International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.10, No. 1 (2017), pp.215-222
http://dx.doi.org/10.14257/ijsip.2017.10.1.21

Text Independent Amharic Language Dialect Recognition: A Hybrid Approach of VQ and GMM

Abrham Debasu Mengistu[1*] and Dagnachew Melesew[2]

[1]*Bahir Dar University, Bahir Dar Institute of Technology, Bahir Dar, Ethiopia, Email: abiyity@gmail.com; abrhamd@bdu.edu.et*
[2]*Bahir Dar University, Bahir Dar Institute of Technology, Bahir Dar, Ethiopia, Email: dagnachew.m@gmail.com; dagnachewm@bdu.edu.et*

Abstract

In Amharic language there are four main different types of dialects these are Gojjam (Gojjamegna), Wollo (Wollogna), Shewa (Shewagna) and Gonder (Gonderegna). In this paper a hybrid approach of VQ(vector quantization) and GMM(Gaussian Mixture Models) have been used for classifying dialects of Amharic language. For our data set a total of 100 speakers for each group of dialects are considered. Mel frequency cepstral coefficients (MFCC) feature vectors 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, MFCC, ΔMFCC and ΔΔMFCC vectors are used. When 25 speakers are considered from areas, 85.9% accuracy achieved. After conducting this experiment, the number of speakers are increased to100, which is the maximum number of dialect speakers for our experiment, 92.7% accuracy achieved for the given dialects.

Keywords*: GMM, VQ, MFCC, Amharic dialects*

1. Introduction

Speech is the most common 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 the difference in spectral properties of sound produced can be observed [1]. A dialect is a regional or social variety of a language distinguished by the way they speech pattern of a region. 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. It is important to know Amharic dialects because different Amharic Dialects are spoken by Amharic speakers. 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 countries. The total number of Amharic Dialects is four these are Gonder, Gojjam, Wolo and shewa [3].

2. Statement of the Problem

Amharic is the working language of the country Ethiopia and it ranks 55 in the number of first number of speakers in the world [4-6]. Despite the fact that there are relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed, and nothing has been done in terms of making the language useful in the area of dialects recognition system.

*Corresponding Author

ISSN: 2005-4254 IJSIP

3. Data Set Collection and Preparation

To collect the data set, directly record from speakers had been used. To have a speech samples of different varieties, speakers are randomly chosen. In addition to this, the data set also contains utterances from both sex. After collecting the data, the next step is preparing it to have the same sampling frequency. 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 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. 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.

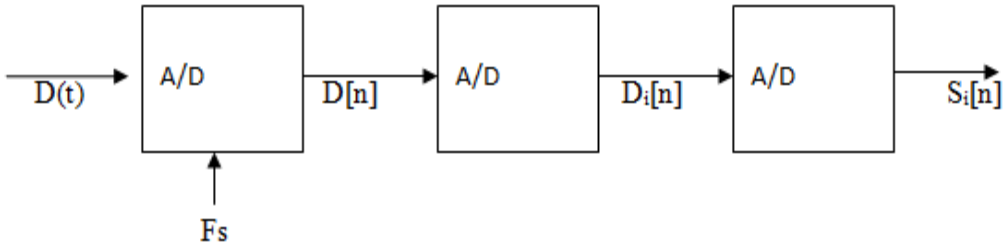

Figure 1. Signal Processing Subsystem

Where D[n] is signal which is converted to digital format where as Di[n] is the signal after pre-emphasis is applied on it. Si[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.

4.1. 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].

4.2. 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

4.3. Silence Removal and End Point Detection

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].

4.4. Segmentation and Overlapping

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 ~ 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].

4.5. 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 from20 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, format 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].

4.6. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is one of the most frequently used features both in speech and speaker recognition [14-16]. 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,

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

4.7. 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].

4.8. 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].

A vector quantizer maps k-dimensional vectors in the vector space Rk into a finite set of vectors Y={yi:i=1,2…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, yi, is a nearest neighbor region called Voronoi region, and it is defined by:

$$V_i=\{X\epsilon R^k:||x-y_i||\leq||x-y_j||, \text{ for all } j\neq i \quad (5)$$

5. Dialects Recognition

For dialects recognition, a group of S speakers S= {1, 2 ...S} is represented by GMM‘s λ1, λ2…λS. The objective is to find the dialects model which has the maximum a posteriori probability for a given observation.

$$\hat{S}=\arg\max_{1\leq K\leq S} \Pr \quad (6)$$

6. Experimentation and Discussion

In this research., two different methods, namely the Vector Quantization (VQ) and Gaussian mixture models (GMMs) 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, ΔMFCC and ΔΔMFCC. Below, the results obtained from the experiments are explained.

Table 1. Recognition Result of Amharic Dialects

# of speakers	#MFCC coefficients	VQ (%)	GMM (%)	GMM & VQ
25	13	65.2%	61.4%	80.2%
	26	67.9%	64.0%	83.7%
	39	69.9%	67.6%	85.9%
50	13	60.1%	71.8%	86.1%
	26	61.4%	73.2%	86.4%
	39	62.7%	73.9%	86.9%
75	13	57.6%	72.9%	88.7%
	26	57.9%	76.3%	88.8%
	39	58.2%	76.8%	89.1%
100	13	48.4%	77.2%	89.7%
	26	52.5%	79.1%	89.9%
	39	53.1%	79.9%	92.7%

Here, we used the first 13 MFCC coefficients for both training and testing. As we can see from the above table, the experiment was conducted for varying number of dialect speakers, the minimum being 25 and the maximum 100. In case of VQ the number of dialect speakers increase, its performance decreases. This is because, as the number of speakers increases, the probability of having similar templates increases. When 13 MFCC coefficients considered with 25 speakers 65.2% accuracy achieved. After experimenting with 13 MFCC coefficients, we 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. After trying the above mentioned experiments, we tried to see what will happen to the result if 39 MFCC coefficients are used. We got 69.9% success for 25 individuals in the given dialects. When the number of speakers increased to 100, which is the maximum number of dialect speakers for our experiment, we got 53.1% success using 39 MFCC coefficients For GMM, 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. For 25 individuals considering the first 13 MFCC coefficients using GMM 61.4% accuracy achieved and when 100 speakers with 39 MFCC coefficients are considered in this experiment 79.9% accuracy achieved.

The last experiment was conducted to see what will happen in the hybrid approaches of both VQ and GMM. In the hybrid approaches as the numbers of speakers increases the identification accuracy also increases. In this experiment, when 25 speakers with 13 MFCC coefficients are considered 80.2% success are achieved. Similarly, when the individuals increased to 100 with 39 MFCC coefficients, 92.7% accuracy achieved.

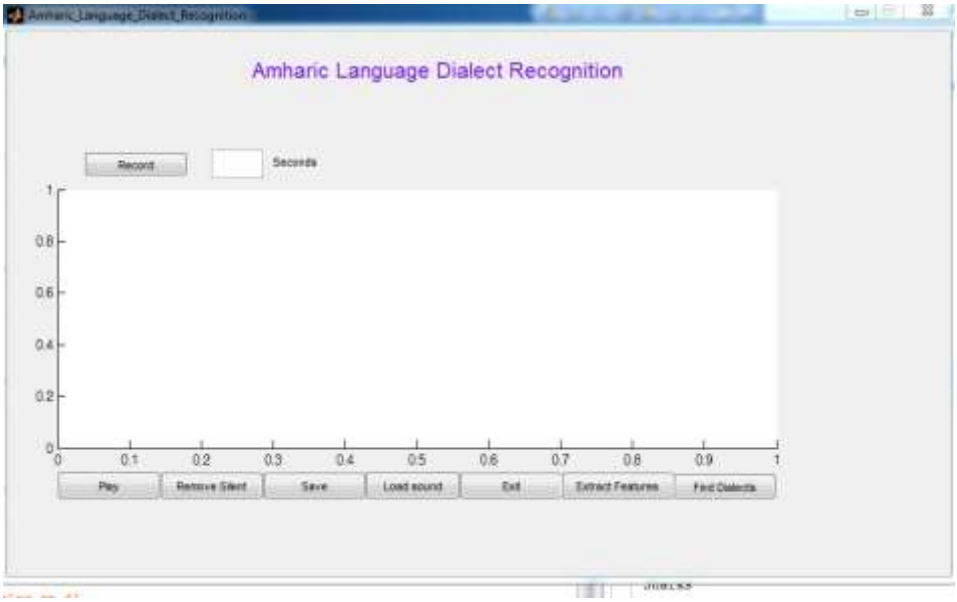

Figure 2. Dialect Recognition Prototype

7. 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 VQ with promising 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 and testing using others techniques.

References

[1] S. Sinha, A. Jain and S. S. Agrawal, "Acoustic phonetic feature based dialect identification in hindi speech", International journal on smart sensing and intelligent systems, vol. 8, no. 1, **(2015)** March.
[2] http://www.ethiopiantreasures.co.uk/pages/language.htm.
[3] http://www.languagecomparison.com/en/amharic-dialects/model-58-6.
[4] B. Gamback and L. Asker, "Experiences with Developing Language Processing Tools and Corpora for Amharic".
[5] http://joshuaproject.net/languages/amh.
[6] http://www.davidpbrown.co.uk/help/top-100-languages-by-population.html.
[7] I. 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 560 012, **(2007)**.
[9] G. Saha, S. Chakroborty and S. Senapati, "A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications," Departmentof Electronicsand ElectricalCommunication Engineering Indian Institute of Technology, Khragpur, 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, vol. 1, **(2009)**.
[11] S. M. Siniscalchi, F. Gennaro and S. Andolina, "Embedded Knowledge-based Speech Detectors for Real-Time Recognition Tasks," Dipartimento di Ingegneria Informatica, Università di Palermo V.ledelleScienze (Edif. 6), 90128 Palermo, Italy.
[12] I. Y. Kelbesa, "An Intelligent Text Independent Speaker Identification using VQ-GMM model based Multiple Classifier System," Universit àdegli Studi di Brescia, **(2014)**.
[13] L. P. Heck, "Automatic Speaker Recognition Recent Progress, Current Applications, and Future Trends," MIT Lincoln Laboratory, **(2000)**.
[14] E.Yücesoy and V. V. Nabiyev, "Gender Identification of a Speaker Using MFCC and GMM".

[15] M. H. Sedaaghi, "A Comparative Study of Gender and Age Classification in Speech Signals", Iranian Journal of Electrical & Electronic Engineering, vol. 5, no. 1, **(2009)** March, pp. 1- 12.
[16] R. Djemili, H. Bourouba and M. C. A. Korba. "A speech signal based gender identification system using four classifiers." Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE, **(2012)**.
[17] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs (N.J.), Prentice Hall Signal Processing Series, **(1993)**.
[18] A. Rajsekha, "Real time speaker recognition using MFCC and VQ," Department of Electronics & Communication Engineering National Institute of Technology Rourkela – 769008, **(2008)**.
[19] S. Selvanidhyananthan, S. kumara, "Language and Text-Independent Speaker Identification System Using GMM," WSEAS Transactions on Signal Processing, vol. 9, no. 4, **(2013)** October.

Authors

Abrham Debasu Mengistu, he is born in February 04, 1985 and received his B.Sc. Degree in Computer Science from Bahir Dar University and also MSc. in Computer Science from Bahir Dar University, School of Computing and Electrical Engineering, BiT, Ethiopia. He has published 06 research papers in international journal. His main research interest is signal processing, image processing and Robotics. He is a life member of professional societies like MSDIWC.

Dagnachew Melesew Alemayehu, he is born in January 15, 1985 and received his B.Sc. Degree in Computer Science from Bahir Dar University and MSc. in Information Technology from Madras University, India. He has published 06 research papers in international journal. His main research interest is Image processing and Robotics.