Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC, MFCC and GFCC

Abrham Debasu Mengistu
Faculty of Computing, Department of computer science
Bahir Dar University, Bahir Dar Institute of Technology
Bahir Dar, Post code 26, Ethiopia
abrhamd@bdu.edu.et

Abstract: Speaker recognition refers to the automated method of identifying the identity of an individual based on verbal communication. This paper presents an automatic text-independent speaker identification system for the Amharic language in noisy environments. VQ (vector quantization), GMM (Gaussian Mixture Models) and a combination of GMM and (BPNN) Backpropagation Artificial neural network had been as techniques for speaker recognition. For feature vectors MFCC (Mel-frequency cepstrum coefficients), LPCC (Linear Predictive Cepstral Coefficients) and GFCC (Gammatone Frequency Cepstral Coefficients had been used. For the identification process, speech signals are collected from different speakers including both sexes; for our data set, a total of 300 speakers' speech samples were collected, and each speech has 10 seconds duration from each individual. 93.7% accuracy achieved when GMM and Backpropagation Artificial neural network with tanh activation function are combined.

Keywords: ANN, VQ, GMM, MFCC, GFCC and LPCC.

1. Introduction

Ethiopia has 83 different languages that have up to 200 different dialects spoken. The largest ethnic and linguistic groups are the Afan Oromos, Amharas and Tigrayans [1]. In line with this, Ge'ez is an ancient language of the country; it was introduced as an official written language during the first Aksumite kingdom when the Sabeans' sought refuge in Aksum. The Aksumites developed Ge'ez, a unique script derived from the Sabean's alphabet, and it is still used by the Ethiopian Orthodox Tewahedo Church today. Regarding to this, Tigrigna and Amharigna (Amharic) are the languages which are derived from Ge'ez. Amharic is the working language of our country. Generally, Ethiopian languages are divided into four major language groups. These are Semitic, Cushitic, Omotic, and Nilo-Saharan. Amharic language which categorized in the Semitic group is the focus of this research paper [2]. 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. Among semantic languages, Amharic is spoken by 30 million people as a first or second language. That is, it is used as the second and most spoken Semitic language in the world (after Arabic). Probably, it is the second largest language in Ethiopia (after Oromo, a Cushitic language), and possibly one of the five largest languages on the African continent. [3].

Speaker recognition or voice recognition refers to the automated method of identifying or confirming the identity of an individual based on his or her voice. For real world use, speaker identification with noise resistance systems is crucial. Within a given a speech sample, speaker recognition is concerned with extracting clues to the identity a person who is the source of that utterance [4].

Besides, Rao etal, speaker recognition is divided into two specific tasks: verification and identification. For speaker verification, its goal is to determine a voice from a given sample to which he or she claims to be. On the other hand, for speaker identification, the goal is to determine or know the voices of one speech perfectly out of the given input voice as a sample.

However, the speech can be constrained to a known phrase (text-dependent) or totally unconstrained (text-independent). Hence, the research mainly focuses on text independent because telephone becomes use as a tool that interacts with computer persistently this day.

2. The Research Method

To collect the data set, audio recorder is used to record the voice directly, and both female and male speakers are included in order to have a good data set form all perspective. The data contains noises because they were record in the café, market and school. Having such types of data set, it was very helpful to determine the potential use of speaker identification in the noise areas.

A total of 100 speakers are considered for this study. Three speech samples were held with each speaker that has 10 seconds duration. That is, form these speakers 300 speeches were record. In addition, each sample is taken at a sampling rate of 16 KHz and 16 bit from these dataset 70% is given to training and the rest 30% is give for testing. Once the data set collected, various sequential steps are performed to achieve the goal of the study through MATLAB, 2013.

3. Related Works

Different researchers have been conducted their researches to find an automated means of identifying the identity of individuals from the speech signal they produce. Regarding to this, related works have performed to identify the speakers' speaking in different languages. These are discussed as follow.

Tech & Bansal, in their work entitled as "Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB" presented text-dependent speaker identification for the Hindi words. They used ek (one), do (two), teen (three), char (four) to train the system and the system have been found to possess a great degree of learning and recognizing accuracy by using MFCC feature extraction technique and Vector Quantization method for pattern matching combination with noise free environment. Then, they found out that 90% success rate in their experiment [5].

Khan, studied in "Hindi Speaking Person Identification Using Zero Crossing Rate" with acoustic measure of voice sources were extracted from 3 utterances spoken by 10 peoples including 5 male and 5 female talkers (aged 19 to 25 years old). They presented a method for isolated Hindi word recognition based on zero-crossing feature. Consequently, the estimation of zero crossing rates reflects more effectively the difference in different people speaking in Hind as their finding [6].

Al-Dahr,etal., conducted a study to develop a system that is capable to identify an individual from a sample of his or her speech for Arabic language, Semitic language. They used a word dependent system using the Arabic isolated word /ns10 as10 cs10 as10 ms10//[unk]/ a single keyword for the test utterance. Speech features are extracted using MFCC, and HTK is used to implement the speaker identification module with phoneme based HMM. The designed automatic Arabic speaker identification system contains 100 speakers and it achieved 96.25% accuracy in recognizing the correct speaker [7].

Das presented the implementation of speaker identification system using artificial neural network with digital signal processing. The system is designed to work with the text dependent speaker identification for Bangla Speech. The utterances of speakers are recorded for specific Bangla words using an audio wave recorder, and acquired by the digital signal processing technique. He used Hamming window and Blackman-Harris window to investigate better speaker identification performance. Then, he found out that the best identification score with 82.5%, which is obtained for one syllable word ―Ami‖ using Hamming window, but the highest false inclusion error is 12.5% [8].

Marciniak, etal, carried out an experiments, used a databases of short Polish utterances. They used fast speaker recognition based on recordings of duration about 1 second, while typically automatic speaker recognition systems need about 7seconds. The result showed that 90% and 84% for the GMM and VQ out of the 25 recorded individuals [9].

Romito & GalatàIn, studied on people who speak Italian language. They recorded and collected the data, reproducing characteristics and instruments usually to be found in legal cases. Then, they evaluate all the Forensic Speaker Recognition (FSR) methods used in Italy using a common data set. Preliminary results demonstrate, however, that much work has yet to be done in order to verify and validate the FSR methods especially when, as happens in Italy, the prosecutions' deductions and conclusions, and subsequently, the verdict, are primarily based on speaker identification [10].

Taabish Gulzar studied on Linear Prediction Cepstral Coefficient (LPCC), Mel Frequency Cepstral Coefficient (MFCC) and Bark frequency Cepstral coefficient (BFCC) feature extraction techniques for recognition of Hindi Isolated, Paired and Hybrid words. On this paper, the authors used Artifical Neural Network is used as back end processor. The experimental results show that the better recognition rate is obtained for MFCC as compared to LPCC and BFCC [11].

4. System Design

The combined feature vector of LPCC, MFCC and GFCC are proposed to identify Amharic speaker identification in the noisy environment. Besides, PCA (Principal Component Analysis) and GA (Genetic algorithm) also used to reduce feature dimension of LPCC, MFCC and GFCC. Hence, BPNN, VQ and GMM for identification of speakers in noise environment and thereby it is possible to find out better results in their combination. In sum, LPCC, MFCC and GFCC have 122 feature vectors each, and then to reduce their dimensions first PCI is applied and it reduced to 59 feature vectors. After that, GA is applied to 59 features. In turn, it comes to 27 feature vectors. These help the study to minimize the training time.

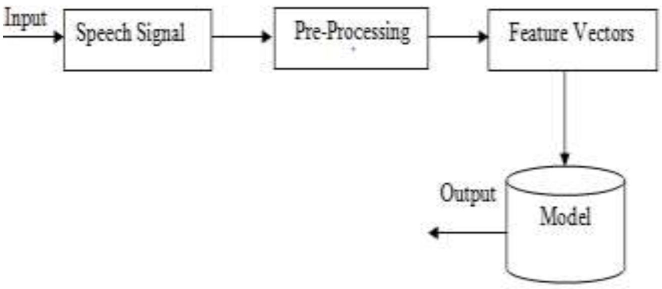

Figure1. Signal processing schematic diagram

5. Signal Pre-processing

5.1. Silence Removal and Filtering

Tanprasert, etal, pointed out that silence removal and filtering is the process of extracting out silence part from the speech signal; otherwise, the training might be seriously biased. We used simple energy based approach to remove the silence part. In this method, the frames having an average energy is below 0.01 times out of the whole utterance are identified and removed [12].

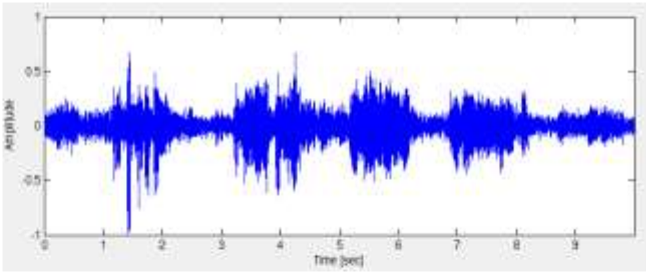

Figure 3: Speech signal

5.2. Feature Extraction

LPCC, MFCC and GFCC feature vector are extracted. The Gammatone Frequency Cepstral Coefficients (GFCC) is speech feature based on a set of Gammatone filter banks. The GFCC is calculated from the Cochleagram of Gammatone Filter bank. Mel-frequency cepstrum coefficients (MFCC) are well-

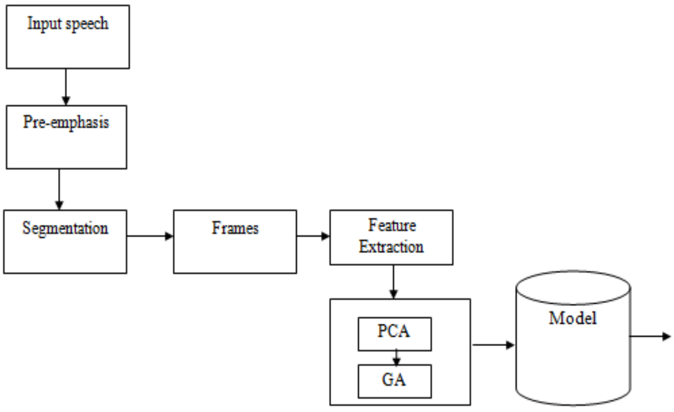

Figure2. Speaker Recognition Model

known features used to describe speech signal. The technique of computing MFCC is based on the short-term analysis, and thus from each frame, an MFCC vector is computed. They are based on the speech processing carried out in the human ear and in the cepstrum of the speech signal [13, 14, 15, and 16].

5.3. Gaussian Mixture Model

Selvani dhyananthan & Kumara, indicated that 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 are not deterministic. A particular sound is not produced by the speaker with exactly the same vocal tract shape, glottal flow, due to the context, co- articulation, anatomical and fluid dynamical variations. One way to represent this variability is probabilistically through multi-dimensional Gaussian probability density function [17, 18].

5.4. Vector Quantization

Rajsekha, revealed that 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). VQ is used because it would be impractical to store every signal feature vector that generate from the training utterance. When VQ algorithm carried out it may take time, but it saves time during testing phase. That is, it is possible to compromise the training and testing time.

A vector quantizer maps k-dimensional vectors in the vector space Rk into a finite set of vectors Y={yi:i=1,2…N}, then each vector y is called a code vector or a codeword, and the set of all the code-words is called a codebook. This is associated with each codeword of yi is a nearest neighbor region called Voronoi region, and it is defined by [19, 20]:

$V_i = \{X \in R^k : \|x-y_i\| \leq \|x-y_j\|$, for all $j \neq i$……… (1)

5.5. BPNN

The neural network needs 27 inputs of the combined feature vectors of LPCC, MFCC and GFCC and 90 neurons in its output layer to classify the speakers. The hidden layer has 17 neurons .This number was picked by trial and error methods, if the network has trouble of learning capabilities, and then neurons can be added to this layer. There is a significant change when we increase the number of hidden layers neurons until 17 but there is no change when the number of hidden layer neurons increases above 17. Each value from the input layer is duplicated and sent to all of the hidden nodes.

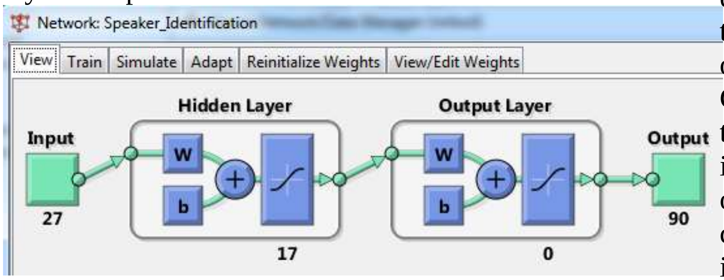

Figure 4. BPNN Model

6. Experimentation and Discussions

In this research, three different methods namely ANN, VQ and GMM with the combined LPCC, MFCC and GFCC feature vectors are used to recognize the speakers.

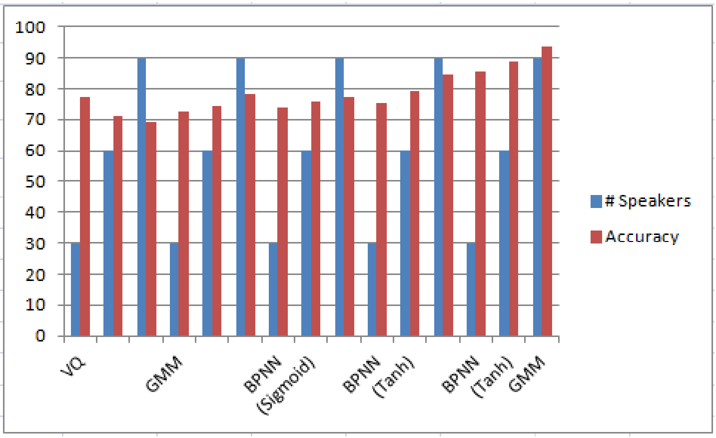

Figure 5: Speaker Identification result

To begin with, the MFCCs are used for both training and testing for VQ, GMM, and BPNN. As the above figure shows, the experiment was conducted for 30, 60 and 90 speakers because this helps us to examine the performance of the methods. VQ is template matching approach, as the number of speakers tends to increase, its performance declines very rapidly. That is when 30, 30, 60 and 90 speakers are tested with the combined feature vectors of LPCC, MFCC and GFCC. The result showed that 77.2%, 70.9% and 69% of them are identified respectively. After first experiment was held the second experiment was conducted in order to see the performance of the system using GMM. This approach develops stochastic models for speakers. As it is a new approach, it is expected to give us better results. As in the VQ approach mentioned in the above, the combined feature vectors are used here and the result revealed that there are some improvements from the first experiment. Here, the percentage of correctly classified speakers tends to increase when we compare it with the first one. Consequently, the third experiment was conducted to see what will happen in the BPNN is used. In BPNN, needs 24 inputs neurons of the combined feature vectors of LPCC, MFCC and GFCC and 90 neurons in its output layer to classify the speakers. The hidden layer has 17 neurons. There is a significant change when we increase the number of hidden layers neurons until 17 but there is no change when the number of hidden layer neurons increases above 17 and the result indicated that there was 84.7% and 77.3% success for 90 individuals' speakers using BPNN with tanh sigmoid activation function. After conducting the above experiments 93.7% success achieved when GMM and BPNN with tanh activation function are combined.

7. Conclusion

The aim of the research paper is to analyze the performance of the algorithms for Amharic language speaker identification under noisy environments. In addition, this research has been focused on text-independent speaker identification since it matches with the original purpose of the research. In this paper, VQ, GMM and BPNN with the combined features of LPCC, MFCCs and GFCCs in speaker identification are tested and the accuracy of the system are presented, and the results of the combined GMM and BPNN with tanh activation function approaches were discussed and encouraging results were obtained.

8. Future Work

The speech processing for the Amharic language can be further investigated. 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. In addition to this, speaker identification for Amharic language in phone calls.

References

[1] Yeshiwas Degu Belay, “Kemant (ness): The Quest for Identity and Autonomy in Ethiopian Federal Polity”, Vol.4, No.18, 2014.

[2] Zelalem Leyew (2000). The Kemantney Language: A Sociolingustic and Grammatical Study of Language. AAU: Unpublished PhD dissertation.

[3] Gamback, B. & Asker, L. (2010). "Experiences with Developing Language Processing Tools and Corpora for Amharic", SICS SWEDISH ICT.

[4] Rao, R., Prasad, A., & Rao, K. (2011). "Robust Features for Automatic Text Independent Speaker Recognition Using Gaussian Mixture Model," International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231 - 2307, Volume-1, Issue-5.

[5] Tech, M.&Bansal, A. (2012). "Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB," International Journal of Computer Applications (0975 – 8887)Volume 45– No.24.

[6] Khan, A., Bhaiya, P., & Banchhor, K. (2012). "Hindi Speaking Person Identification Using Zero Crossing Rate," International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231 -2307, Volume-2, Issue-3.

[7] Al-Dahri, S., Alotaibi, H., & Y. Alsulaiman, A. (2008). " A Word-Dependent Automatic Arabic Speaker Identification," IEEE International Symposium on Signal Processing and Information.

[8] Das, D. (2014). "Utterance Based Speaker Identification Using ANN," International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.4, No.4..

[9] Marciniak, T., Weychan, R., Drgas, Z., Dąbrowski, A., & Krzykowska, A. (2012). "Speaker recognition based on short Polish utterances," Division of Signal Processing and Electronic Systems, Chair of Control and Systems Engineering, Poznań University of Technology, Poznań, Poland.

[10] Romito, L. & Galatà, V. (2008). "Speaker recognition in Italy: Evaluation of methods used in forensic cases," Paper presented at the IV Congreso de Fonética Experimental, Granada – Spain (in press).

[11] Taabish Gulzar, Anand Singh, Sandeep Sharma, “Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks” International Journal of Computer Applications (0975 – 8887) Volume 101– No.12, September 2014. 12

[12] Tanprasert, C., Wutiwiwatchai, C., & Sae-tang, S. (2000). "Text-dependent Speaker Identification Using Neural Network on Distinctive Thai Tone Marks," Technocal Journal, Vol 1, No 6.

[13] Xiaojia, Z., Yuxuan, W., and Wang, D. (2014). “Robust Speaker Identification in Noisy And Reverberant Conditions’ IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 22, No. 4.

[14] S. SELVA NIDHYANANTHAN, R.SHANTHA SELVA KUMARI,” Language and Text-Independent Speaker Identification System Using GMM”, WSEAS TRANSACTIONS on SIGNAL PROCESSING, Issue 4, Volume 9, October 2013.

[15] Ming, J., Hazen, T., Glass, J., and. Reynolds, D. (2007). “Robust Speaker Recognition in Noisy Conditions,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 5.

[16] Utpal Bhattacharjee, “A Comparative Study Of LPCC And MFCC Features For The Recognition Of Assamese Phonemes”, (IJERT), Vol. 2 Issue 1, January- 2013.

[17] Rajsekha, A. (2008). "Real time speaker recognition using MFCC and VQ," Department of Electronics & Communication Engineering National Institute of Technology Rourkela, 2008.

[18] Mr. P, Kumar , S. L. Lahudkar, “Automatic Speaker Recognition using LPCC and MFCC”, IJRITCC April 2015. 11

[19] Tech, M., & Bansal, A. (2012). "Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB," International Journal of Computer Applications, Volume 45– No.24.

[20] Francis, F. and Vishnu Rajan, V. (2015). “novel noise robust speaker identification system”, ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 17.