Comparison of VQ and GMM for Text Independent Speaker Identification System for The Bengali Language

Md Mahadi Hasan Nahid
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Md Ashraful Islam
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Md Saiful Islam
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

ABSTRACT
Speaker identification (SI) is the system to identify the person by the signal pattern of their voices. In recent years, many speaker identification models are proposed, but till now speaker identification technology do not reach their full potential. This paper presents a comprehensive comparative study of VQ and GMM to identify the speaker who speaks in Bengali accent. We consider the problem of text-independent speaker identification. We compare the performance/accuracy of VQ and GMM based Speaker Identification System (SIS). We use Mel Frequency Cepstral Coefficients (MFCC) and Liner Predictive Coding Coefficients (LPCC) for feature extraction.

General Terms
Pattern Recognition, Algorithms, Machine Learning

Keywords
Bengali Speaker Identification, SI, Voice Recognition, MFCC, LPCC, VQ, GMM.

1. INTRODUCTION
The voice signal carries mainly two types of information. Firstly, the voice signal carries the information of the message or word being spoken. Secondly, the signal also carries the information about the speaker. The main goal of speaker recognition process is to automatically and accurately recognize the speaker using his/her voice signal. Speaker recognition is divided into two parts- speaker identification and speaker verification. In speaker identification, the process determines the identity of the speaker that produced the speech from among a population of speakers. In the speaker verification, the process accepts or rejects the identity claim of the speaker. Based on the text to be spoken, speaker recognition system can also be grouped into text-dependent and text-independent speaker recognition system. In text-dependent speaker recognition systems there need same text in both train and test phase, on the other hand in text-independent speaker recognition system text in test phase and train phase may not be same. [1]

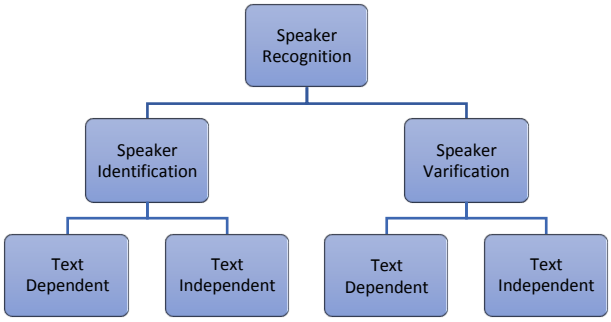

Fig 1: Hierarchy of Speaker Recognition

2. METHODOLOGY
The method for speaker identification is divided into two phases. A training phase (enrolment phase) and a testing phase. In the training phase [Fig 2] we extract the most useful features from speech signal for SI, and train models to get optimal system parameters. The training phase runs in offline.

In identification phase, the same method for extracting features as in the first phase is used for the incoming speech signal, and then the speaker models getting from enrollment phase are used to calculate the similarity between the new speech signal model and all the speaker models in the database. In closed-set case the new speaker will be assigned to the speaker ID which has the maximum similarity in the database. [10]

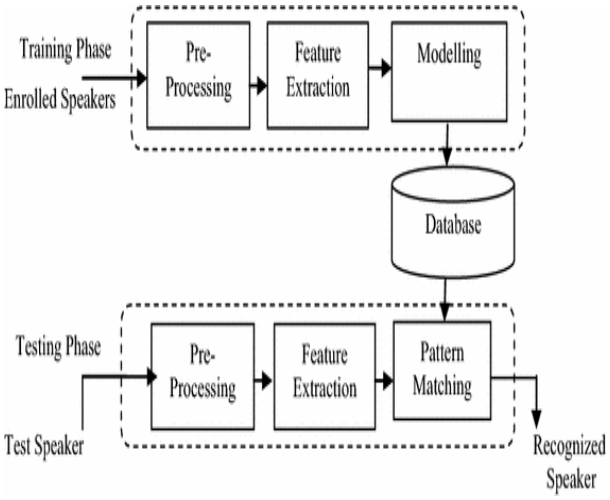

Fig. 2. Training and Testing phases

3. PRE-PROCESSING

We use low pass filter for removing noise from the voice data. A low-pass filter is a filter that passes low-frequency signal and attenuates (reduces the amplitude of) signals with frequencies higher than the cutoff frequency [2][11]. We detected the endpoint of the words from the voice data and removed Silences from the voice data. [2]

4. FEATURE EXTRACTION

4.1 MFCC

The Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression. MFCCs are commonly derived as follows: [3] [9]

$$mel = 2595 * log10\ (1 + \frac{fb}{700}) \quad (1)$$

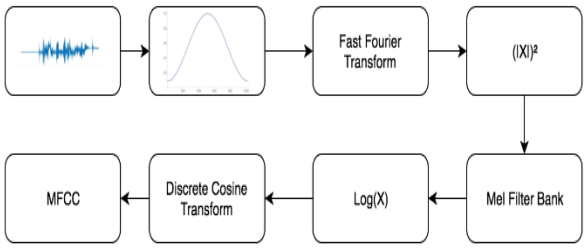

Fig. 3. MFCC Process

• Take the Fourier transform of (a windowed excerpt of) a signal.

• Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows

• Take the logs of the powers at each of the Mel frequencies

• Take the discrete cosine transform of the list of Mel logpowers, as if it were a signal

• The MFCCs are the amplitudes of the resulting spectrum. [12]

4.2 LPCC

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters.

LPC analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples.

LPC is a useful tool for feature extraction as the vocal tract can be accurately modelled and analyzed. Studies have shown that the current speech sample is highly correlated to the previous sample and the immediately preceding samples. LPC coefficients are generated by the linear combination of the past speech samples using the autocorrelation or the auto covariance method and minimizing the sum of squared difference between predicted and actual speech sample y(n) is the predicted x(n) based on the summation of past samples.

$$y(n) = a_1 x(n-1) + a_2 x(n-2) + \cdots + a_n x(n-M) \quad (2)$$

a_i is the linear prediction coefficients M is the number of coefficients and n is the sample. The error between the actual sample and the prediction can then be expressed by

$$\varepsilon(n) = x(n) - y(n) \quad (3)$$

$$\varepsilon(n) = x(n) - \sum_{i=1}^{m} a_i\, x(n-i) \quad (4)$$

$$x(n) = \sum_{i=1}^{M} a_i\, x(n-i) + \varepsilon(n) \quad (5)$$

The speech sample can then be accurately reconstructed by using the LP coefficients and the residual error $\mathcal{E}(n)$. [3]

5. FEATURE MATCHING AND SPEAKER MODELING

5.1 VQ

Vector quantization (VQ) based classification algorithms play an important role in speech independent speaker identification (SI) systems. VQ-based solution is less accurate than the Gaussian Mixture Model (GMM) but it is a simple model for computing. [4]

Fig. 4 shows the VQ based speaker identification system. There are two phases in the VQ based speaker identification system- an offline training sub-system to produce VQ codebooks and an online testing sub-system to generate identification decision. Both sub-systems contain a pre-processing or feature extraction module to convert an audio utterance into a set of feature vectors. [5]

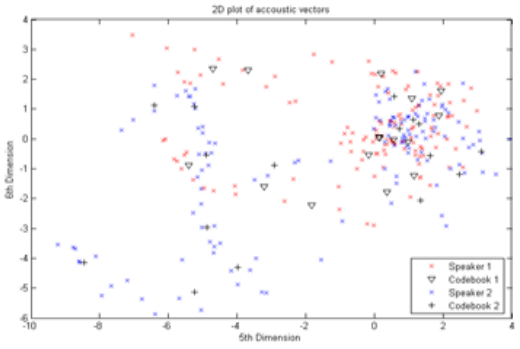

Fig. 4. VQ based speaker identification

In the first phase, the speech signal is decomposed into short fixed-length speech frames, which form the feature vectors. The feature extraction process is described more detailed in the Section 4. The extracted feature vectors are processed further by vector quantization for locating the clusters in the feature space and for reducing the amount of data. The input of vector quantization is the set of feature vectors X and the

output is a codebook C that consists of the cluster centroids, denoted as code vectors. The codebook represents the speaker model by approximating the distribution of the feature vectors in the feature space.

In the second phase, the dissimilarity (d) between the codebook (Ci) and the feature vectors (X) is calculated.

$$C_i = \{C_{i1,}, C_{i2}, \dots\dots, C_{iK}\}; \quad (6)$$

$$X = \{x_1, x_2, \dots\dots, x_L\}; \quad (7)$$

Then we mapped each vector in X to the nearest code vector in Ci and compute the average of these distances:

$$d(X, C) = \frac{1}{L}\sum_{j=1}^{L} min_{k=1}^{K} d_E(X_j, C_{ik}) \quad (8)$$

$$d_E(x, y) = \sqrt{\sum_{i=1}^{dim}(x_i - y_i)^2} \quad (9)$$

Here, d_E is the Euclidean metric. In the recognition part we perform a direct comparison between the set of feature vectors and the codebooks of the known speakers.

5.2 GMM

Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering. A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data. [3][6]

A Gaussian Mixture density is a weighted sum of M component densities, as depicted in Fig. 2 and given by the equation

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i, g((x|\mu_i), \Sigma_i) \quad (10)$$

Where x is a D-dimensional random vector, $b_i(x), i = 1 \dots M$, are the component densities and p_i, $i = 1 \dots\dots M$. are the mixture weights, each component density is a D-variate Gaussian function of the from-

$$g\big((x|\mu_i), \Sigma_i\big) = \frac{1}{2\pi^{D/2}\,|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}\,(x - \mu_i)^T \sum_i^{-1}(x - \mu_i)\right\} \quad (10)$$

Here μ_i is the mean vector and Σ_i is the covariance matrix. The mixture weights satisfy the constraint that

$$\sum_{i=1}^{M} \omega_i = 1 \quad (11)$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation,

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\}, \qquad i = 1, \dots, M \quad (12)$$

There are several variants on the GMM shown in Equation (12). The covariance matrices, Σi, can be full rank or constrained to be diagonal. Additionally, parameters can be shared, or tied, among the Gaussian components, such as having a common covariance matrix for all components, The choice of model configuration (number of components, full or diagonal covariance matrices, and parameter tying) is often determined by the amount of data available for estimating the GMM parameters and how the GMM is used in a particular biometric application. [6]

6. EXPERIMENTS AND RESULTS

We used audio corpus from http://www.voxforge.org/ [13] and Bengali Real Number Audio Dataset [14] for testing this system. There was 10 speaker each has more than 100 samples audio data (16 KHz, .wav format). For testing and simulation, we use MATLAB.

Table 1: System Configuration

Processor	Intel core i5
RAM	8GB
Clock Rate	3.07GHz

Experimental Result of our system is shown in Table 2.

Table 2: Experimental Result

Method	***Accuracy (voxforge dataset) [13]***	***Accuracy (Bengali Real Number Audio Dataset) [14]***	***Average time for identifying a speaker (in seconds)***
GMM +LPCC	87.6 %	89.6 %	0.1505
GMM +MFCC	55.00%	58.0 %	0.1746
VQ +MFCC	97.6%	98.0 %	4.01359
VQ +LPCC	72.0%	73.56%	1.25678

7. CONCLUSION

Experiments show that the method (GMM+LPCC) gives tremendous improvement over the method (GMM+MFCC), and it can detect the correct speaker from much shorter speech samples. Method like VQ+MFCC is highly accurate but slow and it can be applied in security purpose where the number of users is limited. The method GMM+LPCC and VQ+LPCC are very fast, have moderate accuracy and these techniques can be efficiently used in the development of speech recognition systems where the number of users is high.

8. REFERENCES

[1] Ling Feng, "Speaker Recognition", IMM-THESIS: ISSN 1601-233X, Kgs. Lyngby 2004

[2] G. Saha, Sandipan Chakroborty, Suman 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, Khragpur, Kharagpur-721 302, India

[3] Yuan Yujin, Zhao Peihua, Zhou Qun,, "Research of speaker recognition based on combination of LPCC and MFCC", Intelligent Computing and Intelligent Systems (ICIS), IEEE International Conference , vol.3, 29-31 Oct. 2010, pp.765-767. Reynolds, A.D., and Rose, C.R.: "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". IEEE Transactions on Speech and Audio Processing, 3(1): 72-83, 1995.

[4] Ningping Fan, Justinian Rosca, "Enhanced VQ-based Algorithms for Speech Independent Speaker Identification", Siemens Corporate Research Inc., 755 College Road East, Princeton, New Jersey 08540

[5] Douglas Reynolds, "Gaussian Mixture Models" MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.

[6] M.Campbell, D. E. Sturim, D. A. Reynolds: "Support Vector Machines using GMM Super vectors for Speaker Verification", MIT Lincoln Laboratory.

[7] Tomi Kinnunen, Teemu Kilpeläinen And Pasi Fränti "Comparison Of Clustering Algorithms In Speaker Identification", Department Of Computer Science, University Of Joensuu, P.O.Box 111, 80101 Joensuu, Finland.

[8] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617

[9] Evgeny Karpov, "Real-Time Speaker Identification", Master's Thesis, Department of Computer Science, University of Joensuu, Finland, 2003

[10] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617

[11] Kim, Taesun, and Chulhun Seo. "A novel photonic bandgap structure for low-pass filter of wide stopband." IEEE Microwave and Guided Wave Letters 10.1 (2000): 13-15.

[12] Han, W., Chan, C. F., Choy, C. S., & Pun, K. P. (2006, May). An efficient MFCC extraction method in speech recognition. In 2006 IEEE international symposium on circuits and systems (pp. 4-pp). IEEE.

[13] MacLean, K. Voxforge. Ken MacLean. [Online]. Available: http://www. voxforge. org/home. [Acedido em 2016].

[14] Nahid, Md Mahadi Hasan, et al. "Comprehending Real Numbers: Development of Bengali Real Number Speech Corpus." arXiv preprint arXiv:1803.10136 (2018).