IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 6 (June. 2013), ||V2|| PP 32-36

Detailed analysis of Speaker Recognition System and use of MFCCs for recognition

Prerna Puri
Assistant Professor, National institute of Technology, Uttarakhand, India

Abstract: **-** This paper presents the reader with the complete analysis of speaker recognition system with the explanation of each step with their necessity of existence. Every speaker recognition system comprises broadly of two phases namely the enrollment phase and the verification phase. The system in its first phase uses the record function for collecting the voices of different speakers and save them.Later these voices can be used to form a codebook and compare the voice with that in the enrollment phase. Enrollment phase either use feature extraction coefficients(like MFCCs,LPCCs,LPCs etc.)or speaker models like GMM,LDA(Linear Discriminant Analysis) or Factor Analysis(FA)techniques for extracting the features to serve as an identity of the speaker.The verification phase is simply the phase in which the voices in the enrollment and this phase are tested for the likelihood between them and then identify the speaker on the basis of the likelihood ratio.

I. INTRODUCTION

Speaker recognition is a technique which makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Bychecking the voice characteristics of the input utterance one is able to add an extra level of security.The speaker recognition task is basically sub divided into feature extraction and feature matching .Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.

The property that an ideal speaker recognition system should possess includes high inter speaker recognition, low intra speaker variation, features easily measurable, giving correct responses to mimicry, prone to noise and not dependent on other features. Pronunciation pattern, language use comes under the category of high level speaker features but it require a lot of data to get its output. Therefore it is always wiser to go for low level features because not only they can be extracted easily but they do not require a lot of data. In most of the speech and speaker recognition systems MFCC are used because of the fact that these features apart from identifying the frequency distribution also tells the glottal sources and the vocal tract shape and length, which are the features specific to a speaker[1].Speaker recognition is being used a biometric in various security applications where the people are recognized on the basis of their voice.The main tedious task in the speaker recognition is the enrollment phase where the users are asked to input their voice via a microphone or any input device which is further used as a database in the recognition phase for identification of the correct speaker.There are some other techniques also which can be used apart from enrollment phase because it is not always possible to get inputs from various speakers.

MFCC's are used in speaker recognition system because of the ease of implementation and their behavior similar to human speech which is linear at low frequencies and logarithmic at high frequencies.They have high accuracy of recognition and do not vary as recorded sound and while transmitting also their degree of recognition remains unchanged.MFCCs are expressed in the *mel-frequency* scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.In this paper, I have used a feature set, consisting of 12 MFCC, their first and second derivatives, energy and its first derivative. The vector quantization technique is used to obtain the speaker models. Finally, features are sorted according to theirweights, and the K features with greatest average ranks are retained and evaluated.

II. FEATURE SELECTION

There are three sources of variation among speakers:
1. Differences in vocal cords and vocal tract shape.
2. The size **of** the nasal cavity.
3. Differences in speaking style. [2]

For feature set reduction many methods are suggested because on account of identification of a particular speaker we get a lot of features which need to be reduced to only the features capable of representing someone

PCA(Principal Component Analysis) is one such technique where the eigenvectors are calculated which are then sorted in descending order and a projection matrix is built finally known as the Karhunen-Loeve Transforn(KLT)with the largest K eigenvectors.KLT decorrelates the features and gives the smallest possible reconstruction error among all linear transforms, i.e. the smallest possible mean-square error between the data vectors in the original D-feature space andthe data vectors in the projection K-feature [1]

Linear Discriminant Analysis (LDA) attempts to find the transform A that maximizes a criterion of class separability.This is done by computing the within-class and between class variance matrices, W and B, then finding the eigenvectors of W and B, sorting them according to the eigen values, in descending order, and finally building the

projection matrix A with the largest K eigenvectors (which define the K most discriminative hyperplanes). LDA assumes that all classes share a common within-class covariance, anda single Gaussian distribution per class. LDA also assumes that classes are linearly separable. Additionally, as any supervised approach, it requires labelling samples with speaker identities.

Independent Component Analysis (ICA) is a more recent technique that aims to reduce redundancy in the originalfeature space. Whereas PCA removes second order dependencies, ICA removes also higher order dependencies,by minimizing the mutual information between the features, thus projecting them on the directions of maximumindependence. In fact, ICA was originally designed to solvethe problem of blind source separation. Observed signals areassumed to be a linear combination of some unknownstatistically independent non-Gaussian source signals. Thetask of ICA is to recover the source signals from theobserved signals, i.e. to find those directions that are best forseparating the sources. Once the full D*D matrix isestimated, the K projection vectors with greatest L2-normsmay be retained to build the transformationmatrix A.

III. FEATURE EXTRACTION

The purpose of this phase is to convert the speech waveform, using digital signal processing (DSP) tools, to a set of features (at a considerably lower information rate) for further analysis. This is often referred as the *signal-processing front end*.

The speech signal is a slowly timed varying signal (it is called *quasi-stationary*).When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others.

Frame Blocking

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.

Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where *N* is the number of samples in each frame, then the result of windowing is the signal

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

Typically the *Hamming* window is used, which has the form:

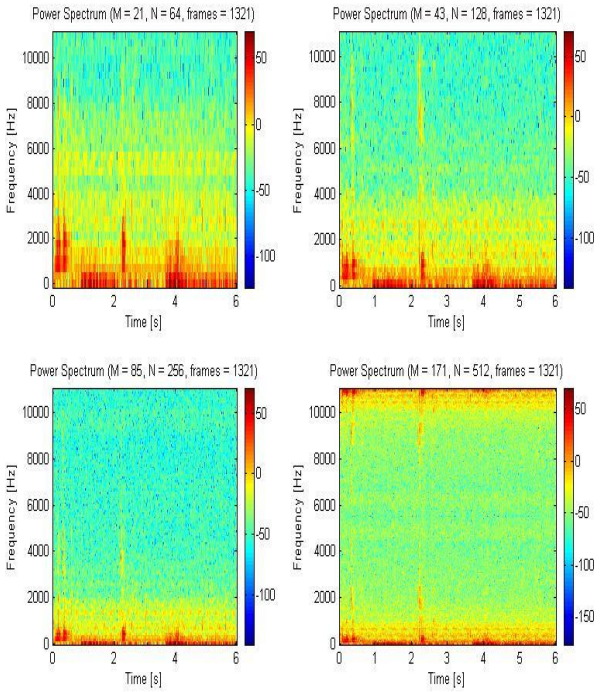
$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

Fast Fourier Transform (FFT)

The next processing step is the Fast Fourier Transform, which converts each frame of *N* samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of *N*samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,..., N-1$$

The result after this step is often referred to as *spectrum* or *periodogram.*The result of my speaker recognition with different values of M and N is shown below.



IV. MEL-FREQUENCY WRAPPING

The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

Feature Matching

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as *feature matching*.

The state-of-the-art in feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this paper I have used the VQ approach, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all codewords is called a *codebook.*

The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance.

Clustering the Training Vectors

After the enrolment session, the acoustic vectors extracted from input speech of each speaker provide a set of training vectors for that speaker. As described above, the next important step is to build a speaker-specific VQ codebook for each speaker using those training vectors. There is a well-known algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of *L* training vectors into a set of *M* codebook vectors.The algorithm is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook $\mathbf{y}_n$ according to the rule

   $$\mathbf{y}_n^+ = \mathbf{y}_n(1+\varepsilon)$$

   $$\mathbf{y}_n^- = \mathbf{y}_n(1-\varepsilon)$$

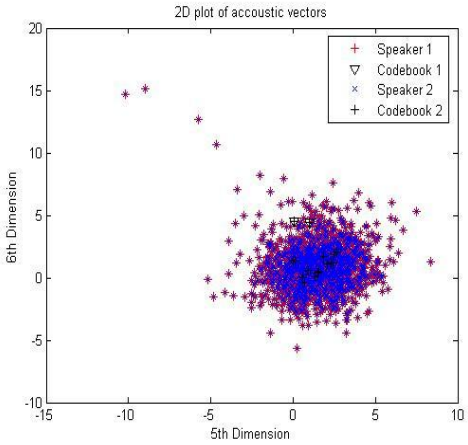
   Where n varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε =0.01).

3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.
5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold
6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed.

Intuitively, the LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained. The longer the recording, the greater the resolution of the codebooks created.For optimum recognition performance, one must ensure that the *.wav voice files are recorded with respect to the following characteristics:

- 8000Hz Sample Frequency
- 8-Bit
- Mono
- Sound file durations ranging from 20 – 30 seconds produce satisfactory results.

Following is the output of the same speaker generating the word “hello” in both testing and enrollment phase. For testing purpose, I have recorded voice files from 6 different speakers with no more than 4 seconds long. The sample rate was 44.1 kHz. First, I trained the system with 5 voice files of different 6 speakers. They all said the same words “Testing 1 2 3 4 5. ”. Among them s1 and s2 are female voices and s3,s4,s5 and s6 are male voices. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. It consists of two main parts.



The first part consists of processing each persons input voice sample to condense and summarize the characteristics of their vocal tracts. The second part involves pulling each person's data together into a single, easily manipulated matrix.The speaker recognition system contains two main modules (i) feature extraction and (ii) feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. The system was tested many times with various databases and found to be very reliable.

V. PERFORMANCE RESULTS

Results from the analysis show that the system is currently operating at a success rate of 83% (10/12). Additional tests with different speakers is required to further increase the accuracy of the performance . Additional tests have revealed that by increasing the threshold value, greatly increases the likelihood that the test

speaker is identified in the codebook, despite not being trained. Tests using short length duration wave files (i.e. 2 - 5 seconds) produce results which are higher in error than their longer duration (> 40 seconds) counterparts. This notion directly correlates to the theory of Vector Quantization where a larger set of vectors produces a more accurate representation of a speakers voice.

VI. CONCLUSION

In this work, I have made the study of Speaker recognition system and also developed a text-independent speaker identification system that is a system that identifies a person who speaks regardless of what is saying. A typical speaker verification system consists of two sections(i) Enrolment section to build a database of known speakers and (ii) Unknown speaker identification system. Enrollment session is also referred to as training phase while the unknown speaker identification system is also referred to as the operation session or testing phase.

REFERENCES

[1] Maider Zamalloa, Germacn Bordel, Luis Javier Rodriguez, Mikel Penagarikano ,Feature Selection Based on Genetic Algorithms for Speaker Recognition(2006)
[2] Ehab F. M. F. Badran, Hany Selim,Proceedings of ICSP2000,Speaker Recognition Using Artificial Neural Networks Based on Vowel phonemes
[3] Xing Fan and John H. L. Hansen, *Fellow, IEEE,*Speaker Identification Within Whispered Speech Audio Streams, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 5, JULY 2011
[4] Mary A. Kohler', Walter D. Andrews, Joseph P. Campbell2, and Jaime Hernhdez-Cordero,
[5] Phonetic speaker Recognition(2001 IEEE)
[6] Ran D. Zilca, *Senior Member, IEEE*, Brian Kingsbury, *Member, IEEE*, Ji˘rí Navrátil, and
[7] Ganesh N. Ramaswamy, *Member, IEEE,*Pseudo Pitch Synchronous Analysis of Speech With Applications to Speaker Recognition,IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 2, MARCH 2006
[8] Pongtep Angkititrakul, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE,*Discriminative In-Set/Out-of-Set Speaker Recognition IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 2, FEBRUARY 2007
[9] Vinod Prakash, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE,*In-Set/Out-of-Set Speaker Recognition Under Sparse Enrollment,IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007
[10] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.
[11] Ji Ming, *Member, IEEE*, Timothy J. Hazen, *ember, IEEE*, James R. Glass, *Senior Member, IEEE*, and Douglas A. Reynolds, *Senior Member, IEEE,*Robust Speaker Recognition in Noisy Conditions
[12] IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 5, JULY 2007