IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331
PP 51-56
www.iosrjournals.org

Speaker Identification system using Mel Frequency Cepstral Coefficient and GMM technique

Om Prakash Prabhakar [1], Navneet Kumar Sahu [2]
[1](*Department of Electronics and Telecommunications, C.S.I.T.,Durg,India)*
[2](*Department of Electronics and Telecommunications, C.S.I.T.,Durg,India)*

ABSTRACT : *The performance of speech recognition systems have improved due to recent advances in speech processing technique but there is still need of improvement. In this paper we present the hybrid approach for feature extraction technique using MFCC & LPC, two classification techniques, Gaussian mixture models (GMM) and Vector quantization (VQ) with LBG design algorithm are used for classification of speakers.The Vector Quantization (VQ) approach is used for 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. After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. LBG algorithm due to Linde, Buzo and Gray is used for clustering a set of L training vectors into a set of M codebook vectors. For comparison purpose, the distance between each test codeword and each codeword in the master codebook is computed. The difference is used to make recognition decision. The entire coding was done in MATLAB and the system was tested for its reliability.*

Keywords - F*eature extraction, feature matching, MFCC,LPC,GMM,VQ*

I. INTRODUCTION

Speech being a natural form of communication advancements in scientific technology have made it possible to use this in security systems. Speaker recognition is a process that enables machines to understand and interpret the human speech by making use of certain algorithms and verifies the authenticity of a speaker with the help of a database. First, the human speech is converted to machine readable format after which the machine processes the data. The data processing deals with feature extraction and feature matching. Then, based on the processed data, suitable action is taken by the machine. The action taken depends on the application. Every speaker is identified with the help of unique numerical values of certain signal parameters called 'template' or 'code book' pertaining to the speech produced by his or her vocal tract. Normally the speech parameters of a vocal tract that are considered for analysis are (i) formant frequencies, (ii) pitch, and (iii) loudness.

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. MFCC is perhaps the best known, robust, accurate and most popular. The Mel frequency scale is linear frequency spacing below 1000Hz and logarithmic spacing above 1000Hz. In other words, frequency filters are spaced linearly at low frequencies and are logarithmically at high frequencies which have been used to capture the phonetically important characteristics of speech. This is an important property of a human ear. Hence the MFCC processor mimics the human ear of perception. This is the process of feature extraction. Pattern recognition does the job of feature extraction which 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.

A generic speaker recognition system is shown in Fig. 1**.** In Fig. 1, the desired features are first extracted from the speech signal. The extracted features are then used as input to a classifier,which makes the final decision regarding verification or identification.

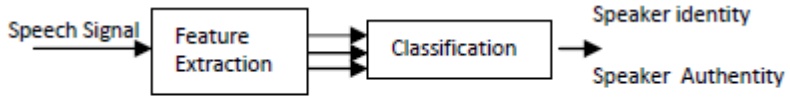


Fig.1. Speaker Recognition system

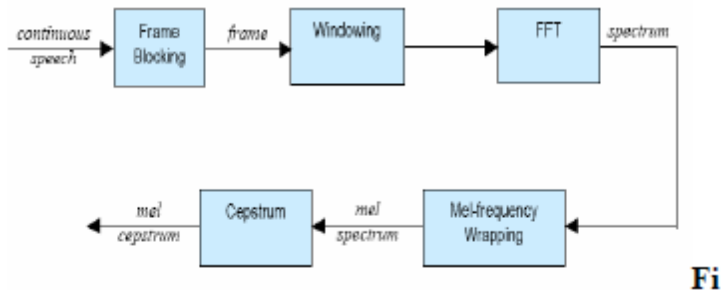
IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331
PP 51-56
www.iosrjournals.org

II. FRONT END PROCESSING / FEATURE EXTRACTION

Speech front-end processing consists of transforming the speech signal to a set of feature vectors. The aims of this process are to obtain a new representation which is more compact, less redundant, and more suitable for statistical modeling. Feature extraction is the key to front-end process; it mainly consists in a coding phase. The attributes of features that are desirable for speaker verification systems are [1]

- Easy to extract, easy to measure, occur frequently and naturally in speech
- Not affected by speaker physical state
- Not change over time and utterance variations (fast talking vs. slow talking rates)
- Not affected by ambient noise
- Not subject to mimicry

In this paper, we are focusing in Mel Frequency Cepstral coefficients (MFCC). Mel Frequency Cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) [2] are the most popular acoustic features used in speech recognition. Often it depends on the task; this method leads a better performance Due to the high performance of MFCC, this technique has been chosen as front-end processing for this research. MFCC are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. Several step of MFCC are described in these following phases show in fig.2.



Fi

Fig 2 MFCC Processing

A. Frame Blocking

Framing is the first applied to the speech signal of the speaker. The signal is partitioned or blocked into N segments (frames).

B. 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.

C. Fast Fourier Transform

Next step is the Fast Fourier Transform which converts each frame of N samples in time domain to frequency domain.

D. Mel-Frequency Wrapping

The spectrum obtained from the above step is Mel Frequency Wrapped; the major work done in this process is to convert the frequency spectrum to Mel spectrum.

E. Cepstrum

In this final step, we convert the log Mel spectrum back to time. The result is called the Mel frequency Cepstrum coefficients (MFCC).

III. back end processing / pattern matching

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern matching. The goal of pattern matching 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.
Many forms of pattern matching and corresponding models are possible. Pattern-matching methods include dynamic time warping (DTW), the hidden Markov model (HMM), artificial neural networks (ANN), and Gaussian Mixture Models (GMM). Template models are used in DTW whereas statistical models are used in HMM. In this paper, we are focusing and discussing in GMM.

IV. GAUSSIAN MIXTURE MODEL APPROACH

This section describes the form of the Gaussian mixture model (GMM) and motivates its use as a representation of speaker identity for speaker recognition. The speech analysis for extracting the MFCC feature representation used in this work is presented first. Next, the Gaussian mixture speaker model and its parameterization are described. The Gaussian mixture model (GMM) is a density estimator and is one of the most commonly used types of classifier. The implementation of the maximum likelihood parameter estimation and speaker
identification procedures is described. The classification stage uses the Gaussian Mixture Model (GMM) shown in Fig. 3

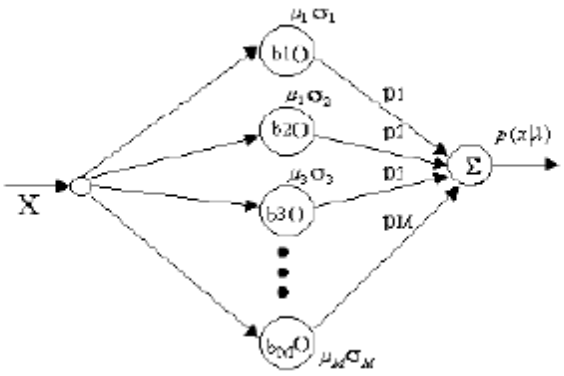


Fig 3. A Gaussian mixture density is a weighted sum of Gaussian densities, where p_i, $i = 1, \ldots, M$, are the mixture weights and $b_i()$, $i = I, \ldots, M$, are the component Gaussians.

Model Description
A Gaussian mixture density is a weighted sum of *M* component densities, as depicted in Fig. 3 and given by the equation

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}), \text{ with } \sum_{i=1}^{M} p_i = 1 \qquad (1)$$

where **x** is a random vector of D-dimension, λis the speaker model, pi are the mixture weights, bi(**x**) are the density components, that is formed by the mean μ and covariance matrix σi to i = 1,2,3,….M., and each density component is a D-Variate- Gaussian distribution of the form

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \qquad (2)$$

The mean vector, μ□ , variance matrix,σi, and mixture weights pi of all the density components, determines the complete Gaussian Mixture Density

$$\lambda = \{\mu, \sigma, p\}, \qquad (3)$$

used to represent the speaker model. To obtain an optimum model representing each speaker we need to calculate a good estimation of the GMM parameters. To do that, a very efficient method is the Maximum-Likelihood Estimation (ML) approach. For speaker identification, each speaker is represented by a GMM and is referred to by his/her modelλ

Maximum Likelihood Parameter Estimation

Given training speech from a speaker, the goal of speaker model training is to estimate the parameters of the GMM, x, which in some sense best matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM [17]. By far the most popular and well-established method is maximum likelihood (ML) estimation. The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM, given the training data. For a sequence of *T* training vectors *X* = {X1... XT*}*, the GMM likelihood can be written

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda). \qquad (4)$$

ML parameter estimates can be obtained iteratively using a special case of the expectation-maximization (EM) algorithm [18]. The basic idea of the EM algorithm is, beginning with an initial model, x, to estimate a new model 1, such that *p(X* Iλ) ≥ *p(X* Iλ). The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. This is the same basic technique used for estimating HMM parameters via the Baurn-Welch re-estimation algorithm. On each EM iteration, the following re-estimation formulas are used which guarantee a monotonic increase in the model's likelihood value:

Mixture Weights:

$$\bar{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \qquad (5)$$

Means:

$$\vec{\bar{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \qquad (6)$$

Variances:

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) x_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \bar{\mu}_i^2 \qquad (7)$$

Where σi^2, XT and μi, refer to arbitrary elements of the vectors σi^2, XT and μi, respectively.

The *a posteriori* probability for acoustic class *i* is given by

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)}. \qquad (8)$$

Two critical factors in training a Gaussian mixture speaker model are selecting the order M of the mixture and initializing the model parameters prior to the EM algorithm.

V. EXPERIMENT

For the experimental results we first recorded the sound of any digit .Then we go for speech detect to identify whether the recording has been done correctly or not . The we train the system to prepare a dat base of different digits. Finally we go in for recognition . We found that the system identifies the correct digit. The results are shown .

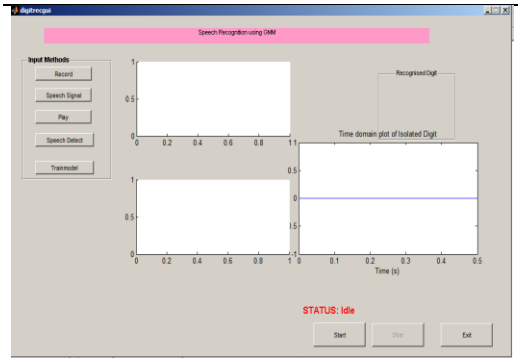


Figure 4 Main GUI window

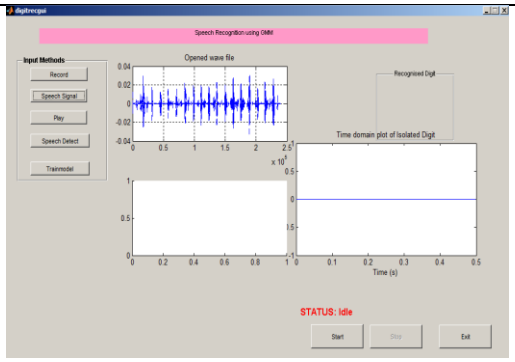


Figure 5 Window showing sampled speech signal

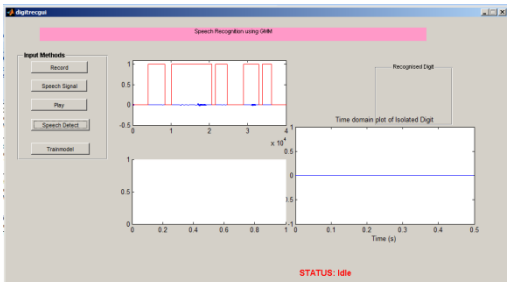


Figure 6 Window showing the detected speech signal

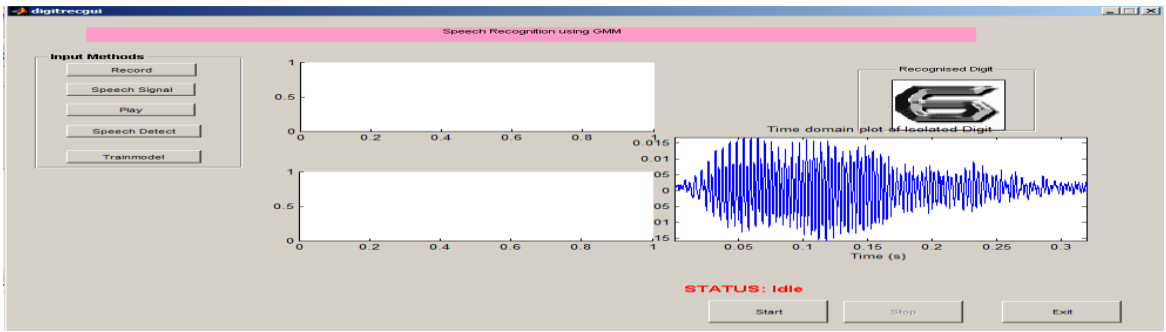


Figure 7 Window showing the correctly recognized digit.

VI. CONCLUSIONS

Thus in this paper we have been experimentally able to recognize the digits correctly. Further work will include to train and recognize other word apart form the digits and we can go in to analyze the efficiency of the system.

References

[1]C.E. Vivaracho, J. Ortega-Garcia, L. Alonso, Q.I. Moro, "A Comparative Study of MLP-based Artificial Neural Networks in Text-Independent Speaker Verification against GMM-based Systems", EUROSPEECH 2001-SCANDINAVIA, Aalborg Denmark, Volume 3, pp. 1753-1756, September 2001
[2] Campbell J.P. and Jr. "Speaker recognition: A Tutorial*" Proceeding of the IEEE. Vol 85, 1437- 1462 1997.*
[3] S.Furui. "Fifty years of progress in speech and speaker recognition," *Proc. 148th ASA Meeting, 2004.*
[4] A. Rosenberg, "Automatic speaker recognition: A review," Proc. IEEE, vol. 64, pp. 475487, Apr. 1976.
[5] G. Doddington, "Speaker recognition-Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651-1664, 1985
[6] Douglas A. Reynolds, *Member, IEEE*, and Richard C. Rose, *Member, IEEE*, "Robust Text- Independent Speaker Identification Using Gaussian Mixture Speaker Models", *TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, 1995*
[7] J. Hertz, A. Krogh, and R. J. Palmer, *Introduction to the Theory of Neural Computation*, Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, Reading, Mass, USA. 1991.

[8] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, NY, USA, 1994.
[9] H. Bourlard and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167– 1178, 1990.
[10] J. Oglesby and J. S. Mason, "Optimization of neural models for speaker identification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '90), vol. 1, pp. 261–264*, Albuquerque, NM, USA, April 1990.
[11] Y. Bennani and P. Gallinari, "Connectionist approaches for automatic speaker recognition," in *Proc. 1st ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 95 102*, Martigny, Switzerland, April 1994.
[12] J. M. Naik and D. Lubenskt, "A hybrid HMMMLP speaker verification algorithm for telephone speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '94), vol. 1, pp. 153–156*, Adelaide, Australia, April 1994.
[13] Reynolds, D., and Heck, L.P., "Automatic Speaker Recognition", *AAAS 2000 Meeting, Humans, Computers and Speech Symposium, 2000.*
[14] Davis, S. B. and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustic, Speech and Signal Processing, ASSP-28, No. 4, 1980.*
[15] G. McLachlan, *Mixture Models.* New York: Marcel Dekker, 1988.
[16] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Stat. Soc., vol. 39, pp. 1-38, 1977.
[17] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing.* Cambridge, MA: MIT Press, 1986.