A Novel Method for Text-Independent Speaker Identification Using MFCC and GMM

M.S.Sinith, Anoop Salim, Gowri Sankar K, Sandeep Narayanan K V, Vishnu Soman
sinith@ieee.org,anoopsalim@yahoo.com,gowrisankar707@gmail.com,
sandeepnarayanankv@ieee.org,vishnushowman@gmail.com
Dept. of Electronics and Communication Engineering,Amrita Vishwa Vidyapeetham, Amritapuri, Kerala, India

Abstract

The area of speaker recognition is concerned with extracting the identity of the person speaking. Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification can be Text-Independent or *Text-Dependent. In this paper we lay emphasis on text-Independent speaker identification system where we adopted Mel-Frequency Cepstral Coefficients (MFCC) as the speaker speech feature parameters in the system and the concept of Gaussian Mixture* Modeling (GMM) for modeling the extracted speech *feature. We used the Maximum Likelihood Ratio Detector algorithm for the decision making process. The experimental study has been performed for various speech time duration and several languages and was conducted around MATLAB 7 language environment. Gaussian mixture speaker model attains high recognition rate for various speech durations.*

1. Introduction

Speech signal provides several levels of information. It conveys the words and messages being spoken and also provides the identity of the speaker. Speaker recognition is the process of automaticallyrecognizing who is speaking by using the speaker specific information included in speech waves to verifyidentities being claimed by people accessing systems; that is, it enables access control of various services by voice[1]. This process finds its application extensively in biometrics, voicedialing telephone banking etc.

Speaker recognition can be classified into speaker identification and speaker verification:Speaker identification is the process of determining from which of the registered speakers a given utterance comes.Speaker verification refers to whether or not the speech samples belong to some specific speaker. The verification is the number of decision alternatives. In identification, the number of decision alternatives is dependent to the size of the population, whereas in verification there are only two choices, acceptance or rejection, regardless of the population size.Speaker recognition methods can also be divided into text-dependent (fixed passwords) and text-independent (no specified passwords) methods. The former require the speaker to provide utterances of key words or sentences, the same text being used for both learning and recognition, whereas the latter do not rely on a specific text being spoken.

The rest of this paper is organized as follows: Section 2 presents a brief description of the suggested architecture for our automatic speaker recognition system. Section 3 describes about the pre-processing done on the speech file. Section 4 is devoted to the mathematical formulation of MFCC parameterization. Section 5 describes about the modeling GMM algorithms used. Section 6 explains about the decision making process process used in the system. Experiments and performances evaluation of our system are presented in Section 7. Section 8 concludes and presents perspectives of this study. The last section lists the main references which we have used in this work.

2. Speaker identification system-architecture

The scheme at figure 1 represents the basic elements of our Speaker Identification System .This scheme

contains four main modules: Pre-processing module: It gives an initial treatment to the input speech signal making it desirable for the next module. Feature extraction module: it is responsible for the acoustic analysis of voice signal. Thus for each time signal, we extracted a matrix equivalent of features vectors Modelization module: it determines the models parameters from those extracted at the previous module. In the decision module following the discrimination between speakers will be made on the basis of these models Decision module: a decision on the identity of a speaker is taken on the basis of a similarity measure between his test model and all models of reference contained in the database.

3. Preprocessing

Preprocessing is considered as the first step of speech signal processing, which involve with the analog signal to digital signal conversion which had been described by E. C. Gordon(1998) [2]. The silence has been removed from the speech signal before any processing has been done on it. The signal is then sampled at a rate of 23000Hz. The speech samples are then segmented into frames of the time length within the range of 20-40msec, also known as Framing. Framing enables the non-stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behavior within the short time period of 20-40msec. each individual frame is windowed in order to minimize the signal discontinuities at the beginning and the end of each frame. Here, hamming window most commonly used as window shape in speech recognition technology, by considering the next block in the feature extraction processing chain, integrates all the closest frequency lines. Impulse response of the Hamming window is shown in the equation below:

$$w(n)=\left\{\begin{matrix}0.54-.046\cos\frac{2\pi n}{N-1},0\le n\le N-1\\0,otherwis\end{matrix}\right\} \quad (1)$$

Where N is the number of samples in each frame

4. Feature extraction using MFCC

Feature Extraction deals with extracting the features of speech from each frame and representing it as a vector . The feature here is the spectral envelope of the speech spectrum which is represented by the acoustic vectors.MFCC (Mel Frequency Cepstral Coefficients) is the most common technique for feature extraction which computed on a warped frequency scale based on known human auditory perception.Based on human perception experiments it is observed that human ear acts as filteri.e.it concentrates on only certain frequency components.Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale.Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale [3],[4].The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz:

$$mel(f) = 2595*\log10(1+f/700) \quad (2)$$

The idea act as follows as in figure 1.

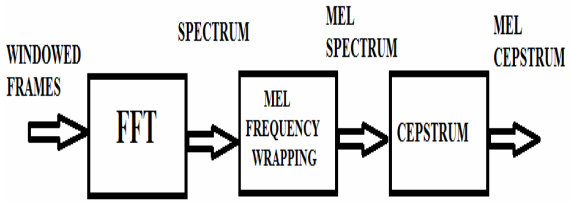

Fig.1. Obtaining mel cepstrum from windowed frames

Particularly, for the filter banks implementation, the magnitude coefficient of each Fourier Transform speech segment is binned by correlating them with each triangular filter in the filterbank. To perform Mel-scaling, 24 triangular filters having triangular frequency response.The triangular filters are used as they cansmoothen the harmonics .There are 10 filters spaced linearly below 1000 Hz, and the remaining filters spread logarithmically above 1000 Hz.The results of the FFT will be information about the amount of energy at each frequency band. Human hearing, however, is not equally sensitive at all frequency bands. It is less sensitiveat higher frequencies, roughly above 1000 Hertz. It turns out that modeling this property of human hearing during feature extraction improves speech recognition performance.

Peak (referred to as formants) in the spectrumdenote dominant frequency components in the speech signal and carry the identity of the sound.Formants are connected by smooth curve-referred to as spectral envelope.The spectral envelope is used for speaker Identification and Spectral details are used for speech identification. We now need to separate spectral envelope and spectral details from spectrum.The next step is to take the logarithm which simply converts the multiplication of the magnitude in the Fourier

transform into addition making the extraction of the formants simpler.

In general, the human response to the signal level of logarithmic. It is because humans are less sensitive to slight differences in amplitude at high amplitudes compared to the low amplitudes. In addition, using a log makes the feature estimates less sensitive to variations in input (for examplepower variations due to the speaker's mouth moving closer or further from the microphone) [5]. The final procedure for the Mel Frequency cepstral coefficients (MFCC) computation is to convert the log mel spectrum back to time domain where we get the so called the mel frequency cepstral coefficients (MFCC). Because the mel spectrum coefficients are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT) and get a featured vector.The dct compresses these coefficients to 13 in number.

This features vector is considered as an input for the next stage, which are concern with training the features vector and pattern recognition. The cepstrum is more formally defined as the DCT of the log magnitude of the DFT of a signal and is given by:

$$C(u) = a(u)\sum_{x=0}^{N-1} f(x)\cos\left[\frac{\pi(2x+1)u}{2N}\right] (3)$$

for u=0,1,2…N-1 ,x=0,1,2…N-1 and $a(u)$ is defined as

$$X = (\vec{x_1}, \vec{x_2}, \ldots \vec{x_T})$$

$$a(u) = \begin{cases} \sqrt{\frac{1}{N}} & for \quad u = 0 \\ \sqrt{\frac{2}{N}} & for \quad u \neq 0 \end{cases}$$

5. Modelization using GMM

In speaker recognition, there are two types of modeling methods: The deterministic methods (Dynamic Time Warping DTW and Quantization Vector QV) and statistics methods(Gaussian Model Mixture GMM and HiddenMarkov Model HMM). These last are the most used inthis field. In this study, we have chosen to use a modeling based on Gaussian mixture model GMM and we have adapted it to the identification of speakers. A Gaussian mixture density is a weighted sum ofM component densities, and is given by

$$P(\overline{X})P(\overline{X}|\lambda) = \sum_{t=1}^{M} P_1 B_1(\overline{X}) \qquad (4)$$

Where *xv* is an *N*-dimensional random vector, $B_i(\overline{X}), i = 1 \ldots M$ are the component densities and *P* ,*i* 1,...,*M*, *i* are the mixture weights. Each component density is a*N*-variate Gaussian function of the form:

$$B_i(\bar{x}) = \frac{1}{(2\pi)^{N/2}\sum |2/z} \exp\{\frac{1}{2}(\vec{x} - \vec{\mu_1})\sum^{-1}(\vec{x} - \vec{\mu_1})\} \quad (5)$$

with mean vector $\vec{\mu_1}$ and covariance matrix $\sum i$ The mixture weights satisfy the constraint that $\sum_{t=1}^{M} P_i = 1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariancematrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$\lambda = \{P_i, \vec{\mu_1}, \sum i\}$ i=1,…..,M.

For Speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ

There are two principal motivations for using Gaussian mixture densities as a representation of speaker identity. The first motivation is the intuitive notion that the individual component densities of a multi-modal density, like the GMM, may model some underlying set of acoustic classes. It is reasonable to assume the acoustic space corresponding to a speaker's voice can be characterized by a set of acoustic classes representing some broad phonetic events, such as vowels, nasals or fricatives. These acoustic classes reflect some general speaker-dependent vocal tract configurations that are useful for charactering speaker identity.

The second motivation for using GMM for speaker identification is the empirical observation that a linear combination of Gaussian basis function is capable of representing a large class of sample distributions. A powerful attribute of a GMM is its ability to form smooth approximations to arbitrary-shaped densities.

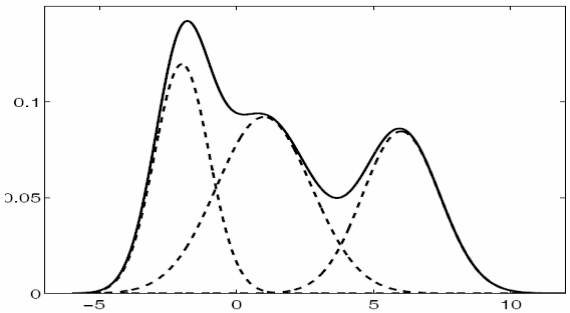

Fig.2. A GMM with three component densities

Given training speech from a speaker, the goal of speaker model training is to estimate the parameters of the GMM λ, which in some sense best matches the distribution ofthe training feature vectors. There are several techniques available for estimating the parameters of a GMM [5]. By far the most popular and well-established method is maximumlikelihood (ML) estimation. The estimation of ML parameter can be obtained iterativelyusing a special case of the expectation-maximization (EM) algorithm. 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 $X = (\vec{x_1}, \vec{x_2}, \ldots \vec{x_T})$ the GMM likelihood can be written as

$$(\vec{x_1}, \vec{x_1}, \ldots \vec{x_T}) \quad p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x_t}|\lambda) \qquad (6)$$

But, this expression is a non-linear function of the parameters and direct maximization is not possible. However, ML parameter estimates can be obtained iteratively using a special case of the expectation –maximization (EM) algorithm.On each EM iteration the following reestimation formulas are used which guarentee a monotonic increase in the model's likelihood value:

Mixture Weights: $$\vec{p_1} = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda) \qquad (7)$$

Means: $$\vec{\mu_1} = \frac{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)\vec{x_t}}{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)} \qquad (8)$$

Variances: $$\vec{\sigma_1}^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)} - \vec{\mu_1}^2 \qquad (9)$$

Where $\vec{\sigma_i}^2, x_t$ and μ_i refer to arbitary elements of the vectors $\vec{\sigma_i}^2$, $\vec{x_t}$,and $\vec{\mu_i}$ respectively .

The *aposteriori* probability for acoustic class i is given by

$$p(i|\vec{x_t}, \lambda) = \frac{p_i b_i(\vec{x_t})}{\sum_{t=1}^{T} p_k b_k(\vec{x_t})} \qquad (10)$$

The two critical factors in training a Gaussian Mixture Model are selecting the order M of the mixture and initializing the order parameters prior to the EM algorithm. Initialising the order parameters is done by using K-Means algorithm.This algorithm clusters the vectors based on attributes into k partitions.

For Speaker identification, a group of S speakers S={1,2,S} is represented by GMM's $\lambda_1, \lambda_2, \ldots, \lambda_S$ The objective is to find the speaker model which has the maximum posteriori probability for a following observation sequence $(\vec{x_1}, \vec{x_1}, \ldots \vec{x_T})$

6. Decision making process

In the test phase GMM models for different speakers are made models are loaded and input waveform from anyone among the speakers in the database is read. Now by using the maximum likelihood algorithm, scores for all speakers are computed and these models collectively form the database. In the training phase, these different speakers with the maximum score is identified as the one who has spoken. The test phase is as shown in fig 3.

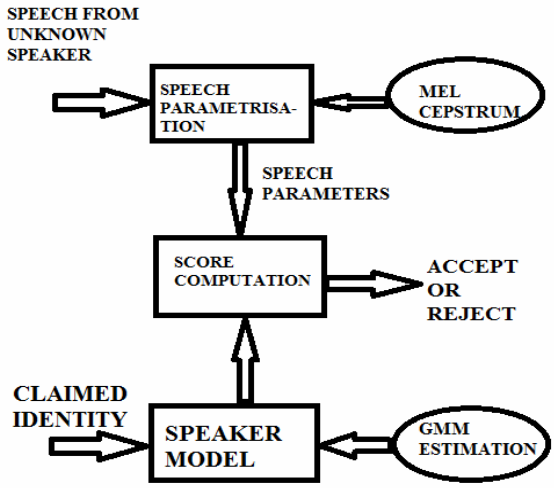

Fig.3. Test phase

7. Experiments and results

This section presents the experimental evaluation of Gaussian Mixture Model implemented using Matlab7 language environment for Text-Independent speaker Identification. To investigate speaker identification performance of the GMM four speakers were trained with 4,8 and 16 component Gaussians using 60 seconds speech of several languages such as English, Hindi, Tamil and Malayalam. The speech signals were divided into 25 ms frames. 13 Mel Frequency Cepstral coefficients were obtained for each frame. The following table shows the identification performance of the GMM.

Table 1. Identification performance of GMM

SPEECH(Sec)	NUMBER OF GAUSSIANS	TEST LENGTH		
		1Sec(in %)	5Sec(in%)	10Sec (in%)
30	4	48.3	74.8	79.4
	8	55.2	80.3	83.6
	16	64.2	88.4	91.2
60	4	56.7	84.6	88.2
	8	65.8	91.2	97.4
	16	75.2	95.8	98.8

8. Conclusion

This paper has evaluated the use of GMM, made using the feature vectors obtained by doing MFCC on the pre-processed speech signal,for Robust Speaker Identification.The Gaussian Mixture speaker Model attains high recognition rate for various speech durations.The Recognition rate is maximum(98.8 %) when the speech is of 60 seconds duration and the number of Gaussians is 16.

References

[1] S. Furui "An Overview of speaker recognition technology" In Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pages 1-9, Martigny, Switzerland, April 1994

[2] E.C. Gordon "Signal and Linear System Analysis". Copyright © 1998 John Wiley &Sons Ltd., New York, USA

[3] Reynolds, Douglas A. Thomas F. Quatieri, and Robert B.Dunn "Speaker Verification Using Adapted Gaussian Mixture Models" Digital Signal Processing. vol. 10, pp. 19-41, 2000.

[4] S.S. Stevens, J.Volkmann and E.B. Newman, 1937 "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of American,8, 185-190.

[5] S.S. Stevens and J.Volkmann, 1940 "The relation of pitch to frequency: A revised scale". The American Journal of Psychology,53(3),329-353.

[6] G.McLachlan " Gaussian Mixture Models" NewYork: Marcel Dekker, 1988

[7] Daniel Jurafsky & James H. Martin, 2007. "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition".

[8] Reynolds,Douglas A "Speaker Identification and Verification Using Gaussian Mixture Speaker Models" Speech Communication. vol. 17, pp. 91-108, 1995

[9] Reynolds, Douglas A. Robust "Text-Independent Speaker Identification Using Gaussian Mixture Speaker, Model" IEEE Transactions on Speech and Audio Processing. vol.3, n. 1, pp. 72-83, January, 1995