ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Performance Improvement of Human Voice Recognition System using Gaussian Mixture Model

Om Prakash Prabhakar [1], Navneet Kumar Sahu [2]

Department of Electronics and Telecommunications, C.S.I.T., Durg, India[1]

Department of Electronics and Telecommunications, C.S.I.T., Durg, India[2]

Abstract: The voice recognition system is most prominent technique of identification of human voice. This is used for security purposes. Recent research concentrates on developing systems that would be much more robust against variability in environment, speaker and language. Hence today's researches mainly focus on ASR systems with a large vocabulary that support speaker independent operation with continuous speech in different languages. The performance of Speaker recognition systems has improved due to recent advances in speech processing techniques but there is still need of improvement. In this paper we present the recognition performance of numeric digits and speech alphabets. The Gaussian mixture model is used for classification of the speech signal. To improve, we use hybrid concentration of Mel Frequency Cepstral Coefficients (MFCC) and Linear predictive coding (LPC). The entire coding was done in MATLAB and the system was tested for its reliability.

Keywords: Feature extraction, feature matching, MFCC, LPC, Gaussian mixture model (GMM)

I. INTRODUCTION

Voice recognition is the translation of spoken words into text. It is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Some SR systems use "speaker independent speech recognition" while others use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent" systems.

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 using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model. Speech signal contains several levels of information. At first it contains information about the spoken message. At second level speech signal also gives information about the speaker identity, his emotional state and so on. The task of speaker recognition can be divided into two parts: speaker identification and speaker verification. Speaker identification is answering the question which one of the group of known voices best matches the input voice. Speaker verification is answering the question is really this person who claims to be. Also speaker recognition can be text dependent or text independent. In text dependent speaker recognition, speech recognition is performed too and there are used the same methods as in speech recognition.

In this paper Gaussian mixture model [1] [2] classifiers is used for classifying speakers into their respective classes. Prior to construction of GMM for each speaker, speech signal is first transformed into a set of spectral vectors which is a convenient representation of a person's vocal tract structure and would constitute an important factor distinguishing one person's voice from another. MFCC [3] features have been used by various researchers for speech recognition, speaker recognition and languages identification purposes**.** Probability density function consisting of 4 mixtures is illustrated in Fig.1b, [10].

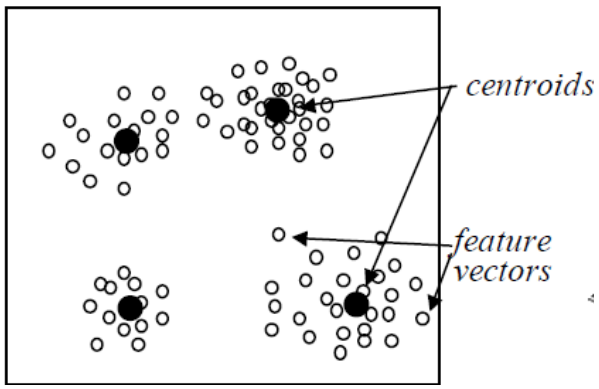


Fig-1(a)

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

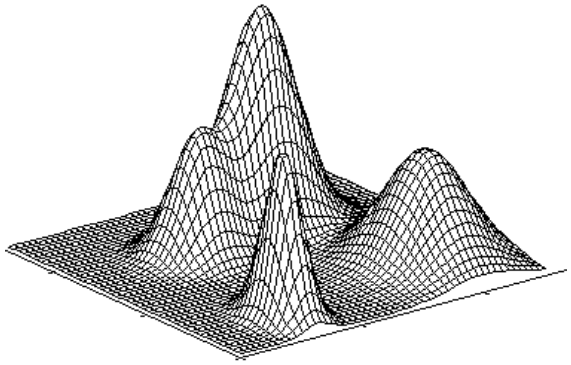


Fig-1(b)
Fig-1. GMM – Gaussian mixture model

a) illustration of feature vectors distributed in space,
b) corresponding model with 4-mixtures

II. FEATURE EXTRACTION

MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency Cepstral these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as

$$m=2595 \log_{10} (1+f/ 700) \quad (1)$$

Mel scale and normal frequency scale is referenced by defining the pitch of 1000 Mel to a 1000 Hz tones, 40 db above the listener's threshold. Mel frequency are equally spaced on the Mel scale and are applied to linear space filters below 1000 Hz to linearized the Mel scale values and logarithmically spaced filter above 1000 Hz to find the log power of Mel scaled signal [4]. Mel frequency wrapping is the better representation of voice. Voice features are represented in MFCC by dividing the voice signal into frames and windowing them then taking the Fourier transform of a windowing signal. Mel scale frequencies are obtained by applying the Mel filter or triangular band pass filter to the transformed signal. Finally transformation to the discrete form by applying DCT presents the Mel Cepstral Coefficients as acoustic features of human voice.

III. PATTERN MACHTING

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. 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.

In this final step, the log mel spectrum is converted 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. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$C_n = \sum_{k=1}^{k} (\log S_k) \cos \left\{ n\left(k-\frac{1}{2}\right) * \frac{\pi}{k} \right\}, \quad n = 1, 2, \ldots k \quad (2)$$

GAUSSIAN MIXTURE MODEL APPROACH

The Gaussian mixture model (GMM) is a density estimator and is one of the most commonly used types of classifier [5], [6], [9]. In this method, the distribution of the feature vector x is modeled clearly using a mixture of M Gaussians. A Gaussian mixture model is modeled by many different Gaussian distributions. Each of the Gaussian distribution has its mean, variance and weights in the Gaussian mixture model.

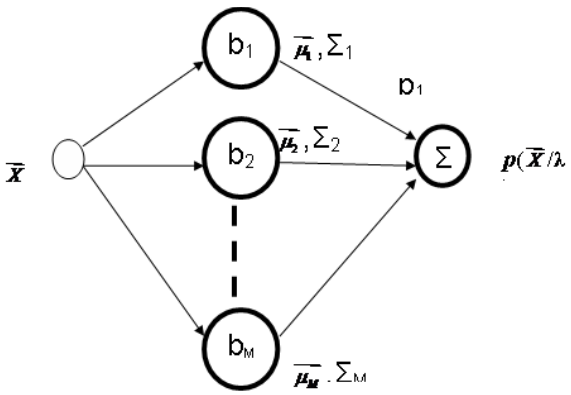


Fig- 2. M probability densities forming a GMM

Gaussian Mixture density is weighted sum of M component densities and can be expressed:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \quad (3)$$

where x is D dimensional vector, pi is the component weight, bi (x) - component densities, that can be written:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left| \sum_i \right|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \sum_i^{-1} (\vec{x}-\vec{\mu}_i)}, \quad (4)$$

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

where *i* - mean vector, *i* - covariance matrix. Mixture weights must satisfy constraint:

$$\sum_{i=1}^{M} p_i = 1 . \qquad (5)$$

Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights. All these parameters are represented by notation:

$$\lambda = \left\{ p_i, \mu_i, \sum_i \right\} \quad i = 1,2,...,M . \qquad (6)$$

So, each speaker is represented by his/her GMM and is referred by his/her model λ The other task is to estimate the parameters of GMM λ, which best matches the distribution of the training feature vectors, given by speech of the speaker. There are several available techniques for GMM parameters estimation [7]. The most popular method is maximum likehood (ML) estimation [8]. The basic idea of this method is to find model parameters which maximaze the likehood of GMM. For a given set of T training vectors X {$x1$,..., xT } □ □ □ GMM likehood can be written:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) . \qquad (7)$$

ML parameter estimates can by obtained iteratively using special case of expectation-maximization (EM) algorithm. There the basic idea is, beginning with initial model □ , to estimate a new model □ , that $p(X \mid \square) \square p(X \mid \square)$. The new model then becomes the initial model for the next iteration. This process is repeated until some convergence threshold is reached. On each iteration, following reestimation formulas are used: mixture weights are recalculated

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) . \qquad (8)$$

Means are recalculated

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} . \qquad (9)$$

Variances are recalculated

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)(x_t - \mu_i)^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} . \qquad (10)$$

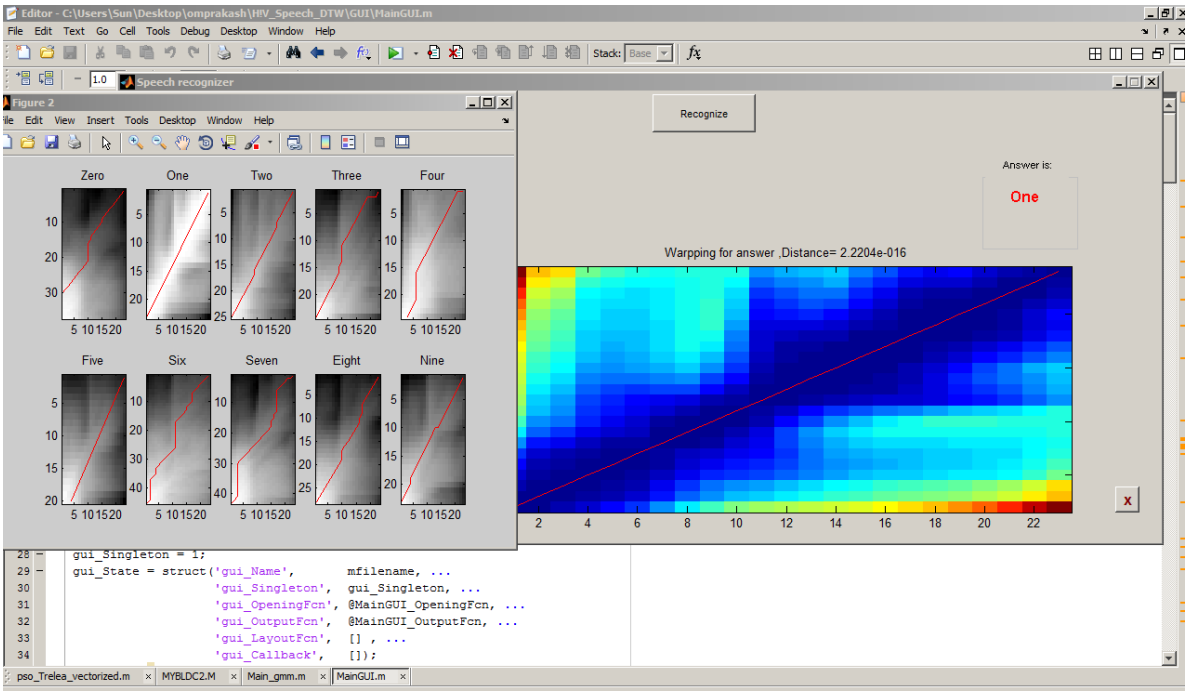
The a posteriori probability for acoustic class i is given by:

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} . \qquad (11)$$

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.

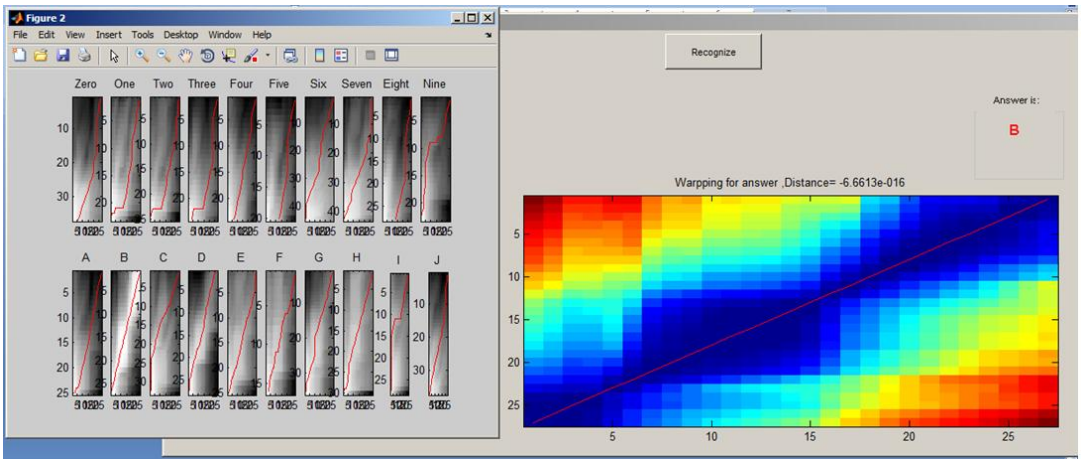
IV. EXPERIMENTAL RESULTS

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. Thee we train the system to prepare a data base of different digits. Finally we go in for recognition. We found that the system identifies the correct digit and alphabet. The system is then used to identify the digits from 0 to 9 and alphabets from A to Z the results are shown.



ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014



The experimental result shows the recognition rate of digits 0 to 9 is from 92% to 99% and also the recognition performance of speech alphabet A to Z is showing 92%to 100%. The graph shows the recognition performance of numeric digit 0 to 9 and alphabet A to Z.

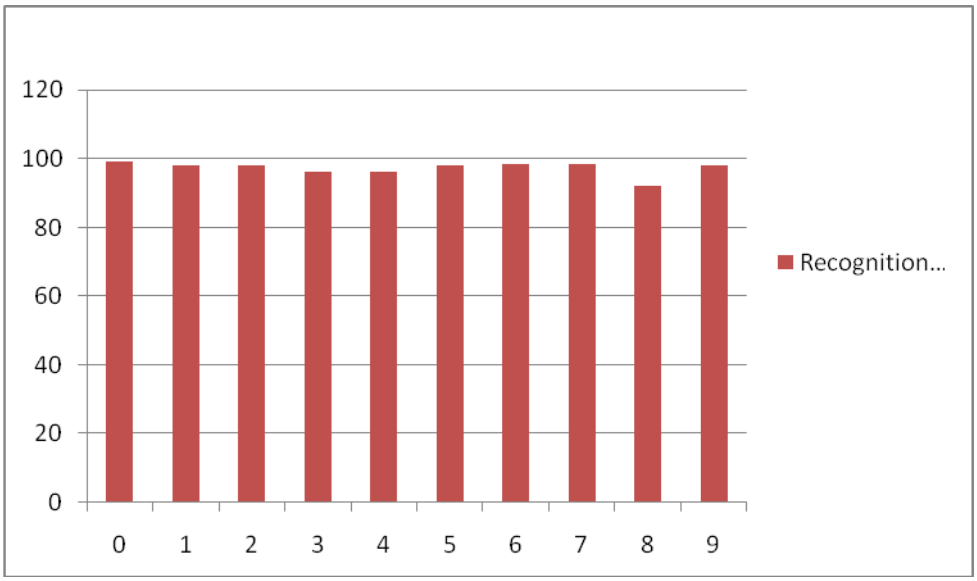


Fig-3. Recognition rate of digit 0 to 9

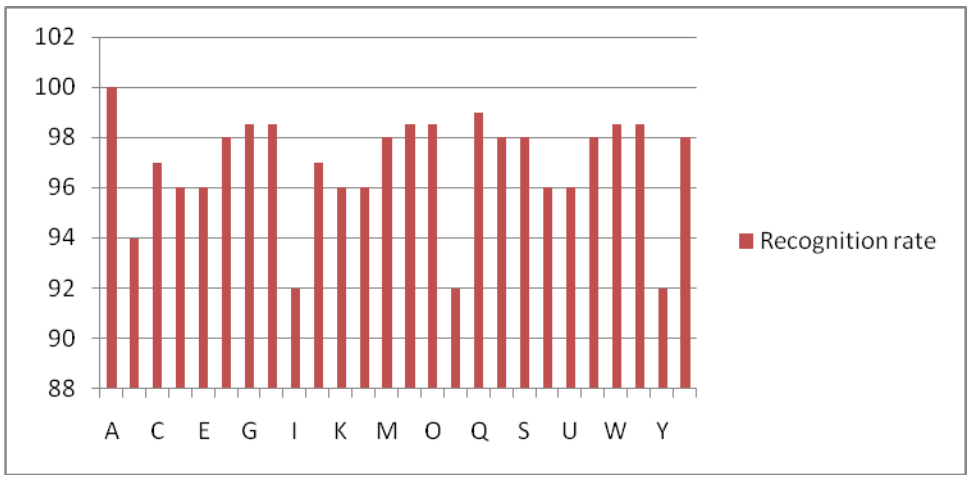


Fig-4. Recognition rate of speech alphabet A to Z

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

V. CONCLUSION

Thus in this paper using the proposed methods we were able to use the system to recognize the digits from zero to nine and alphabets from A to Z. The proposed system is suitable for highly secured environments. Even with this high population the system performed well since it has produced comparatively good performance than the existing algorithms. The proposed system can be extended to recognition of sentences and there by completing the communication system.

REFERENCES

[1] Bing Xiang, and Toby Berger, "Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Model and Neural Network" IEEE Transactions On speech and Audio Processing, Vol. 11, No. 5, pp. 448-449, 2000.
[2] Douglas A. Reynolds, and Richard C.Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp. 74-77, 1995.
[3] K.R. Aida–Zade, C. Ardil and S.S. Rustamo, "Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems" World Academy of Science, Engineering and Technology, pp . 74-77, 2006.
[4] Anjali Bala, Abhijeet Kumar and Nidhika Birla, "Voice Command Recognition System Based On MFCC and DTW" Anjali Bala et al. / International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342
[5] Reynolds, D. A. and Rose, R. C. "Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans.Speech Audio Process. 3, 1995, pp 72–83
[6] Charles B.de Lima,Abraham Alcaim and Jose A.Apoloniyaro Jr "Text independent speaker verification using GMM"
[7] McLachlan G. Mixture Models. – New York: Marcel Dekker, 1988.
[8] Dempster A., Laird N., and Rubin D. Maximum likelihood from incomplete data via the EM algorithm // J. Royal Srar. Soc. – 1977. – Vol. 39. – P. 1–38.
[9] Cheang Soo Yee and Abdul Manan Ahmad, "Mel Frequency Cepstral Coefficients for Speaker Recognition Using Gaussian Mixture Model-Artificial Neural Network Model" University of Technology Malaysia.
[10] Petr David, "Experiments with Speaker Recognition using GMM" Technical University of Liberec, Hálkova 5, 461 17 Liberec.
[11] M. Pandit and J. Kittler, "Feature selection for a dtw-based speaker verification system, in *Proceedings of IEEE Int. Conf. Acoust. Speech and Signal Processing*, 2**:** 769-772 (1998).
[12] G. Doddington, "Speaker recognition-Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651-1664, 1985

.