International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:10, 2014

An Intelligent Text Independent Speaker Identification Using VQ-GMM Model Based Multiple Classifier System

Cheima Ben Soltane, Ittansa Yonas Kelbesa

***Abstract*—**Speaker Identification (SI) is the task of establishing identity of an individual based on his/her voice characteristics. The SI task is typically achieved by two-stage signal processing: training and testing. The training process calculates speaker specific feature parameters from the speech and generates speaker models accordingly. In the testing phase, speech samples from unknown speakers are compared with the models and classified. Even though performance of speaker identification systems has improved due to recent advances in speech processing techniques, there is still need of improvement. In this paper, a Closed-Set Tex-Independent Speaker Identification System (CISI) based on a Multiple Classifier System (MCS) is proposed, using Mel Frequency Cepstrum Coefficient (MFCC) as feature extraction and suitable combination of vector quantization (VQ) and Gaussian Mixture Model (GMM) together with Expectation Maximization algorithm (EM) for speaker modeling. The use of Voice Activity Detector (VAD) with a hybrid approach based on Short Time Energy (STE) and Statistical Modeling of Background Noise in the pre-processing step of the feature extraction yields a better and more robust automatic speaker identification system. Also investigation of Linde-Buzo-Gray (LBG) clustering algorithm for initialization of GMM, for estimating the underlying parameters, in the EM step improved the convergence rate and systems performance. It also uses relative index as confidence measures in case of contradiction in identification process by GMM and VQ as well. Simulation results carried out on voxforge.org speech database using MATLAB highlight the efficacy of the proposed method compared to earlier work.

***Keywords*—**Feature Extraction, Speaker Modeling, Feature Matching, Mel Frequency Cepstrum Coefficient (MFCC), Gaussian mixture model (GMM), Vector Quantization (VQ), Linde-Buzo-Gray (LBG), Expectation Maximization (EM), pre-processing, Voice Activity Detection (VAD), Short Time Energy (STE), Background Noise Statistical Modeling, Closed-Set Tex-Independent Speaker Identification System (CISI).

I. Introduction

SPEECH signal is basically meant to carry the information about the linguistic message. But, it also contains the speaker specific information. It is generated by acoustically exciting the cavities of the mouth and nose, and can be used to recognize (identify/verify) a person. This paper deals with the speaker identification task (SI); i.e., to find the identity of a person using his/her speech from registered speaker voice stored in the database. SI can be text-dependent and text-independent [1]. Text-independent SI system is not limited to recognize speakers on the basis of same sentences stored in the database. While text-dependent SI system only can recognize speakers by uttering the same sentence every time [2]. SI can be further divided into closed set SI and open set SI [3]. In closed set speaker identification, unknown speech signal came from one of the registered speakers. Open-set speaker identify unknown signal from either the set of the registered speakers or unregistered speakers. The basic structure of Speaker Identification system (SIS) is shown in Fig. 1.

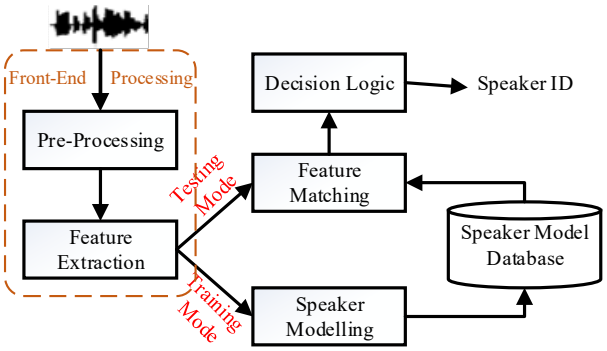


Fig. 1 Basic structure of Speaker Identification (SI)

In spite of impressive advances in the field of speaker identification in recent years, it is still the area of an active research because of uncertainties involved due to unknown environments in real world scenarios. These uncertainties are due to like mimicking voice, background noise, recording, stress condition of individuals, etc.

In the recent years commercial applications of speaker recognition systems have become a reality. It is starting to gain increasing acceptance in both government and financial sectors as a method to facilitate quick and secure authentication of individuals. Potential applications of speaker recognition include access security, phone banking, web services, personalization of services and customer relationship management. When combined with speech recognition, speaker recognition has the potential to offer most natural to human-computer means of communication.

Moreover, biometric applications of speaker recognition provide very attractive alternatives to Biometrics based on finger prints, retina scans and face recognition. The advantages of speaker recognition over these techniques include: low costs and non-invasive character of speech

Cheima Ben Soltane & Ittansa Yonas Kelbesa are with Department of Information Engineering, University of Brescia, Via Branze, 38 – 25123 Brescia, Italy (e-mail: y.ittansa@studenti.unibs.it, ch.bensoltane@gmail.com).

acquisition, no need for expensive equipment. As an access security tool, speaker recognition can potentially eliminate the need for remembering PIN numbers and passwords for bank accounts and security locks and various online services. Moreover, speaker identification and verification is the only biometric technique that can be viably used over the telephone without the user having dedicated hardware.

The two widely used feature extraction techniques include: Mel-frequency Cepstral Coefficients (MFCC) [4] and Linear Prediction Coefficients (LPC) [5]. To solve the above mentioned challenges for feature extraction, MFCC is found more accurate compared to LPC.

Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Dynamic Time warping (DTW) and Vector Quantization (VQ) [6]-[8] are prevalent techniques for features matching in SIS.

In this paper, a model that combines two modeling methods is proposed i.e. VQ and GMM. This is due to the fact that speaker identification made by a single decision making scheme is always a risky because each type of features are not suitable for all environments. Thus, this paper describes a Multiple Classifier System (MCS) for CISI which reduces errors and wrong identification. The basic idea is to analyze the results obtained by different classifiers. Then, these classifiers are integrated such that their reliability is enhanced due to a proper combination technique. In the proposed approach, overall decision is based on agreement or disagreement by individual models. In case of agreement, speaker identification is simple but in case of difference of opinion, confidence ratio have been used - ratio of best score to the second best score - as a secondary measure that shows the confidence of the given model for particular identification task. This overcomes the disadvantage of individual VQ and GMM methods. In the proposed system, parameters for feature extraction like filter type and size, number of MFCC and for modeling like number of Gaussians and codebook size are fine tuned by performing experimentation.

The rest of paper is organized as follows. Section II discusses components of the SIS. Section III explains the proposed Multiple Classifier System (MCS). Section IV depicts the results obtained from systems testing and experimentations. Finally we conclude our work in Section V.

II. Components of a Speaker Identification System

A. Front-End Processing

The aim of the front-end processing is to extract the speaker discriminative features. The speech signal needs to undergo various signal conditioning steps before being subjected to the feature extraction methods, as depicted in Fig. 1, in order to generate the feature vectors.

1. Pre-Processing

Before extracting the features of the signal various pre-processing tasks must be performed. This task includes Analogue-to-Digital conversion (A/D), Noise Removal, Pre-emphasis, silence removal etc. Fig. 2 shows the pre-processing steps.

a) Pre-Emphasis

Due to the characters of the human vocal system, glottal airflow and lip radiations make the higher frequency components of the voiced sounds dampened. To eliminate this effect and prevent lower frequency components from dominating the signal, pre-emphasis should be performed. Fig. 3 shows effect of pre-emphasizing.

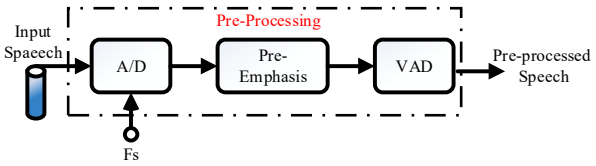
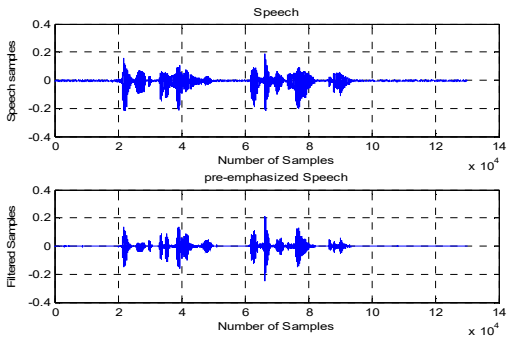


Fig. 2 Speech Pre-Processing subsystem

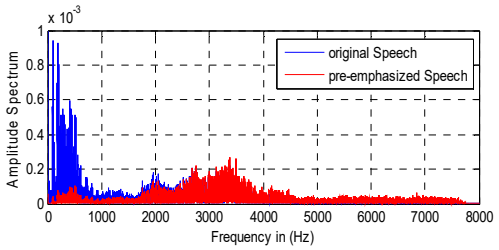
By pre-emphasizing, dynamic range will be decreased so as to let spectral modeling methods capture details at all frequency components equally. Generally pre-emphasis is performed by filtering the speech signal (original signal) with the first order FIR filter, which has the form as follow:

$$y(z) = 1 - az^{-1} \quad (0 < a < 1) \tag{1}$$

where 'a' is the pre-emphasis factor.



(a)



(b)

Fig. 3 (a) Original and pre-emphasized speech, (b) Amplitude spectral plots for the original and pre-Emphasized speech

b) Voice Activity Detection (VAD)

VAD is the fundamental step for applications like Speaker Identification. There are three main activities/events in speech,

i.e., Silence (S), Unvoiced (U) and Voiced (V). The information which is more important from the prospective of speaker identification is generally contained inside the voiced part of the speech signal. Therefore the process of isolating the redundant information especially in the unvoiced part in the pre-processing step bears a lot of importance. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part. Below the proposed speech silence removal by hybrid algorithm will be discussed.

The hybrid algorithm first classifies the input signal using short time energy [11], [12] into smaller parts; each part is either voiced or unvoiced. The transitional frames between voiced and unvoiced part are then classified properly according to their statistical behavior [13]. Silence removal using this algorithm has nine steps:

1. Classify the input signal into smaller segments by using STE algorithm. Each segment is either voiced or unvoiced.
2. Calculate the mean (μ) and variance (σ) of each segment.
3. Select the transitional frames by taking 3 frames on either side of the transition region between voiced and unvoiced segment.
4. Calculate the mean (μ_t) and variance (σ_t) of all transitional frames.
5. The segment present on the left side of transitional frames is named as left segment with its mean μ_l and variance σ_l and the segment present on the right side of transitional frames is named as right segment with its mean μ_r and variance σ_r.
6. Calculate the Bhattacharyya distance between left segment and transitional frame, which is denoted by $dist_{left}$. Analytically,

$$dist_{left} = \frac{1}{2}\ln\frac{\sigma_l+\sigma_t}{2(\sigma_l\sigma_t)^{1/2}} + \frac{1}{8}(\mu_l-\mu_t)^2\left(\frac{\sigma_l+\sigma_t}{2}\right)^{-1} \quad (2)$$

7. Calculate the Bhattacharyya distance between right segment and transitional frame, which is denoted by $dist_{right}$. Analytically,

$$dist_{right} = \frac{1}{2}\ln\frac{\sigma_r+\sigma_t}{2(\sigma_r\sigma_t)^{1/2}} + \frac{1}{8}(\mu_r-\mu_t)^2\left(\frac{\sigma_r+\sigma_t}{2}\right)^{-1} \quad (3)$$

8. Classify the transitional frame on the basis of following rule:

if ($dist_{left} < dist_{right}$) transitional frame belongs to the left segment
else it belongs to the right segment.

9. Eliminate all unvoiced frames from the input speech signal.

Fig. 4 shows the speech silence removal by proposed method.

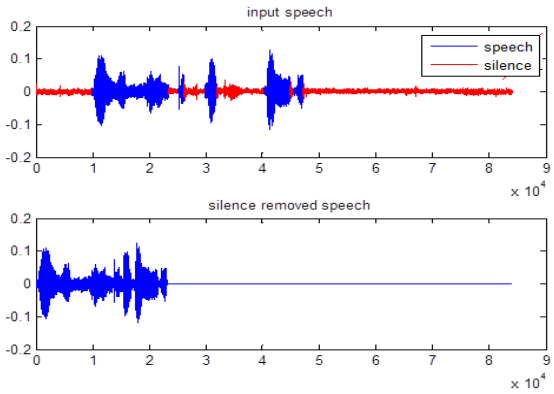


Fig. 4 Silence removal by hybrid algorithm

2. Feature Extraction Using Mel Frequency Cepstrum Coefficient (MFCC)

To recognize the speaker, extraction of the features from speaker's speech is required. MFCCs [4] are commonly used feature vectors for speaker identification. The computation of MFCC is shown in Fig. 4.

a) Frame Blocking

Speech signal is quasi-periodic in voiced segment, and it can be viewed as short-time stationary within 10 - 30 ms, see Fig. 5. Hence, the pre-emphasized speech signal should be framed in to short overlapping segment before further processing. The usefulness of overlapping is given by the fact that a single frame in uniform segmentation can contain a non stationary transition between two frames. This effect can be reduced by using overlapping, as the probability that a window is centered on the middle. Fig. 5 illustrates the blocking into frames.

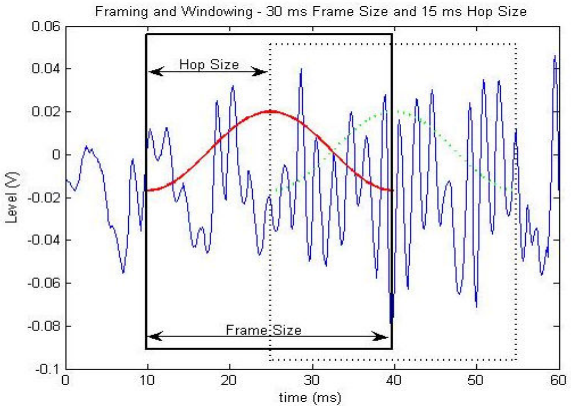


Fig. 5 Speech frame blocking and windowing

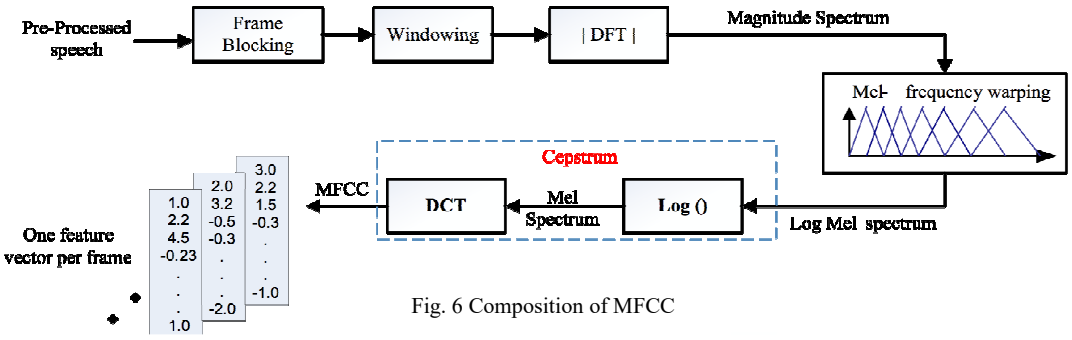


Fig. 6 Composition of MFCC

b) Windowing

The framed signal is multiplied by a window function as shown in Fig. 5. The window function is used to smooth the signal for the computation of the DFT. The DFT computation makes an assumption that the input signal repeats over and over. If there is a discontinuity between the first point and the last point of the signal, artifacts occur in the DFT spectrum.

By multiplying a window function to smoothly attenuate both ends of the signal towards zero, this unwanted artifacts can be avoided. The window function that is applied is preferably not rectangular, as this can lead to distortion due to vertical frame boundaries. The hamming window is usually used in speech signal spectral analysis, because its spectrum falls off rather quickly so the resulting frequency resolution is better, which is suitable for detecting formants.

$$w(n) = 0.54 - 0.4\cos\left(\frac{2\pi n}{N-1}\right) \qquad n = 0,1,...,N-1 \tag{4}$$

c) Discrete Fourier Transform (DFT)

After segmenting the speech signal into overlapping frames and windowing, as depicted in Fig. 7, the frequency response of each frame is computed by Discrete Fourier Transform (DFT). Then the spectrogram of the speech signal is obtained.

$$X[k] = \sum x[n] e^{-j2\pi nk/N}, \qquad k \in Z \tag{5}$$

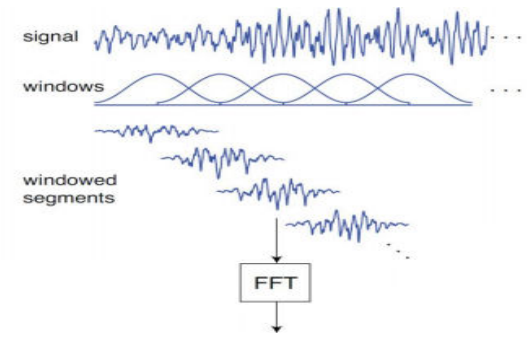


Fig. 7 DFT computation on framed and windowed speech segments

d) Mel-Frequency Warping

Mel (melody) is a unit of special measure or scale of perceived pitch of a tone. The relation between linear frequency (f) and mel frequency $mel(f)$ can be approximated:

$$mel(f) = \begin{cases} 2595\log_{10}(1 + f/\ 700) & f > 1KHz \\ f & f < 1KHz \end{cases} \tag{6}$$

Mel-frequency scale, based on human auditory perception experiments, is approximately linear up to the frequency of 1000 Hz and then becomes close to logarithmic for the higher frequencies. It is observed that human ear acts as filters that concentrate on only certain frequency components. Thus the human auditory system can be modelled by a set of band-pass filters. Since the relationship between frequency scale and Mel-frequency scale is nonlinear, these filters are non-uniformly spaced on the frequency scale, with more filters in the low frequency regions and less filters in the high frequency regions. Mel Filter Bank filters an input magnitude spectrum through a bank of number of Mel-filters. The output is an array of filtered values, typically called Mel spectrum, each corresponding to the result of filtering the input spectrum through an individual filter. It can be achieved by:

$$Y[n] = \sum_{k=0}^{N/\ 2} |X[k]|^2 \times MelWeight[n][k] \qquad 0 < n < M \tag{7}$$

where M is the number of filters.

Fig. 8 shows a 24-band Mel-frequency filter bank.

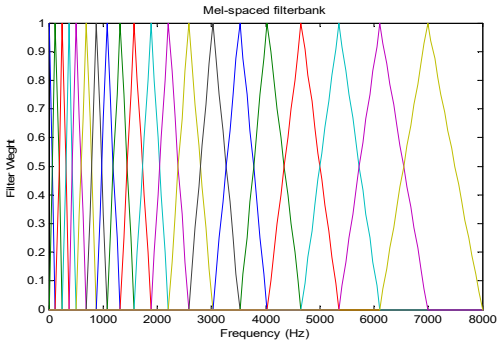


Fig. 8 A 24-band Mel-frequency filter bank

e) Cepstrum

The goal is to obtain the spectral envelope, because it conveys information about the formants. Cepstrum analysis can be used to extract the spectral envelope from the spectrum. The MFCC features are obtained by taking log of the outputs of a Mel- frequency filter bank. And conduct the Discrete Cosine Transform (DCT) rather than IFFT as in the case for computing the Cepstral coefficients to convert the log Mel spectrum back to time. The result is called the MFCC which can be calculated as:

$$C_k = \alpha_k \sum_{n=0}^{N-1} \log\left(Y[n]\right)\cos\left[k\left(n-\frac{1}{2}\right)\frac{\pi}{N}\right] \qquad \forall k = 0,\ldots,\ M-1$$

$$\alpha_k = \begin{cases} \sqrt{1/N} & k = 0 \\ \sqrt{2/N} & k \neq 0 \end{cases} \qquad (8)$$

Note that the first component, C_0, is excluded from the DCT since it represents the mean value of the input signal, which carries little speaker specific information. Fig. 9 shows 19 dimensional MFCCs projection into two principal components, to observe the separation between two speakers.

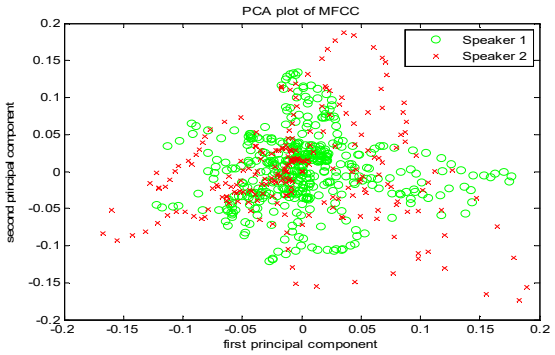


Fig. 9 MFCCs projection into two principal components

B.Speaker Modeling

Speaker modeling algorithms have been used for speaker identification by compressing feature vectors but retaining most prominent characteristics. Generally there are two major types of models for classification: stochastic models (includes: GMM) and template models (includes: VQ)

1. Vector Quantization (VQ)

It is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution. A vector quantizer maps k-dimensional vectors, $X=\{x_1, x_2, \ldots, x_T\}$, in the vector space R^k into a smaller finite set of vectors, by clustering it in to a set of M vectors, $C=\{c_1, c_2, \ldots, c_M\}$, called a codebook in the vector space R^k, which are representative feature vectors as an efficient means of characterizing the speaker specific features. Speaker recognition can be done using the code book generated for each registered user. One speaker can be discriminated from another based on the location of centroids. Fig. 10 best describes the process.

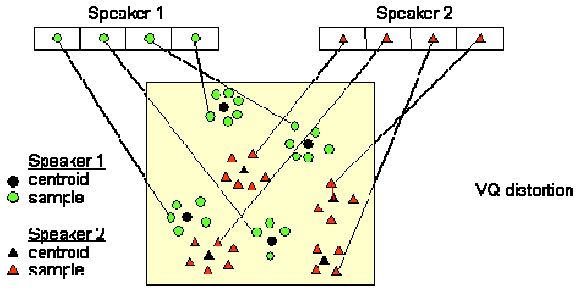


Fig. 10 Conceptual diagram illustrating VQ codebook formations

The clustering is done by a clustering algorithm. Comparison of different clustering techniques is provided in [9]. The binary split algorithm proposed by Linde, Buzo and Gray (LBG) is most frequently used VQ technique [10]. The LBG algorithm is shown in Fig. 11:

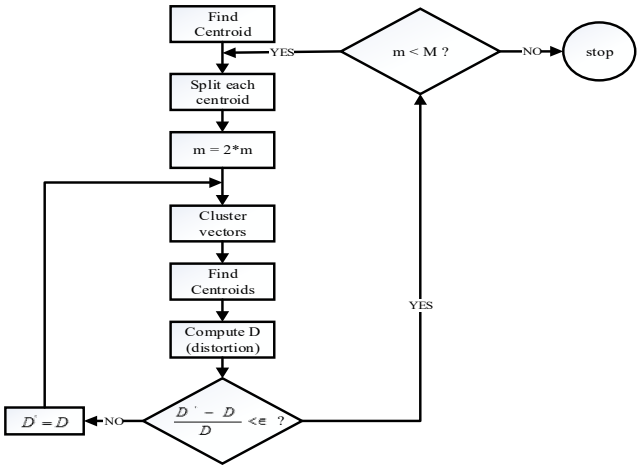


Fig. 11 Flow diagram of the LBG algorithm

Hence for a given number of users N, codebooks are generated for each speaker during the training phase using VQ method to build a speaker-database, $C_{database} = \{C_1, C_2, \ldots, C_N\}$ consisting of N codebooks, one for each speaker in the database.

2. Gaussian Mixture Model (GMM)

GMM is a density estimator and is one of the most commonly used types of classifier [6]. When feature vectors are displayed in D-dimensional feature space after clustering, they some-how resemble Gaussian distribution as shown in Fig. 12.

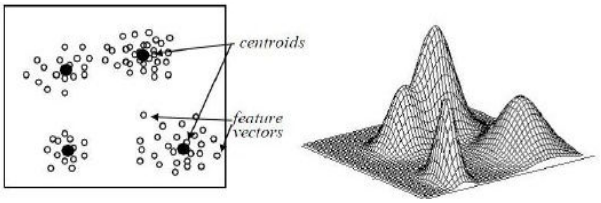


Fig. 12 GMM models showing a feature space and corresponding Gaussian model in 2D

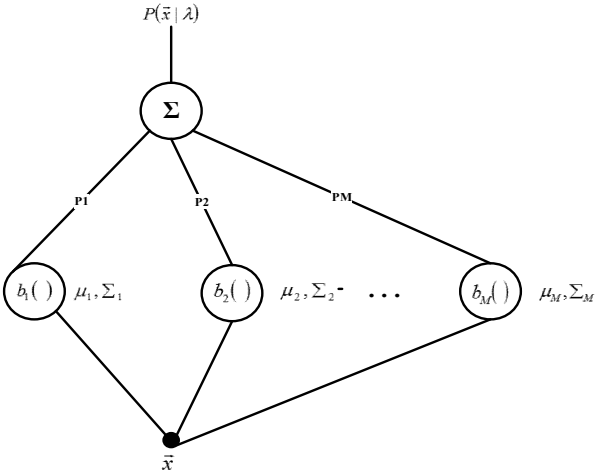


Fig. 13 Description of M-component Gaussian densities

In this method, the distribution of the feature vector x is modeled clearly using a mixture of M Gaussians. A Gaussian Mixture density is a weighted sum of M component densities, and is given by:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{9}$$

where $\vec{x}$ refers to a feature vector, p_i stands for mixture weight of i[th] component and $b_i(\vec{x})$ is the probability distribution of the i[th] component in the feature space. Diagrammatically it is shown in Fig. 13.

As the feature space is D-dimensional, the probability density function $b_i(\vec{x})$ is a D-variate distribution. It is given by:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{\left\{\frac{-1}{2}(\vec{x}-\vec{\mu}_i)' \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\}} \tag{10}$$

where μ_i is the mean of i[th] component and Σ_i is the co-variance matrix.

The complete Gaussian mixture density is represented by *mixture weights* p_i, *mean* μ_i and *covariance* Σ_i of corresponding component and denoted as:-

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \qquad i = 1,..., M \tag{11}$$

For Speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ. GMM uses the Expectation Maximization (EM) algorithm to determine the underlying parameters, i.e., the means, covariances and mixing coefficients. The EM algorithm for Gaussian Mixtures works as follows:

1. Initialize Gaussian parameters: means μ_i, covariances Σ_i and mixing coefficients p_i for each cluster i.

$$\mu_i \leftarrow mean(cluster(i)) \qquad \Sigma_i \leftarrow \text{cov}(cluster(i))$$

$$p_i \leftarrow \frac{\#\, po\,\text{int}\, s \quad in \quad i}{total \quad \#\, po\,\text{int}\, s}$$

2. **E Step:** Assign each point $\vec{x}_t$ an assignment score (posteriori probability) $\gamma(z_{ti})$ for each cluster i.

$$\gamma(z_{ti}) = \frac{p_i b(x_t \mid \mu_i, \Sigma_i)}{\sum_{j=1}^{M} p_j b(x_t \mid \mu_j, \Sigma_j)} \tag{12}$$

$\gamma(z_{ti})$ is called a "*responsibility":* how much is this Gaussian i responsible for this point $\vec{x}_t$

3. **M Step:** Given scores, adjust μ_i, p_i and Σ_i for each cluster i.

Mean of Gaussian i:

$$\mu_i^{new} = \frac{1}{N_i}\sum_{t=1}^{N} \gamma(z_{ti}) x_t \quad \text{Where: } N_i = \sum_{t=1}^{N} \gamma(z_{ti}) \tag{13}$$

Covariance of Gaussian i:

$$\Sigma_i^{new} = \frac{1}{N_i}\sum_{t=1}^{N} \gamma(z_{ti}) \left(x_t - \mu_i^{new}\right)\left(x_i - \mu_i^{new}\right)^T \tag{14}$$

Mixing coefficient of Gaussian i:

$$p_i^{new} = \frac{N_i}{N} \qquad \text{N: total number of points} \tag{15}$$

4. **Evaluate likelihood.** If likelihood or parameters converge stop.

$$\ln P(X \mid \mu, \Sigma, p) = \sum_{t=1}^{N} \ln\left\{\sum_{i=1}^{M} p_i b(x_t \mid \mu_i, \Sigma_i)\right\} \tag{16}$$

EM algorithm for Gaussian Mixtures procedure must be initialized with some starting point $\lambda^{(0)}$, preliminary clustering of the feature vectors. The EM algorithm is guaranteed to find a local maximum likelihood model regardless of the starting point, but the likelihood equation for a GMM has several local

maxima and different starting models can lead to different local maxima. K-Means and K-Means++ are some of the initialization methods employed. Unfortunately none of them are satisfactory and it took them lots of iterations to converge. In the contrary, VQ using LBG method of initialization has shown a better performance compared to the above mentioned techniques. Fig. 6 gives a better understanding. In this paper, VQ_{LBG} are primarily used for GMM initialization since they showed a better identification performance.

C. Feature Matching

a) Speaker Identification Employing VQ

During the recognition (testing) phase, after generating codebooks for each user during the training phase consisting of N codebooks, $C_{database} = \{C_1, C_2, \ldots, C_N\}$ one for each speaker in the database, the feature vectors $\{y_1, y_2, \ldots, y_T\}$ representing the test utterance are encoded in terms of their nearest code vectors from the codebook of each of the N speakers. The total distortion for the i^{th} speaker is computed by:

$$D^i = \sum_{t=1}^{T} \min_{1 \le j \le M} d\left(y_t, C_j^i\right) \qquad i = 1, \ldots, N \tag{17}$$

where C_j^i is the j^{th} code vector of the i^{th} speaker's codebook.

Once these N distances are computed, the speaker identification system classifies the test utterance to a speaker whose VQ codebook results in the least distortion; i.e.,

$$i^* = \arg\min_{1 \le j \le M} D^i \tag{18}$$

b) Speaker Identification Employing GMM

After modeling each user's Gaussian mixture density, we will have a set of models, each representing Gaussian distribution of all the components present. For K number of speakers it is denoted as $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_k\}$. The objective culminates in finding the speaker model λ having maximum posterior probability for a given test utterance [14]. Mathematically it can be represented as:

$$\hat{S} = \arg\max_{1 \le k \le S} \Pr(\lambda_k \mid X) = \arg\max_{1 \le k \le S} \frac{\Pr(X \mid \lambda_k)\Pr(\lambda_k)}{p(X)} \tag{19}$$

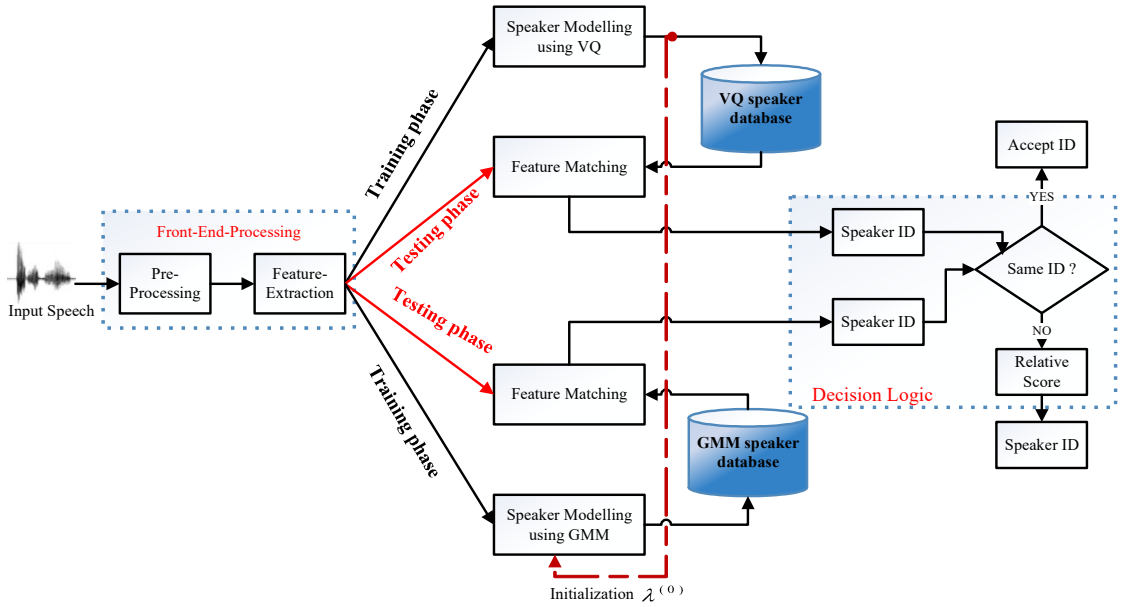


Fig. 14 Proposed Hybrid method using GMM and VQ

III. Proposed Multiple Classifier System (MCS)

The proposed MCS is shown in Fig. 14. Once extracted features are modeled using both VQ and GMM blocks, they are compared with database of known speaker to identify the user. If both models agree, the system directly accepts speaker. Otherwise, in order to take the final decision relative scores for both methods are computed as under:

$$relative\_index = \left| \frac{best\_score - second\_best\_score}{best\_score} \right| \times 100\% \tag{20}$$

Then the correct Speaker ID will be the one generated by the classifier which resulted in greater relative score.

IV. Simulation Results

A. Data Description

The Automatic Speaker Identification System (ASI) presented in this paper has been tested using *voxforge.org* speech database, using 80 speakers for both training and testing, for evaluating the effectiveness of the speaker identification system. The database consists of plenty of speakers, both female and male, that contains ten speech

utterances for each speaker in different languages. During the experiments the files were concatenated to produce one 12s utterance which contains seven sentences for each speaker. The remaining three files were coupled and used as tests segment. All experiments used 12s of English language speech with a sampling rate of 16 KHz to train the system, whereas the testing sessions was done using three different test shot lengths.

V. Simulation Results

A. Data Description

The Automatic Speaker Identification System (ASI) presented in this paper has been tested using *voxforge.org* speech database, using 80 speakers for both training and testing, for evaluating the effectiveness of the speaker identification system. The database consists of plenty of speakers, both female and male, that contains ten speech utterances for each speaker in different languages. During the experiments the files were concatenated to produce one 12s utterance which contains seven sentences for each speaker.

The remaining three files were coupled and used as tests segment. All experiments used 12s of English language speech with a sampling rate of 16 KHz to train the system, whereas the testing sessions was done using three different test shot lengths.

B. Performance of the Speaker Identification System

Since this is a speaker identification system and it is ultimately concerned with its ability to identify speakers, the performance of the system is measured using the identification rate. The identification rate can be described as:

$$Identification\_rate(\%) = \frac{\#correctly\_identified\_segments}{total\_\#of\_segments} \times 100\% \quad (21)$$

In the final modeling, identification rate affecting parameters like codebook size (number of Gaussians), number of MFC coefficient and number of Filters in filter bank are fine-tuned after performing several experiments.

C. The Effect of Codebook and Mixture Component Size on Identification Rate

TABLE I
Codebook Size vs. Identification Rate

Number of centroids (C)	*VQ Identification rate (%)*
16	77.5
32	86.25
64	93
128	93.75
256	95

TABLE II
# of Mixture Components vs. Identification Rate

Number of Gaussian Mixture components	GMM Identification rate (%)
2	86.25
4	93.75
8	97
16	97.5
32	92.5

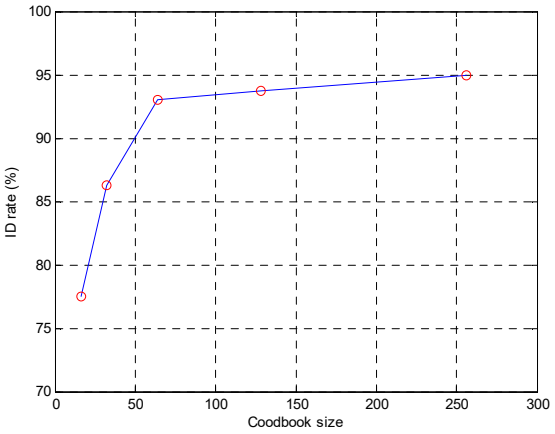


Fig. 15 Codebook size Vs. Identification rate With 19 MFCC 29 filter-banks and 16 Gaussian Mixture components

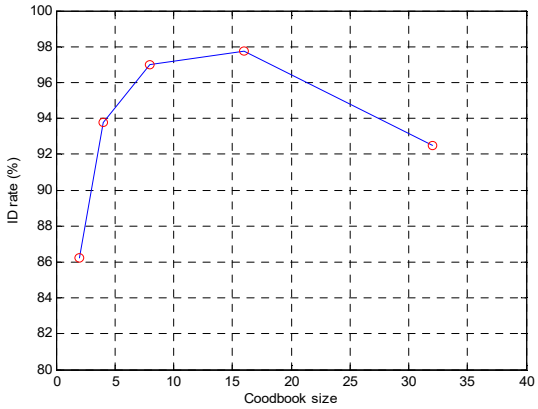


Fig 16 Number of Gaussian Mixture components Vs. Identification rate With 19 MFCC, 29 filter-banks and 128 codebook size (# centroids)

From the table shown above, it is obvious that increasing the number of centroids (codebook size) results in increasing the identification rate, but the computational time will also increase. In the case of GMM, there is a sharp increase in identification performance from 1 to 8 mixture components and leveling off above 16 components. This indicates there is a lower limit on the number of mixture components necessary to adequately model the speakers. Models must contain at least this minimum number of increasing components to maintain a good speaker identification performance.

D. Number of the MFCC Coefficients

TABLE III
Identification Rate as a Function of the Number of the MFC Coefficients

No. of MFC coefficients	*Identification Rate (%)*		
	VQ	GMM	Proposed (GMM + VQ)
5	60%	65%	65%
12	92.5%	92.5%	95%
19	93.75%	95.75%	97%
24	94%	97.5%	98.5%

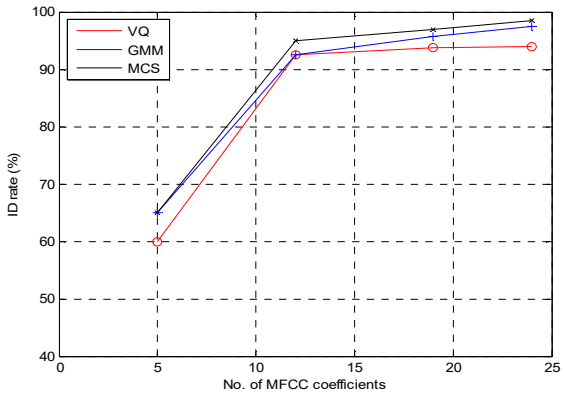


Fig. 17 Identification rate as a function of the number of the MFC coefficients. Using 29 filter, a codebook size of 128 and 16 Gaussian mixture components

Increasing the number of mel frequency cepstral coefficients (MFCC) results in improving the identification rate on the expense of the computational time. MFCCs are typically in the range (12-20).

E. The Number of Filter-Banks

TABLE IV
IDENTIFICATION RATE FOR DIFFERENT VALUES OF THE FILTER

Number of Filter-banks	***Identification Rate (%)***		
	VQ	GMM	Proposed (GMM + VQ)
20	90.75%	93%	93.75%
24	91.25%	93.75%	95%
29	93.75%	95.75%	97%
36	93.75%	96.25%	97.5%

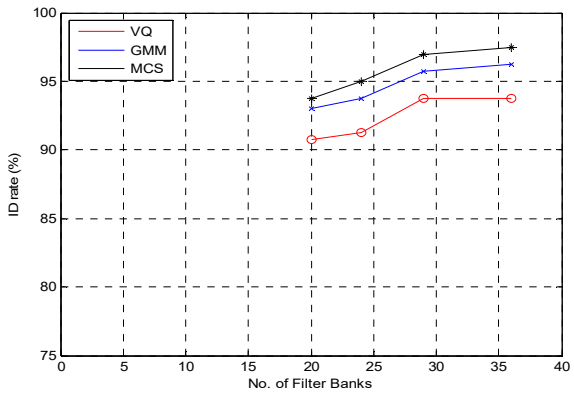


Fig. 18 Performance evaluation as a function of number of the filter-banks. With 19 MFCC, a codebook size of 128 and 16 Gaussian mixture components

It is obvious that number of the filter-banks plays a major role for the purpose of improving the identification accuracy. From the figure, there is a leveling off above 29 filter banks. Thus it is possible to obtain a very good identification rate using the 29 filter-banks.

F. The Performance of the System on Different Test Shot Lengths

To study the performance of different test shot lengths, three tests were conducted using all test speakers uttering the same test speech sample with three different lengths. And the results were as the following:

TABLE V
IDENTIFICATION RATE FOR DIFFERENT TEST SHOT LENGTHS

Test speech length	***Identification Rate (%)***		
	VQ	GMM	Proposed (GMM + VQ)
1sec	57.5 %	61.25 %	67.5 %
2.5sec	78.75%	86.25%	88.75 %
4sec , Full test shots	93.75%	95.75%	97%

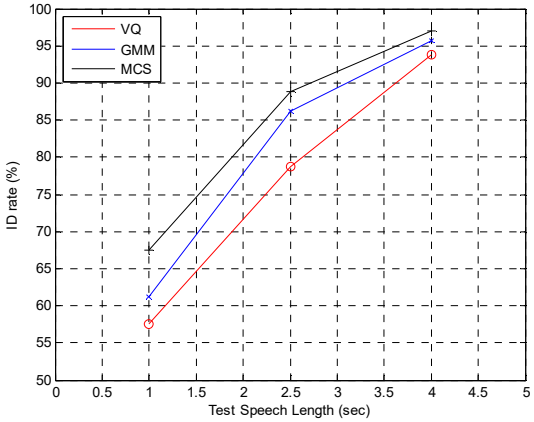


Fig. 19 Identification rate Vs test speech sample length Using 29 filter, 19 MFCC, a codebook size of 128 and 16 Gaussian mixture components

G.Effect of VAD on the Identification Rate

Removal of silence/unvoiced portion of a speech is the fundamental step for Speaker Identification, since most important information is contained inside the voiced part. This is evident from the experiment as shown in Fig. 20 below. In the case of our data, using a database of 80 speakers, following results were obtained.

For 12 (sec) training shot and 4 (sec) tests shot length with 29 Filter Banks, 19 MFCC, 128 Codebook size and 16 Gaussian Mixture Components, the proposed Multiple Classifier System's (MCS) identification rate was 97%.

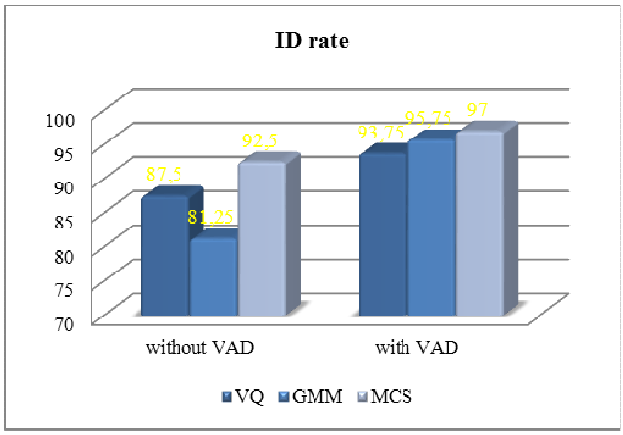


Fig. 20 Effect of VAD on the systems Identification rate

VI. Conclusion and Possible Future Developments

A. Conclusion

The study reveals that as the number of centroids increases, the identification rate of the system increases. Also, the number of centroids has to be increased as the number of speaker's increases. It is also observed that as the number of filters in the filter-bank increases, the identification rate increases. The experiments conducted using voxforge database, showed that reducing the test shot lengths reduced the identification accuracy. In order to obtain satisfactory result for real time application, the test data usually needs to be more seconds long. All in all, during this work, hybrid modeling methods of VQ and GMM is found to be an efficient and simple way to do speaker identification. The system is 97% accurate in identifying the correct speaker when using 12 seconds for training session and 4 seconds long for testing session.

B. Possible Future Developments

The current methods of feature extraction, even though they appear to function reasonably well, are inadequate in representing a speech. The cepstrtal coefficients are extremely good at representing vocal tract properties. However, they are unable to represent voicing information.

Thus it can be concluded that though cepstral coefficients as features are reasonably suited for speaker identification, their performance can be improved through the addition of voicing information.

References

[1] Furui S, "Recent advances in speaker recognition",Pattern Recognition Letters, vol. 18, no. 9, (1997). September, pp. 859–872.H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.

[2] K. Chen, L. Wang, and H. Chi., "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification". Journal on Pattern Recognition and Artificial Intelligence, 11(3):417–445, 1997.

[3] Reynolds, D.A., "An overview of automatic speaker recognition technology". Proc. IEEE Acoustics Speech Signal Processing 4,4072–4075 (2002).

[4] Godino-Llorente, J.I., Gómez-Vilda, P., Sáenz Lechón, N., Velasco, M.B., Cruz Roldán, F., Ballester, M.A.F., "Discriminative Methods for the Detection of Voice Disorder". In: A ISCA Tutorial and Research Workshop on Non-Linear Speech Processing, The COST- 277 Workshop (2005).

[5] Xugang, L., Jianwu, D., "An investigation of Dependencies between Frequency components and speaker characteristics for text–independent speaker identification". Speech Communication 2007 50(4), 312–322 (2007).

[6] D. A. Reynolds and R. C. Rose, "Robust text independent speaker identification using Gaussian mixture speaker models". IEEE Trans. on Speech and audio processing, vol. 3(1), pp. 72–83, 1995.

[7] Yuk,C.C.Q.L.D.-S., "An HMM approach to text independent speaker verification",. In IEEE international conference on Acoustics, Speech and signal processing, 1996.

[8] F. K. Soong, et. al., "A vector quantization approach to speaker recognition", AT & T Technical Journal, Vol.66, No.2, pp. 14-26, 1987.

[9] T. Kinnunen, T., Kilpeläinen,T., Fränti P.: "Comparison of clustering algorithms in speaker identification", proc. Lasted Int. Conf. Signal Processing and Communications (SPC): 222-227, Marbella, Spain, 2000.

[10] Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantizer Design,"IEEE Transactions on Communications, vol. COM-28, pp. 84-95, January 1980.

[11] Atal, B.; Rabiner, L., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", Acoustics,Speech, and Signal Processing (see also IEEE Transactions on Signal Processing), IEEE Transactions on, Volume: 24 , Issue: 3 , Jun 1976, Pages: 201 - 212.

[12] D. G. Childers, M. Hand, J. M. Larar, " Silent and Voiced/Unvoied/ Mixed Excitation(Four-Way), Classification of Speech", IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.

[13] G. Saha, Sandipan Chakroborty, Suman Senapat , "A New Silence Removal and end point detection algorithm for speech and Speaker Recognition Applications", Proceedings of the NCC 2005, Jan.

[14] 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.