

# Speech Processing for Text Independent Amharic Language Dialect Recognition

**Abrham Debasu Mengistu\***, Dagnachew Melesew Alemayehu

Bahir Dar University, Bahir Dar Institute of Technology, Faculty of Computing

\*Corresponding author, e-mail: abrahamd@bdu.edu.et

## Abstract

*Dialect is a difference of verbal communication spoken by people from a particular society or geographic area so the paper focuses on Amharic language dialect recognition. In this paper, the authors have used backpropagation artificial neural network, VQ (vector quantization), (Gaussian Mixture Models) and a combination of GMM and backpropagation artificial neural network for classifying dialects of Amharic language speakers. In this research, a total of 100 speakers for each group of dialects are considered each having about 10 seconds duration is collected. The feature vectors of Mel frequency cepstral coefficients (MFCC) had been used to recognize the dialects of speakers. In this research paper the recognition model that uses a tanh activation function have a better result instead of using the Logistic Sigmoid activation function in backpropagation artificial neural network. After conducting the above experiments 95.7% accuracy achieved when GMM and backpropagation artificial neural network with tanh activation function are combined.*

**Keywords:** GMM, VQ, MFCC, ANN, Ethiopian dialects

**Copyright © 2017 Institute of Advanced Engineering and Science. All rights reserved.**

## 1. Introduction

Speech is the majority universal and natural means of communication among humans. A language when used by people from different regions can be analyzed to see the usage of words with different expressions and even if they speak some standard form of the word with the difference in spectral properties of sound produced can be observed [1]. A dialect is a local or social variety of a language distinguished by the method they speech, Ethiopia has 83 different languages with up to 200 different dialects spoken [2]. The largest ethnic and linguistic groups are the Oromos, Amharas and Tigrayans. Like other languages in the world, Amharic language also has many varieties these Amharic dialects are spoken over the entire Amharic speaking regions. Amharic Language has different dialects and is most commonly spoken language in Amharic speaking regions. The four major Amharic Dialects are Gonder, Gojjam, Wolo and shewa [3].

## 2. Statement of the Problem

Amharic is the working language of the country Ethiopia and spoken by about 65 million people. Among this, about 63400 speakers are living abroad in other countries [5]. It ranks 55 in the number of first number of speakers in the world [6]. In spite of the relatively large number of speakers, nothing has been done in the area of dialects recognition system.

## 3. Data Collection and Implementation of Tools

To collect the data set, we used directly record from speakers. To have a speech samples of different varieties, speakers are randomly chosen. We have tried to include both male and female speakers so as to have a good data set containing both sexes. After collecting the data, the next step is preparing it to have the same sampling frequency and bit with the one we recorded from speakers and having a data set of such types is very helpful to us to determine the potential use of dialects identification on the different speech samples. A total of 100 speakers for each group of dialects are considered for this study. 100 speech samples each having about 10 seconds duration is collected from each individual. Each sample is taken

at a sampling rate of 16KHz and 16 bit. After being collected, all these data is properly preprocessed and the necessary features are extracted.

#### 4. Overview of Amharic Language Dialects Recognition

The task of dialect recognition occurs in wide range of activity. The problem of recognition is concerned with the construction of a procedure that will be applied to recognize dialects of speakers, in which each new item must be assigned to one of a set of predefined classes on the basis the extracted features. For dialects recognition, a group of  $S$  speakers  $S = \{1, 2 \dots S\}$  is represented by GMM's  $\lambda_1, \lambda_2 \dots \lambda_S$ . The objective is to find the dialects model which has the maximum a posteriori probability for a given observation.

$$\hat{S} = \arg \max_{1 \leq K \leq S} \Pr \quad (1)$$

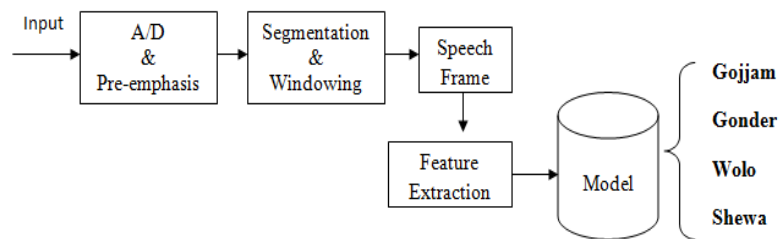


Figure 1. Dialect Recognition Model

As shown in the above Figure 1, classification of Amharic language dialects recognition involves the following activities:

#### 5. Signal Preprocessing

In dialects recognition, the first phase is preprocessing which deals with a speech signal which converts an analog signal at the recording time to digital. The properties of a signal changes with time, so that the speech can be divided into a sequence of uncorrelated segments or frames and process the sequence as if each frame has fixed properties. First, the continuous dialect speech signal  $D(t)$  produced by the speaker and sensed by the microphone has to be converted to the discrete domain. Secondly, the speech signal is segmented into frames. This is done to obtain quasi stationary units of speech. Finally, a pre-emphasis filter is applied to each frame generated in the previous step. Once all this procedure has been performed, the speech frames are ready to enter the feature extraction subsystem. Diagrammatically, it can be represented as follows.

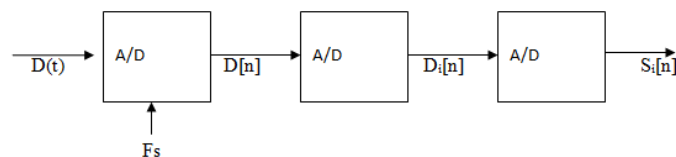


Figure 2. Signal processing subsystem

Where  $D[n]$  is signal which is converted to digital format where as  $D_i[n]$  is the signal after pre-emphasis is applied on it.  $S_i[n]$ , which is the final output of pre-processing phase is the signal which is segmented in to frames and overlapped. Next, each of the steps are to be discussed.

### 5.1. Input Speech

The speech signal is a continuous air-pressure signal that can be captured by a microphone. The microphone converts this pressure-signal into a continuous electrical signal. The A/D converter target is to convert this continuous representation into the discrete domain so that it can be processed in the digital domain [7].

### 5.2. Pre-Emphasis

Due to the structure of voice production system, damping occurs in high-frequency regions. For that reason, the spectrums of voiced regions are compensated by pre-emphasis which amplifies high-frequency regions and performs filtering [8]. Widely used pre-emphasis ranges from 0.95 to 0.97 and filter is given as,

$$Y[n]=x[n]-a*[n-1], a \approx (0.95 - 0.97) \quad (1)$$

In this study we took  $a = 0.95$

### 5.3. Silence Removal

Silence/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 [9], as well as silence removal are 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 [10]. The picture below shows the speech signal before and after of silence removal and endpoint detection.

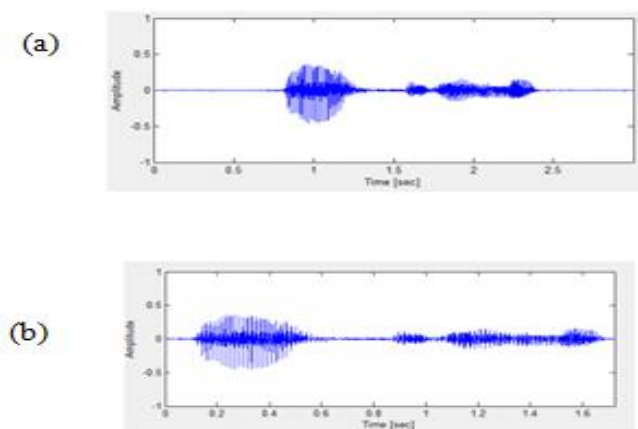


Figure 3. Signal before (a) and after silence removal (b)

### 5.4. Segmentation

In this step the continuous speech signal is blocked into frames of  $N$  samples, with adjacent frames being separated by  $M$  ( $M < N$ ). The first frame consists of the first  $N$  samples. The second frame begins  $M$  samples after the first frame, and overlaps it by  $N - M$  samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for  $N$  and  $M$  are  $N = 256$  (which is equivalent to  $\sim 30$  ms windowing and facilitate the fast radix-2 FFT) and  $M = 100$  [11]. The voice signal cannot be considered as a long-term stable signal as its properties vary considerably along time. However, if that signal is analyzed in a very short period of time (order of milliseconds), the properties of voice do not change so drastically and it can be considered as a quasi-stationary signal. This lack of stability is produced by the movement of the articulators which vary their position to produce different

phonemes. The transition between two phonemes involves the transition of the articulator organs from one position to another. This transition is not immediate and this is reflected in the waveform signal. Generally these transitions are problematic in speech analysis especially when a speech frame is centered in that transition. To avoid this effect, frame overlapping can be applied to the speech signal. The period of time the articulators remain stable is about 80-200 ms. Segmentation is necessary to divide the speech signal into short-enough frames with quasi-stationary properties. Each of these frames will be individually analyzed and used to generate a feature vector [12].

### 5.5. Windowing

The pre-emphasized signal is divided into short frame blocks, and a window is applied to these frames. The frame length can vary, but based on empirical results, is often chosen from 20 to 30ms [13] with an overlap of 10ms. This length depends on the specific feature extraction method that is applied. The window function that is applied is preferably not rectangular, as this can lead to distortion due to vertical frame boundaries [8].

The output signal of windowing block  $xw[n]$  can be calculated as:

$$Xw[n]=x[n].w[n] \quad (2)$$

In all speech signal recognition systems, signal is firstly converted to some measurement values representing the speech that are called as features. To represent speech signal various features are used. Most important features are energy, pitch frequency, formant frequency [14], linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), Mel-Frequency cepstral coefficients (MFCC) and their derivatives. Speech signal converted to features vectors are modeled by using various classification methods. Neural Networks (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Support Vector Machine (SVM) are the most commonly used classification methods in speech recognition [15, 16]. Speech signal does not contain speech information only. At the same time, it contains information like age, gender, and emotional state that are related to the speaker [17].

### 5.6. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is one of the most frequently used features both in speech and speaker recognition [14-16], [20-21]. Stevens and Volkman (1940) experimentally showed that human hearing system perceives the frequencies linearly up to 1 KHz and logarithmically above it. Relationship between perceived frequency which is called Mel and actual frequency is given in as,

$$\text{Mel}(f)=2595*\log(1+f/700) \quad (3)$$

## 6. Gaussian Mixture Model

GMM can smoothly approximate the probability density function of arbitrary shape, portray distributed characteristic of different speaker's speech feature in the feature space. Speech production is not deterministic. A particular sound is not produced by a speaker with exactly the same vocal tract shape, glottal flow, due to context, co articulation, anatomical and fluid dynamical variations. One way to represent this variability is probabilistically through multi-dimensional Gaussian probability density function [19].

## 7. Vector Quantization

Vector quantization (VQ) is the process of taking a large set of feature vectors and producing a smaller set of feature vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point. We use vector quantization since it would be impractical to store every single feature vector that we generate from the training utterance. While the VQ algorithm does take a while to compute, it saves time during the testing phase, and therefore is a compromise that we can live with [18, 20, 21].

A vector quantizer maps  $k$ -dimensional vectors in the vector space  $R^k$  into a finite set of vectors  $Y=\{y_i:i=1,2,\dots,N\}$  Each vector  $y$  is called a code vector or a codeword and the set of all

the codewords is called a codebook. Associated with each codeword,  $y_i$ , is a nearest neighbor region called Voronoi region, and it is defined by:

$$V_i = \{X \in R^k : \|x - y_i\| \leq \|x - y_j\|, \text{ for all } j \neq i\} \tag{4}$$

**8. Back Propagation Artificial Neural Network**

The neural network needs 13 MFCC inputs and 4 neurons in its output layer to classify the dialects of speakers. The hidden layer has 10 neurons. This number was picked by trial and error methods, if the network has trouble of learning capabilities, then neurons can be added to this layer. Each value from the input layer is duplicated and sent to all of the hidden nodes.

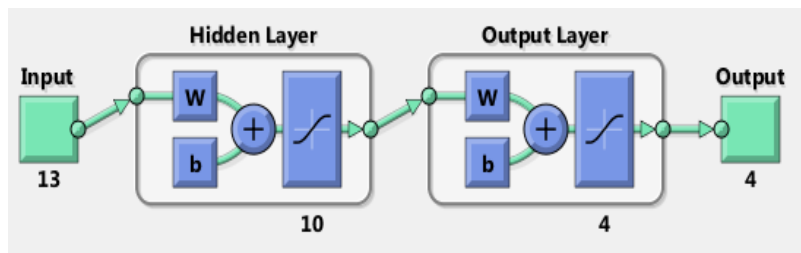


Figure 4. Artificial Neural Network

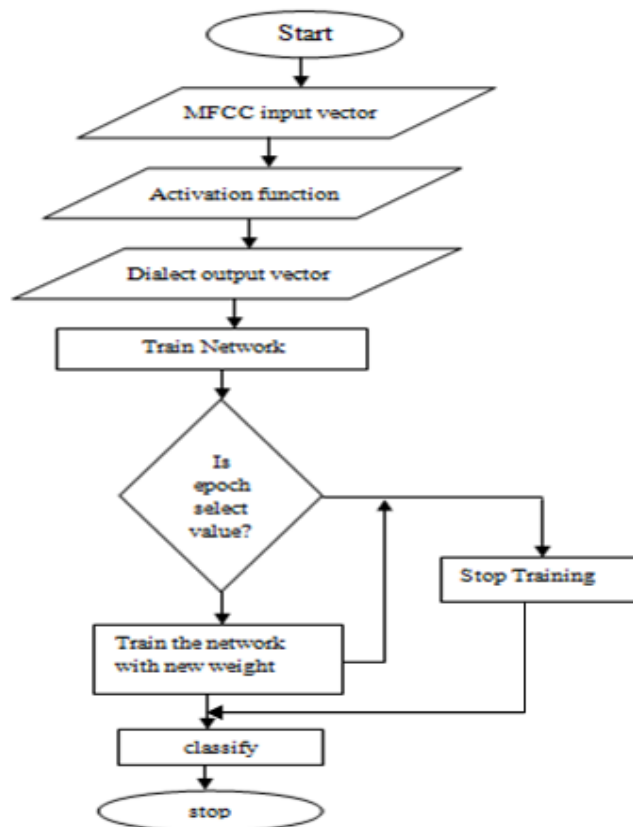


Figure 5. Artificial neural network algorithm

## 9. Experimentation and Discussion

In this research., four different methods, namely the Vector Quantization (VQ), Gaussian mixture models (GMMs), backpropagation artificial neural network and the combination of GMM and backpropagation neural network are used. Mel frequency cepstral coefficients (MFCC) are used to recognize the dialects of speakers. To see the effect of the number of these feature vectors on the performance of the system, a number of 13, 26 and 39 vectors is used. Here, 13 is simply the MFCC extracted from each frame of a given sample and 26 is a vector space obtained by adding 13 delta coefficients on MFCC coefficients where as 39 is obtained by adding 13 acceleration coefficients on delta coefficients. These three coefficients are also known as MFCC,  $\Delta$ MFCC and  $\Delta\Delta$ MFCC. Below, the results obtained from the experiments are explained.

Table 1. Recognition result

# of speakers	#MFCC coefficients	VQ (%)	GMM (%)	ANN (sigmoid) (%)	ANN (tanh) (%)	GMM & ANN %
25	13	73.2	81.4	83.7	84.5	89.2
	26	75.7	82.0	86.1	89.6	89.7
	39	76.0	83.8	88.9	90.2	89.9
50	13	70.7	81.8	84.0	85.1	91.4
	26	72.8	83.2	86.9	89.9	92.0
	39	74.3	83.9	88.9	91.0	93.5
75	13	67.9	82.9	87.2	89.4	92.7
	26	69.0	86.3	87.7	91.1	94.0
	39	68.7	84.0	89.2	92.6	94.1
100	13	63.9	84.0	88.0	92.0	94.9
	26	65.2	86.5	88.0	92.4	95.2
	39	65.4	86.9	90.1	93.2	95.7

Here, we used the first 13 MFCC coefficients for both training and testing. As shown from the above table, the experiment was conducted for varying number of dialect speakers, the minimum being 25 and the maximum 100. Because VQ is a template matching approach, as the number of speakers tend to increase, its performance decreases. When we tested for 25 speakers using the four mentioned classifiers we got 73.2%, 81.4%, 83.7%, 84.5 and 89.2% respectively for dialects of Amharic language speakers.

After experimenting with 13 MFCC coefficients, authors conducted another experiment to see the performance of the system by increasing the number of coefficients to 26 and got some improvements from the first experiment. Here, the percentage of correctly classified dialect speakers tend to increase when we compare it with the first one. And also like in the one above, for VQ as the number of dialect speakers increase, its performance decreases because in VQ probability of having similar templates increases. For the other classifier, as the number of speakers increases, the classifier's accuracy also increases. In addition to this, as the number of speakers increases, this increment in similarity makes the system to pass a correct decision on the recognition of dialects speakers. After trying the above, 76.0%, 83.6%, 88.9%, 90.2% and 89.9% success for 25 individuals using VQ, GMM, ANN with sigmoid activation function and ANN with tanh activation function in Amharic language dialects and 65.4%, 86.9%, 90.1%, 93.2% and 95.7% when the number of speakers is 100 using the above mentioned algorithms respectively from our experiment. The convergence rate of back propagation is very low. This is because the researchers is required to arbitrarily select parameter such as the network topology, initial weights and biases, a learning rate value and the activation function. We have used both sigmoid and tanh activation function in back propagation learning technique. The tanh function is another activation function used by artificial neural network and mathematically defined by:  $\tanh(X) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ . The tanh activation function has a larger range than the sigmoid functions this range is comparable to twice over that of the Sigmoid this means that sigmoid functions ranges from [0, 1] and tanh function ranges [-1, +1] therefore, the tanh function is more sensitive to a change in learning rate, which also ensures that weight changes have a greater range.

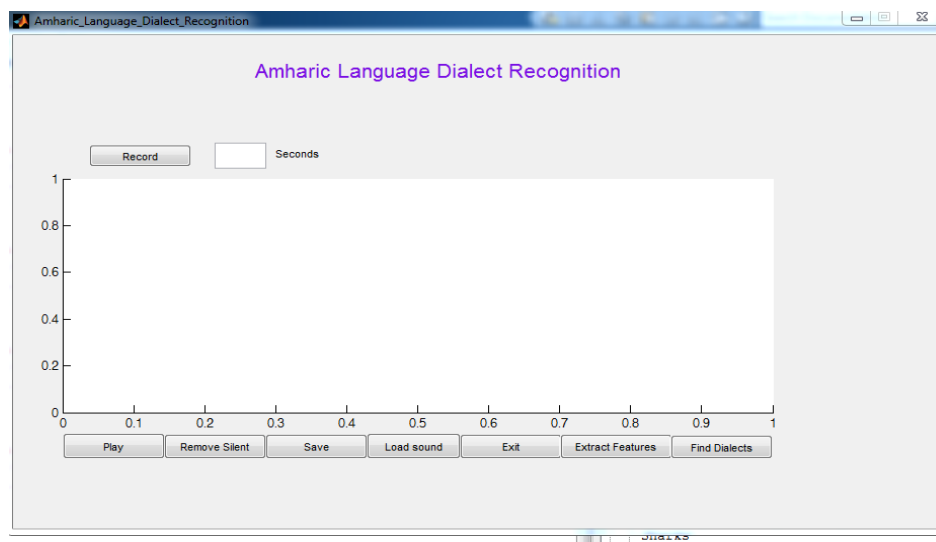


Figure 6. Dialect Recognition Prototype

## 10. Conclusion and Future Work

In this paper, we have shown that the four Amharic language dialects can be recognized using a hybrid approach of GMM and ANN with hopeful result. Table 1 shows the accuracy results of Amharic language dialects. Because this is the first work in dialects identification for the Amharic language, there are many things to be performed to increase the perfection of the system. These includes developing a noise robust system, the one capable of identifying the speaker's dialect with a speech utterance of high noise. The other task that we need to recommend is preparation of large dialects database so that a researcher can easily access it to save time which will be elapsed in collecting speech of dialects.

## References

- [1] Shweta Sinha, Aruna Jain and SS Agrawal. Acoustic phonetic feature based dialect identification in hindi speech. *International journal on smart sensing and intelligent systems*. 2015; 8(1).
- [2] <http://www.ethiopian treasures.co.uk/pages/language.htm>
- [3] <http://www.languagecomparison.com/en/amharic-dialects/model-58-6>
- [4] B Gamback, L Asker. Experiences with Developing Language Processing Tools and Corpora for Amharic. <http://joshuaproject.net/languages/amh>.
- [6] <http://www.davidpbrown.co.uk/help/top-100-languages-by-population.html>.
- [7] Y Kelbesa. An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System. Universit àdegli Studi di Brescia. 2014.
- [8] S Patra, Robust Speaker Identification System. Super Computer Education and Research Centre, Indian Institute of Science Bangalore. 2007; 560 012.
- [9] G Saha, S Chakroborty, S Senapati. *A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications*. Department of Electronics and Electrical Communication Engineering Indian Institute of Technology, Kharagpur-721 302, India. 2014.
- [10] R Islam and F Rahman. Improvement of Text Dependent Speaker Identification System Using Neuro-Genetic Hybrid Algorithm in Office Environmental Conditions. *JCSI International Journal of Computer Science Issues*. 2009; 1.
- [11] SM Siniscalchi, F Gennaro, S Andolina. *Embedded Knowledge-based Speech Detectors for Real-Time Recognition Tasks*. Dipartimento di Ingegneria Informatica, Università di Palermo V.Iedelle Scienze (Edif. 6), 90128 Palermo, Italy.
- [12] Y Kelbesa. *An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System*. Universit àdegli Studi di Brescia. 2014.
- [13] LP Heck. *Automatic Speaker Recognition Recent Progress, Current Applications, and Future Trends*. MIT Lincoln Laboratory. 2000.
- [14] Ergün Yücesoy & Vasif V Nabiyev. Gender Identification of a Speaker Using MFCC and GMM.

- 
- [15] MH Sedaaghi. A Comparative Study of Gender and Age Classification in Speech Signals. *Iranian Journal of Electrical & Electronic Engineering*. 2009; 5(1): 1- 12.
- [16] Djemili, Rafik, Hocine Bourouba, and Mohamed Cherif Amara Korba. A speech signal based gender identification system using four classifiers. 2012 International Conference on Multimedia Computing and Systems (ICMCS). *IEEE*. 2012.
- [17] L Rabiner and BH Juang. *Fundamentals of Speech Recognition, Englewood Cliffs (N.J.)*. Prentice Hall Signal Processing Series. 1993.
- [18] Rajsekha. *Real time speaker recognition using MFCC and VQ*. Department of Electronics & Communication Engineering National Institute of Technology Rourkela – 769008. 2008.
- [19] S Selvanidhyanathan, S Kumara. Language and Text-Independent Speaker Identification System Using GMM. *WSEAS Transactions on Signal Processing*. 2013; 4(9).
- [20] Inggih Permana, Agus Buono, Bib Paruhum Silalahi. Similarity Measurement for Speaker Identification Using Frequency of Vector Pairs. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2014; 12(8): 6205- 6210.
- [21] TB Adam, MS Salam, TS Gunawan. Wavelet Cesprtral Coefficients for Isolated Speech Recognition. *TELKOMNIKA*. 2013; 11(5): 2731- 2738.