio after noise is removed. This process is very useful for audio signals that are recorded in uncontrolled environment as background noise is inevitable when recording sound in such environments. All frequencies above the predefined useful frequency spectrum are stopped from passing.

After all the audio signals are preprocessed in an appropriate way, the whole audio signal is saved in one file and make up the dataset for the next step of the model development task. Then the next task is extracting useful features that are relevant for the dialect identification task from raw audio signals. Having many feature extraction methods, Mel Spectrograms are selected for this study.

3.3.4. Mel Spectrogram Generation

Spectrograms are graphic visualization of the spectrum of a signal as it changes with time. When a spectrogram of a specific audio data is converted into Mel scale it forms Mel spectrogram. To make audio data relevant for deep learning, it is converted into Mel spectrogram images. As stated in audio preprocessing section Audacity software is used to convert audio data from default mp3 into wav format. Then all the audio data is exported as wav file format and saved in a specified folder.
All the audio files are loaded or read from the folder by using Librosa which is commonly used python package for sound processing. When recorded the audio data has sampling frequency of 48000HZ, but Librosa down sample the original audio signal into its default sampling frequency of 16000HZ.

![spectrogram images](image1.png)  ![spectrogram images](image2.png)

- Gojjam
- Wollo

![spectrogram images](image3.png)  ![spectrogram images](image4.png)

- C) Gondar
- D) Shewa

Figure 3.2-5: spectrogram images

After feature extraction phase is performed the next step is the construction of appropriate model. There are a number of different calcification methods available for this purpose. However, in this research we used two supervised machine learning classification methods such as convolutional neural network and support vector machines for modeling.
1.

3.3.5. Image size adjustment

During the process of spectrogram extraction, the resulting images representing the spectrograms of an audio signal are saved as a form of image. But the image is high dimensional which makes it computationally expensive to process it. Therefore, the high dimensional image is resized into uniform and smaller dimensions (224x224), height of 224 and width of 224 respectively. This will reduce the computational cost while training and testing the model by reducing the pixel values the algorithm is working with.

3.3.6. Feature extraction

Dialect identification requires distinctive features of each sound wave pattern that comprise the dialect classes to be learned correctly. These features are extracted by a special network structure of convolutional neural network that comprises of series of convolutions along with the pooling layer for down-sampling the resulting feature vector.

Once the size of the image is adjusted to suitable dimension, the next step to be followed is extracting relevant spatial features that uniquely identify each spectrogram image from each other. The spatial features are relevant for capturing the different chromatic features that are available in the spectrogram image for expressing the different color values of the image. CNN is a very special network structure that is very powerful in extracting regionally applicable features from a spectrogram image.

To extract the distinguishing features of a spectrogram image, CNN is applied in this particular study owing to its better ability to automatically extract deep features from the image. It detects edges from raw pixel data, use these edges to detect shapes, and, use these shapes to detect higher level features. All these reasons make CNN a particular choice for extracting useful features from a spectrogram image. CNN extract features by convolving the image with different filter size and down sampling the dimension of the image using pooling layer. Several Convolutional and pooling layers are used to extract useful features from the image.
3.3.7. Feature learning and Classification

Feature learning phase in this case involves different layers of CNN arranged on top of each other. Convolutional Neural Network (CNN) is used for extracting distinctive features from an image and performs classification of spectrogram images to predefined dialect category. To perform classification, the characteristic properties of a spectrogram image need to be learned during the feature learning phase.

Features of the corresponding image are learned during the training phase of the model development process. In this phase the model is trained by feeding locally relevant features of an image that are extracted by series of convolution layers as input to the network structure. The training phase includes different CNN layers such as the convolution layer, activation layer, and pooling layer for learning the most distinguishing features from the image. The testing phase has similar procedures with the training phase but with a different dataset from the training phase.

Once the learning phase is complete, the next stage is classifying each spectrogram image into corresponding dialect category based on the learned features. Therefore, each input audio sample is classified into one of the four dialect categories as Gojjam, Gondar, Shewa and Wollo. For this study a total of four (4) classes are used and all audio samples are mapped to one of these classes.

3.3.8. Feature learning phase

Feature learning phase includes sequence of successive convolution, activation, pooling, dropout and batch normalization layers that are used to detect the relevant features from spectrogram image.

Convolution Layer

As shown in the proposed architecture, there are three convolution layers used for learning the relevant spatial features from the input speech signal. The input to the first convolution layer has a full shape of (224, 224, and 3). Successive convolution operation is applied to the input by using the filter kernel to obtain the most representative features that are used as an input to the next convolution process. The input image is convolved by a filter function and important
features in the region of an image are captured forming a feature map. During the next convolution process, the feature map discovered in the previous phase is used for performing the next convolution. Parameters such as number of filters, filter size, padding, stride size and activation function are considered in each convolution operation. We used 32 filters consecutively from the first layer to the second convolution layer. Each filter has filter size of 3 and stride of one. The four parameters are adjusted during model construction, where number of filters represent and control the depth of the output volume, filter size determines the size of each kernel, number of strides determine the number of pixel positions skipped both horizontally and vertically while performing convolution operation and padding determines the output volume.

We used smaller filters at the first stage and increased progressively down to detect as many features as possible. We have used stride size of 1 so that none of image regions are skipped during convolution operation. Same padding is used so as to preserve the size of the input matrix. ReLU activation function is used at each convolution layer. After applying convolution operation, pooling operation is applied to reduce the feature map for computational purpose. So, Max-Pooling with pool size of 2 by 2 is used. Once Max-Pooling is applied in each convolution layer, the pooled features are used as input to the next convolution layer.

**Activation Layer:** We have used ReLU activation function in the activation layer throughout the convolution process. ReLU returns zero if the value of the input is negative, otherwise the current value is returned. It is used to introduce non-linear mappings from input to output that helps CNN to learn and model more complex functional mappings that existed in the input features and hence improved the accuracy of the model.

**Pooling Layer:** Once the important features are obtained, there is a need to down-sample the size of the input volume. Pooling layer is applied between successive convolution operations to reduce the size of the output volume. While reducing the data volume, the pooling layer will preserve important information from an input image. Two parameters are required by the pooling operation, pool size and strides. Pool size helps to determine by how much we want to reduce the input volume dimension thereby controls the window size the
pooling operation is applied. Stride size on the other hand is used to determine how much pixels need to be skipped while performing pooling operation.

In designing the model, we used Max-Pooling that selects the maximum value among input features after each convolution layer to reduce the size of the input volume. The pooling layer can reduce the data layer while saving feature information. Throughout the convolution stage of the proposed model, we have applied a pooling size of 3 to reduce the dimension of the input volume by 3 and stride size of 2 to instruct the model to skip 2 pixels horizontally and vertically when performing pooling operations.

**Dropout Layer:** After each pooling operation, we have applied dropout layer with dropout probability of 0.3. It is a method of regularization that helps to reduce overfitting by explicitly altering the network architecture at run time. It is a method of disconnecting inputs randomly from the previous layer to the next layer. Random disconnection of inputs is used to make the CNN architecture independent of any single node.

**Flatten Layer:** During convolution and pooling operations, important features are extracted from an image. The next stage after feature extraction is to use extracted features to classify each image into predefined class label. During classification stage SoftMax classifier is applied, but SoftMax classifier is only one (linear). Flatten layer is applied to compute the final output probabilities for each class before applying the Softmax classifier. The classification is carried out in the fully connected layer. Since the flattening layer accepts only one-dimensional data, the 2D data from convolution and pooling operation must be converted into one-dimensional data. Softmax classifier holds four (4) nodes which correspond to the number of classes we are working on.

**Batch Normalization:** we applied batch normalization immediately after the final convolution layer. It is used to normalize the activations of the given input volume before passing into the next layer. It helps to make the loss and accuracy curve to be more stable.
Finally, the flatten layer merges all the features and provided them to the SoftMax classifier. (The output of the final flatten layer is given as input to the SoftMax classifier). A 4-way SoftMax that corresponds to the number of classes is used for recognizing the four specific Amharic dialects.

### 3.3.9. Validating the model

Validation data is a set of data separated from the training set to be used to validate the model during training. This validation process gives information that may assist us with adjusting our hyper parameters. It helps to ensure that the model is not overfitting or underfitting to the data in the training set. 20% of the dataset is managed as validation set used to validate the model. So, the model is validated in the validation data set aside from training set.

#### 3.3.9.1. Testing phase

The proposed model is developed during the training phase, and the performance of the model should be evaluated on testing data to ensure to what extent it works correctly on unseen data. Therefore, testing phase involves measuring the performance of the developed model by...
exposing it to new data than the one used in the training phase. This is done by using testing
data that is set aside for the evaluation purpose. This phase is done in similar manner to that
of the training phase with all preprocessing, feature extraction and classification steps
performed in the same way. One thing that distinguish testing phase from training phase is the
size and the fact that the data is not included in the training set. Therefore, twenty percent of
the total dataset is used to evaluate the performance of the model and measure the extent to
which it works.

The proposed architecture above shows the schematic diagram for recognition process. In this
process the first thing is preparing the input signal. In this case the relevant datasets are
collected from both controlled and uncontrolled environments in Amhara media corporation
archive system where programs, news and entertaining events are stored. For each dialect 300
data are collected which is a total 1200.

After collecting the data, the next step is preparing the data to have the same sampling
frequency and bit. The default sampling frequency of the recorded audio is 48000 Hz, in order
to decrease the dimensionality of the audio data for computational purpose, each audio data is
resampled to 16000Hz.
Chapter Four

4. Experiments, Results and Discussion

4.1. Introduction

In the previous chapter, the overall design of the proposed model and its detailed architecture are discussed. In this chapter we will discuss the implementation and experimental evaluation of the proposed Amharic dialect identification model. Furthermore, the experimental details such as dataset used for training and testing the proposed model, implementation tools used, and evaluation and test results are presented in detail.

4.2. Dataset description

For speech based Amharic dialect identification task, the first requirement is collecting audio data comprising of the different Amharic dialects. These audio data need to be collected from appropriate source. In this study speech data recorded from uncontrolled environment is required.

Amhara Media Corporation broadcasts programs and news in diverse languages such as Amharic, English, Arabic, Awi, Affan Oromo, Tigrigna and other languages. Broadcasted programs and news consists of diverse speech recorded from individuals from various geographic locations and ethnic groups. When preparing different programs in Amharic language, various individuals from different parts of the country speaking Amharic language belonging to the different Amharic dialect categories are invited, interviewed, and their speech recorded to be broadcasted to the air. All the audio and video data are archived and made available to the public through their TV and YouTube channel.

Therefore, Amhara Media Corporation is considered as our source of data for this study. It contains archived Amharic audio and video data consisting of speech recorded in uncontrolled environment which is one of the main reasons that attract our attention, as our focus is building speech based Amharic dialect identification model by using audio data in uncontrolled environments. During data collection, different groups of individuals from both sexes (Male and Female) who are at different age group are considered.

Thus, a total of 300 utterances of audio data that belongs to the four Amharic dialect categories (Gojjam, Gondar, Wollo, Shewa) that are being spoken in different parts of Amhara region are collected from Amhara Media Corporation (AMCO) archive system. A total of 1200 utterances of audio data which has duration of 20 second is collected, and then preprocessed in an appropriate way
to make it suitable for the dialect recognition. First the speech is edited and preprocessed by audacity software and its sampling frequency is set unaltered and each audio clip is converted from .mp3 ,.wma format into .wav and exported as .wav file format. This is done because python is really good in processing .WAV file formats for audio. The audio file then undergoes preprocessing operations such as silence detection and noise removal to improve the quality of the data for the next recognition tasks. Once the audio is preprocessed, features are extracted from it. To do so, appropriate feature extraction mechanisms such as Mel spectrogram is used. Therefore, Mel spectrogram is extracted from each audio sample and the spectrogram is saved as .jpg image format.

**Experimental environment setup**

To evaluate the detection model, we conduct the experiment on TensorFlow 2.3, Python 3.8 framework and use Scikit-learn, NumPy, Pandas, and Keras and other python packages. Our experiments are performed on windows 10 systems equipped with 8GB RAM, 1600 MHz DDR3, Intel® Core™ i7-8250U CPU.

To develop the model and evaluate it, we used Intel® Core™ i7-8250U HP computer with processor speed of 1.8GHz and 8GB RAM with windows 10 operating system. The implementation is done on python environment with Keras front end and Tensor Flow backend. For writing the code for implementation of the proposed model, we used Pycharm and Jupiter Notebook editors. We used python 3.8 and Tensor Flow 2.3 to conduct the simulation of the model with the addition of other packages.
Table 4.3-1: Experimental environment setup

<table>
<thead>
<tr>
<th>Tools used</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>8GB RAM, 1600 MHz DDR3, Intel® Core™ i7-8250U CPU</td>
</tr>
<tr>
<td>Programming language</td>
<td>Python</td>
</tr>
<tr>
<td>Libraries</td>
<td>Keras (Tensor Flow 2.3.0 as a backend) and Open CV are used, Moviepy, Librosa</td>
</tr>
<tr>
<td>Epoch</td>
<td>50 epochs</td>
</tr>
<tr>
<td>Batch Size</td>
<td>32</td>
</tr>
<tr>
<td>Dataset partition</td>
<td>80% Training, 20% Testing set</td>
</tr>
<tr>
<td>Editors</td>
<td>Jupyter notebook, Pycharm editor</td>
</tr>
</tbody>
</table>

Amharic audio data is required to conduct the experiment for Amharic dialect identification model by using the above-mentioned setup and experimental tools. Therefore, Amharic audio data that consists of equal amount of audio from equal number of speakers each belonging to the four Amharic dialects is collected from Amhara Media Corporation archive system. The dataset consists of audio data collected from both male and females who are at different age group. The percentage distribution and number of audio data collected is presented in table 4.3-2.

Table 4.3-2: Audio data Distribution for each Dialect

<table>
<thead>
<tr>
<th>Dialect category</th>
<th>Number of speakers</th>
<th>Total audio data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Gojjam</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Gondar</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Wollo</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Shewa</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Total Audio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3. Evaluation Metrics

In general, five separate experiments are conducted by using generated spectrogram images of an acoustic speech signal that represents audio containing the different Amharic dialect categories. The first experiment is conducted using CNN-SVM technique, the second experiment is conducted using CNN algorithm. The other three experiments are performed by using state of the art deep learning models with the purpose of comparing their performance with CNN and CNN-SVM models.

To conduct the experiment the available data which is 1200 utterance of speech data is partitioned into samples by following a train-test split of 80%, 20% of training and testing sets respectively. After proposed a model, experiments are performed to classify Amharic dialects to their dialect category. We used accuracy, recall, precision and loss as measure of performance metrics to see the classification performance of the proposed model. Accuracy of the model describes the number of correctly classified data instances per the total training instances. According to Martin & Powers (2015), precision shows the amount of predicted positive cases that are truly positive and Recall is the amount of actual positive labels that are appropriately labeled as positive. F-score is defined as the harmonic mean of precision and recall. Mathematically, precision, recall, accuracy and F-score are calculated as follows.

- Accuracy = \( \frac{TP+TN}{TP+FP+FN+TN} \)
- Precision = \( \frac{TP}{TP+FP} \)
- Recall = \( \frac{TP}{TP+FN} \)

\[ F1 \text{ Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \]

\( F1 \) Score = 2*(Recall * Precision) / (Recall + Precision) where: TP (True Positive) represents positive tuples that were correctly classified by the model, TN (True negative) represents negative tuples that were appropriately classified by the model, FP (False positive) shows negative tuples that were misclassified as positive and FN (False Negative) refer to positive tuples that were classified as negative. Once we specified the performance metrics for the model, the actual experiment is conducted by using Amharic speech prior to the application of
different preprocessing operations on the raw input signal, and then the proposed model is evaluated by comparing its performance with state-of-the-art architectures.

The experiment for Amharic dialect identification model is performed in two ways, spectrogram images extracted from each audio signal corresponding to the four dialect categories are fed to either CNN or CNN-SVM model, finally the performance of the two approaches are compared and the model that performed better in identifying Amharic dialects is adopted as a base model for Amharic dialect identification.

4.4. CNN-SVM Dialect Identification Model

Support Vector Machine (SVM) is machine learning technique mostly applied for classification and regression problems which handles both continuous valued and categorical variables (Tzotsos, 2005). It is one of the supervised classification techniques used for classifying data into set of classes (Mathur, 2004). The main aim of this study is to classify each audio consisting of Amharic dialect to the correct dialect category.

The model consists of CNN layer used for extracting features and SVM layer used for classifying extracted features into corresponding dialect categories. The CNN layer consists of four convolution layers followed by pooling layer and dropout layer. The input layer provides input images to the next convolution layers. The shape of the input image is 224, 224, 3 obtained from the preprocessing step. The convolution layer consists of 64, 32, 32 and 16 neurons and filter size of 3 by 3. The network consists of padding and RELU activation function.

The convolution layer learns the features of an image, by applying sequence of convolution and pooling layers. Features are extracted using the convolution layer and down sampled applying 2 by 2 pooling layers and ReLU activation function. Finally, the dense layer reduces the dimensionality of the features extracted through the convolution layer and the features are then saved to file and used as input for the next classification process. SVM uses features that are saved from the final dense layer of the convolution process. During SVM training, features are extracted using Convolution and dense layer of the CNN network, and then the training data is feed to the SVM network for classification.

```python
#Extract Features for SVM
model_feat = Model(inputs=model.input, outputs=model.get_layer('dense_1').output)
feat_train = model_feat.predict(trainx)
```
The dataset is first divided into training and testing during data set preparation stage with train test split of 80 percent of the data for training the model, 20 percent for testing the model.

Figure 4.4-1: Confusion Matrix for CNN-SVM

Figure 4.5-1 shows the confusion matrix of the testing phase of CNN-SVM model where the number of correctly and incorrectly classified samples are reported. It indicates how many data samples are correctly classified and incorrectly categorized.
Figure 4.6-2 shows the classification report of CNN-SVM network. It is achieved the overall accuracy of 82%.

4.5. CNN Dialect Identification Model

CNN is supervised deep learning algorithm widely used in image processing and computer vision (Gales, 2015). Recently it is also applied in speech recognition and achieved promising results. In this study we plan to see the applicability of CNN for developing dialect identification model.

Before performing the actual dialect, processing operations using deep learning algorithms, it is very important to load the necessary packages to perform the intended task. Therefore, all the necessary python packages such as NumPy for numerical manipulation, Matplotlib used for visualization purpose, TensorFlow and Keras that are used for developing deep learning models are loaded first. Not only these, to accomplish the overall task there also other packages that are imported.

Once the required libraries are loaded, the next step to be performed is loading spectrogram images from the directory. After being imported, some preprocessing operations are
performed on the input spectrogram image. Preprocessing operations include image resizing for adjusting the image uniform dimensions (width and height), rescaling the image to common scaling factor (between 0 and 1). To transform into uniform dimensions, the image is adjusted into to height of 224 and width of 224 which results in 224x224 pixel image. This resizing serves the purpose of down sampling the image to reduce computational cost. After resizing, the data in Mel spectrogram image is normalized into an integer having a value between 0 and 1.

When developing CNN dialect classification model, different combination of convolution and pooling operations are used and optimal performance is achieved when using four convolutions each with 64, 32, 32 and 16 neurons with 3 by 3 kernels and three two by two maximum pooling layers followed by two dropout layers with 0.3 dropout rate. In each convolution layer ReLU activation function is used except the last SoftMax layer. Finally, a four-way SoftMax layer is used to map each input data into the corresponding class label. The four convolution layers are designed to extract locally relevant texture features from the input data. In the experiment Adam is used as optimizer and compile and sparse categorical cross entropy as loss function. Same padding is used and the model is trained for 50 epochs, each sample has a batch size of 32 and learning rate of 0.0001.
**Figure 4.5-3:** the architecture of CNN model

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output Shape</th>
<th>Param #</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv2d (Conv2D)</td>
<td>(None, 224, 224, 32)</td>
<td>896</td>
</tr>
<tr>
<td>max_pooling2d (MaxPooling2D)</td>
<td>(None, 112, 112, 32)</td>
<td>0</td>
</tr>
<tr>
<td>conv2d_1 (Conv2D)</td>
<td>(None, 112, 112, 32)</td>
<td>5248</td>
</tr>
<tr>
<td>max_pooling2d_1 (MaxPooling2D)</td>
<td>(None, 56, 56, 32)</td>
<td>0</td>
</tr>
<tr>
<td>flatten (Flatten)</td>
<td>(None, 100352)</td>
<td>0</td>
</tr>
<tr>
<td>dense (Dense)</td>
<td>(None, 128)</td>
<td>12045184</td>
</tr>
<tr>
<td>dense_1 (Dense)</td>
<td>(None, 4)</td>
<td>516</td>
</tr>
</tbody>
</table>

Total params: 12,855,844  
Trainable params: 12,855,844  
Non-trainable params: 0
Training of the model in this case involves series of convolution layers for extracting locally relevant features from raw audio signal and pooling layers for down-sampling the information. In this method the CNN model by itself extracts relevant features from Mel spectrogram image using sequence of convolution layers. The training consists of series of convolution, pooling, activation and fully connected layers arranged in a sequential order. The experiment is conducted by using total audio dataset of 1200 utterance that belong to 4 classes. Then the dataset is divided into 80%, 20% of training and testing set. A total of 300 Mel spectrogram images per class which make up 960 images and 240 images are used for training and testing the proposed model. It is trained for 50 epochs and achieved a training accuracy of 98.20% and validation accuracy 78%.
As shown in figure 4.5-4, the CNN model is trained and validated on 960 and 240 Mel spectrogram images respectively. The training accuracy and validation accuracy keep increasing progressively with epoch number, the loss on the other hand keep decreasing.

To achieve the above performance, different techniques like dropout are applied to mitigate over fitting problem. As shown in 4.7-2, the model has good performance resulting in lower gap between training and validation accuracy and loss that clearly shows model is not suffered from either under fitting or over fitting. Accuracy, recall, precision and F1-score are Evaluation metrics that are used to evaluate the performance of the model.

![Figure 4.5-5: Accuracy to Loss Curve](image)

As shown in the figure 4.5-5, representing the accuracy to loss curve of the CNN model, it is revealed that when the training accuracy increases, so does the validation accuracy and the
same trend is perceived in training and validation loss of the model. This shows that the model is performing well and do not expose to either over fitting or under fitting.

![Confusion Matrix of CNN model](image)

**Figure 4.5-6: Confusion Matrix of CNN model**

Figure 4.5-6 shows the confusion matrix of the testing phase of CNN model, which shows the number of correctly and incorrectly classified samples are. It indicates how many data samples are correctly classified and incorrectly categorized.
4.6.1 Comparing CNN dialect identification model with state-of-the-art models

The proposed speech based Amharic dialect identification model is evaluated by comparing its performance with state-of-the-art architectures such as VGGNet, AlexNet and LeNet. The selected state of the art architectures is evaluated using the same data, number of epochs, similar experimental environment with the base speech based Amharic dialect identification model. Similar evaluation metrics such as accuracy, precision, recall and F-measure are used during evaluation procedure and the value associated with state-of-the-art models is compared with the proposed model. These state-of-the-art models are image recognition models, as stated in the preceding section of this study.

4.7.2 VGGNet model

This is one of the famous convolutional neural network models in deep learning (Guan et al., 2019). It is applied for Amharic dialect identification keeping its normal architecture consisting of 13 convolution layers, 5 pooling layers and 3 fully connected layers. It is trained
for 50 epochs using similar dataset used to develop our model and an accuracy of 24% is achieved as shown in the figure 4.6.1-1

![Image showing training results](image.png)

**Figure 4.6.1-1: VGGNet Model Training**

![Image showing training and validation accuracy and loss](image.png)

**Figure 4.6.1-1: Training and Validation accuracy and loss**
4.7.4. ALexNet model

Alex Net model is one of the famous convolutional neural network models in deep learning that is applied for recognition of large datasets (M et al., 2020), (Sutskever & Hinton, 2012). It is considered to be the most successful CNN architecture and it is applied for Amharic dialect identification keeping its normal architecture consisting of 5 convolution layers, 3 pooling layers and 3 fully connected layers. It is trained for 50 epochs using similar dataset used to develop our model and an accuracy of 24% is achieved as shown in the figure 4.6.1-2.

![Figure 4.6.1-2: Training of ALexNet network](image)
4.7.5. LeNet model

LeNet model is one of the famous CNN models in deep learning proposed by (M et al., 2020). It is applied for Amharic dialect identification keeping its normal architecture consisting of 2 convolution layers, 2 pooling and fully connected layers. It is trained for 50 epochs using similar dataset used to develop our model and an accuracy of 26.6% is achieved as shown in the figure 4.6.1-6.
After training for 50 epochs, LeNet model achieved training accuracy of 26.6% and validation accuracy of 21.30%. When compared with the CNN model, it performs poorly that implies the CNN model is the best model chosen.

Figure 4.6.1-6: Training of LeNet Model
Figure 4.6.1-7: training and Validation accuracy and loss curve

LeNet model is evaluated on 240 testing samples and achieved testing accuracy of 24%. Compared with the proposed model, the testing accuracy of LeNet model is lower this implies that the proposed model is performing better than LeNet model.

As shown in the diagram above, the loss varies for the first 15 epochs and it goes uniformly beyond the 15th epoch. Validation and training accuracy of the model keep increasing progressively for some epochs and both become uniform showing no wide gap between the training and validation accuracy and so did the training and validation loss of the model. This shows that the model does not be exposed to either over-fitting or under-fitting problems.
4.8. Summary of comparison of Models

To evaluate the performance of the proposed Text independent Amharic dialect classification (TIADI) model, comparison is made between the proposed model and available state of the art models. When compared with other techniques the proposed model has achieved a training accuracy of 98%, test accuracy of 85% when RELU activation function is used, Recall of Precision of 85, and F1-score of 85. From the experimental results obtained above, it is possible to generalize that the proposed model achieved better accuracy, precision, recall and F1-score measures and signifies it performs better than other techniques.

Table 4.3-3: Comparison of Amharic dialect identification Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-SVM model</td>
<td>90%</td>
<td>82.4%</td>
<td>83</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>CNN Model</td>
<td>98%</td>
<td>85%</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>LeNet model</td>
<td>26.60%</td>
<td>21.30%</td>
<td>21</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>VGG Net model</td>
<td>26.53%</td>
<td>20.73%</td>
<td>21</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Alex Net model</td>
<td>26.79%</td>
<td>20.83%</td>
<td>20</td>
<td>19</td>
<td>20</td>
</tr>
</tbody>
</table>
4.9. Result and Discussion

In this study speech based Amharic dialect identification model is proposed and its performance is evaluated on test data. The model development process consists of activities that are intermediate to the final result. These activities are aimed at accomplishing different tasks of the dialect identification process. These tasks include audio preprocessing, Mel spectrogram generation, image adjustment and model building. To develop dialect identification model raw spectrogram images are used and two techniques such as CNN and CNN-SVM are applied. Finally, comparison between CNN and CNN-SVM is made, CNN out performed CNN-SVM. The accuracy of CNN model is compared with state-of-the-art models and it achieved better recognition accuracy.

The results obtained in the study were presented in the previous section. As discussed CNN-SVM model achieved a training accuracy of 90% and a test accuracy of 82% and CNN based dialect identification model achieved a training accuracy of 98% and test accuracy of 85%. As shown in the result section of the study, CNN approach outperforms CNN-SVM approach with recognition accuracy. This is because CNN has good ability to learn features from raw input images (Abdel-Hamid et al., 2014).

Prior to model development, different speech and spectrogram image processing operations are performed on the raw image and audio signal; these include using moving average filter for noise removal, thresholding method for removing signals having energy below a certain threshold, image size adjustment and normalization are among the many operations performed. Silence part of an audio signal contains quieter waves with lower energy than the corresponding voice audio. The quieter part has nothing to do for dialect recognition task as it may degrade the recognition rate and quality; thus, applying thresholding removes signals below certain energy threshold.

In addition to silence, the speech data collected from uncontrolled environment contain different types of noise that may halt dialect recognition. Moving average filter is applied to remove audio signals that exhibit properties of noise.

Once the audio data is preprocessed in an appropriate way, the following task is converting raw audio signals into Mel spectrogram images. To convert audio into spectrogram images,
the audio data is loaded by using librosa a python library and further processed to make the analysis work easier.

For CNN model we have used different combination of convolution and pooling operations and optimal performance is achieved when using four convolutions each with 64, 32, 32 and 16 neurons with 3 by 3 kernels and three two by two maximum pooling layers followed by two dropout layers with 0.3 dropout rate. In each convolution layer ReLU activation function is used except the last SoftMax layer. Finally, a four-way SoftMax layer is used to map each input data into the corresponding class label. The four convolution layers are designed to extract locally relevant texture features from the input data. In the experiment Adam is used as optimizer and compile and sparse categorical cross entropy as loss function. Same padding is used and the model is trained for 50 epochs, each sample has a batch size of 32 and learning rate of 0.0001. The accuracy of the different methods is discussed as shown in figure 4.9-1 below. The training accuracy of CNN model is 98% with testing accuracy of 85%.

The second experiment is performed with CNN as feature extractor and SVM used for classifying extracted features from spectrogram images into corresponding class labels. The CNN in this case includes the same convolution layers that include 64, 32, 32 and 16 neurons with 3 by 3 kernels having 2 by 2 strides. After extracting locally relevant features from the input spectrogram images, then the features are flattened to convert it into one dimensional feature vectors so that it can be suitable for classification task. In the experiment ReLU activation function is applied in each convolution operation. A dropout probability of 0.3 is used and batch size of 32 is applied and trained for 50 epochs. Then extracted features are flattened into single dimensional feature vectors and final fully connected layer is used for predicting the input feature maps. Extracted features are then fed to SVM network to map each input feature into their corresponding class label. This model achieved training accuracy of 90% and test accuracy of 82%.

In the proposed model learning rate ranging from 0.0025 to 0.0001 are tried and learning rate of 0.0001 has achieved better recognition accuracy. ReLU and SoftMax activation functions are used for convolution, hidden layers and final classification layers of the model. As shown in the figure below the accuracy of CNN model has achieved better accuracy as compared to the other models. CNN model has been tested with 240 spectrogram images for the purpose of
evaluating the model and has achieved a test accuracy of 85%, when compared with CNN-SVM model.

Additionally, the performance of the state-of-the-art models such as LeNet, VGGNet and AlexNet are tested on spectrogram images. But the performance of these models is extremely lower than CNN and CNN-SVM models with their recognition accuracy. It’s training and testing accuracy is lower, this is because state of the art model is trained on large amount of dataset usually millions of datasets. But the model in this study is trained using 1200 audio data which are smaller as compared to the data used to build these state-of-the-art models. In addition, deep learning models perform better when trained on large dataset. In this study the architecture of the state-of-the-art models is used and the 1200 spectrogram images are used in this study to evaluate the performance of those state-of-the-art models.

![Training Accuracy vs Testing Accuracy](image)

Figure 4.6.1-8: chart for accuracy of different models
4.10. Summary

In this chapter, the experimental evaluation of the proposed architecture for Amharic dialect identification model is discussed. The dataset used for training and testing the proposed model is discussed in detail. Different experiments are conducted using the dataset collected and the performance result for each experiment is presented. Comparison is done between the different techniques for Amharic dialect identification and state of the art CNN architectures. The comparison result has revealed that the proposed CNN Text independent Amharic dialect identification model has performed better than other techniques and it is selected for Amharic dialect identification task.

Prior to actual model development task, different preprocessing operations are performed on the input of Amharic speech representing individual dialect category. The overall task is started by converting the audio format that allows mapping all audio signals into a common format. Silence removal is also performed to remove unvoiced part of the audio. Following
silence removal, the audio signal pass through noise removal operations to remove noise introduced into the audio signal during recording where moving average filter is applied to simplify this task.

Once the audio signal is preprocessed in an appropriate manner, the next task will be to convert the audio signal into condensed representation using spectrogram images. Therefore, each audio signal is converted into spectrogram image and some image preprocessing operations such as image resizing and normalization are applied on the spectrogram image. Then appropriate features are extracted from spectrogram images by applying CNN thereby all the identifying features are learned by applying sequence of convolution and pooling operations.

During the classification stage Adam is used as optimizer, ReLU as an activation function and SoftMax is used as a classifier. Learning rate between 0.0025 and 0.0001 are tested and better recognition performance is obtained when ReLU activation function, SoftMax and learning rate of 0.0001 are combined together. Later the performance of text independent Amharic dialect identification model is compared with both state-of-the-art techniques and registered better recognition performance. This achieves a training accuracy of 98% and testing accuracy 85%.
Chapter Five

Conclusion and recommendation

5.1. Conclusion

There are so many languages in the world to share ideas, feelings and opinions between individuals. Amharic is one of these languages with relatively large number of speakers next to Affaan oromo. It is spoken in different parts of the country and has different dialect categories such as Gojjam, Gondar, Wollo and Shewa that differ in some way from each other such as speaking pattern, pronunciation and grammar with the dialect of Addis Ababa as standard way of speaking (Derb et al., 2019). Developing Amharic dialect identification model which automatically identifies Amharic dialects serve different purposes. It has greatest advantage in speech recognition, speaker identification, and also assists in crime investigation.

In this study an attempt is made to develop a model that can recognize and identify the different Amharic dialect categories. The study is undertaken with the goal of developing text independent Amharic dialect identification model using data from the four Amharic dialect categories and different techniques are applied to achieve better recognition accuracy. Two separate experiments are done by using one deep learning (CNN) and hybrid approach (CNN-SVM) algorithm and the performance of the two techniques are compared and the model having better performance is kept as a baseline model and its performance is compared against state-of-the-art models. Speech features extraction in both experiments is performed by using CNN.

Prior to actual model development different tasks, that are intermediate steps to the final result are performed with the aim of improving the quality of data and accuracy of the model including tasks like audio preprocessing activities such as audio format conversion and silence removal, audio de-noising using moving average filter for removing the background noise introduced into the audio signal either during recording or at different time, audio feature extraction methods such as Mel spectrogram generation performed with the aim of extracting the most relevant and salient audio features and convert them as spectrogram image for further analysis. In addition, image size adjustment and normalization are performed to reduce the information space and computational cost during model development.
Dialect recognition may not be effective when using inappropriate data for modeling. In this study the data to develop the dialect recognition model is collected from uncontrolled environments. The uncontrolled data consists of noise introduced into the signal and will reduce the performance of dialect classification techniques. Therefore, the noise should be removed from the audio in order to improve the recognition performance.

Finally, classification is done using CNN and CNN-SVM, their performance is compared and CNN model outperform CNN-SVM model and retained as a base model for dialect recognition task. The performance of speech based Amharic dialect identification CNN model is compared with state-of-the-art models such as VGGNet, AlexNet, LeNet and it achieved better recognition than all state-of-the-art models considered.

State of the art models are trained on large dataset (consisting millions of data) collected for particular purpose and achieved better recognition accuracy. But in this study a total of 1200 dataset that belong to four dialect categories such as Gojjam, Gondar, Wollo and Shewa that is 300 data samples per class. As compared to the dataset of state-of-the-art models developed, the data we used is smaller and leads to lower performance.

Speech data consisting of 1200 sounds is collected from Amhara Media Corporation. Then a CNN dialect identification model is developed by using a total of 1200 audio samples collected from 300 individuals. Regularization techniques such as dropout, batch normalization are applied during model development to mitigate over fitting problem and yield better accuracy. The CNN model achieved training accuracy of 98% and test accuracy of 85%. Then this dialect recognition model is applied for classifying Amharic speech into their corresponding dialect category.
5.2. Contribution

The main contributions of this study are described as follows:

- We have seen the performance of CNN Amharic dialect identification model for classifying Amharic speech into dialect classes with better accuracy than CNN-SVM and state of the art models.

- We performed preprocessing operation, like thresholding method to remove signals with energy quieter than -26dBFS or below -26dBFS to filter out the relevant feature.

- Since our Spontaneous speech data for dialect classification is obtained in uncontrolled environments, so, we applied moving average filter to isolate noise. To enhance the performance of the system.

- We made experiments on state-of-the-art CNN architectures on this limited data to see the effectiveness of the model on Amharic dialects.

5.3. Recommendation

In this study, we have accomplished our task of developing speech-based Amharic dialect identification model using comparative analysis of different techniques and the research meet its objective. Based on the finding and results of the study, the following are recommended by the researcher for future work:

- The current model is trained on limited data. Future researches can focus on improving recognition rate with more diverse data that considers dialectal difference between speakers to make dialect recognizer robust enough.

- It is also recommended to develop and deploy a system that can run on android devices to assist users in their day-to-day activities.

- Consider a robust system which handles background noises from data collected in uncontrolled environment by itself.
6. References


https://doi.org/10.1109/TASLP.2014.2339736


http://213.55.95.56/bitstream/handle/123456789/15597/Mengistu Tadesse.pdf?sequence=1&isAllowed=y


Demircan, Örnek (2020). Comparison of the effects of mel coefficients and spectrogram images via deep learning


Appendices

Appendix A

<table>
<thead>
<tr>
<th>Name</th>
<th>Their contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMECO archive and system admins</td>
<td>Our data source/helped me to get the data</td>
</tr>
<tr>
<td>AMECO software development teams</td>
<td>Checked the model(tasted the model)</td>
</tr>
<tr>
<td>AMECO camera mans</td>
<td>captured the archived data</td>
</tr>
<tr>
<td>Advisor/Abraham Debasu</td>
<td>Advised me and gave me relevant comments</td>
</tr>
</tbody>
</table>

Table 1. Contacted agents during the research process.

Appendix B

<table>
<thead>
<tr>
<th></th>
<th>Fonts</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>ከ,</td>
<td>እ</td>
<td>እ,</td>
</tr>
<tr>
<td>Middle</td>
<td>እ,</td>
<td>እ</td>
<td>እ</td>
</tr>
<tr>
<td>Low</td>
<td>እ</td>
<td>እ</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Categories of Amharic vowel
### Table 3. Categories of Amharic consonants

<table>
<thead>
<tr>
<th>Manner of Articulation</th>
<th>Voicing</th>
<th>Place of Articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Voiceless</td>
<td>Labials</td>
</tr>
<tr>
<td>Stops</td>
<td>Voiceless</td>
<td>ꠇ</td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>ꠇ</td>
</tr>
<tr>
<td></td>
<td>Glottal zed</td>
<td>ꠞ</td>
</tr>
<tr>
<td>Fricatives</td>
<td>Voiceless</td>
<td>ꠐ</td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>ꠐ</td>
</tr>
<tr>
<td></td>
<td>Glottal zed</td>
<td>ꠐ</td>
</tr>
<tr>
<td>Nasal</td>
<td>Voiced</td>
<td>ꠔ</td>
</tr>
<tr>
<td>Liquids</td>
<td>Voiced</td>
<td>ꠌ</td>
</tr>
<tr>
<td>Glides</td>
<td>Voiced</td>
<td>ꠒ</td>
</tr>
</tbody>
</table>

**Appendix C**

![Voice Recognition Diagram](image)

*Figure 1. Speech recognition process*
Appendix D

Figure 2. Different types of machine learning.

Appendix E

Figure 3. K-Fold cross-validation