17

Application of the Vector Quantization Methods and the Fused MFCC-IMFCC Features in the GMM based Speaker Recognition

Sheeraz Memon, Margaret Lech, Namunu Maddage and Ling He
School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia

1. Introduction

Speaker recognition system which identifies or verifies a speaker based on a person's voice is employed as biometric of high confidence. Over three decades of research, voice prints have established very important security applications for the authentication and recognition from voice channels. Recent years, speaker recognition community is putting more efforts to further improve main factors such as robustness and the accuracy in the context independent speaker recognition systems. Signal segmentation where the temporal properties such as energy and pitch with in the speech signal frame is ideally considered stationary, is a major step in speaker recognition systems. Another important area where robustness can be acheived is identifying speaker characteristic sensitive feature extraction methods. However the segmentation and feature extraction stages are examined by modelling methods, thus speaker characteristic modelling is also an important state which should be carefully designed. Effective improvements in above key steps subsequently improve the robustness and accuracy of the speaker recognition system.

In this book chapter we evaluate the performances of the speaker recognition systems when different feature settings and modelling techniques are applied for above mentioned step 2 and step 3 respectively. In general content sensitive features play a vital role in achieving the globally optimized classification decisions. State of the art speaker recognition systems extract acoustic features which capture the characteristics of the speech production system such as pitch or energy contours, glottal waveforms, or formant amplitude and frequency modulation and model them with statistical learning techniques. However *Mel frequency cepstral coefficients* (MFCCs) have commonly being used to characterize the speaker characteristics. In this chapter we compare effectiveness of Inverted MFCC and fused MFCC-IMFCC features against solo MFCC feature for speaker recognition systems. It is commonly assumed that the speaker characteristic distribution is Gaussian. Thus Gaussian Mixture model is effectively used for speaker characteristics modelling in the literature. In this chapter we examine different learning techniques for the representation of the parameters in the GMM based speaker models. Vector Quantization (VQ) techniques effectively cluster the information distributions and reduce the effects of noise. Its found VQ techniques improve the robustness of speaker recognition systems which are deployed at

different noisy environments. We propose several VQ methods to optimize GMM parameters (mean, covariance, and mixture weight). However expectation maximization (EM) algorithm is commonly used in the literature for the GMM parameter optimization. Thus we compare the performances of VQ based GMM –speaker modelling algorithms, K-means, LBG (Linde Buzo and Gray) and Information theoretic vector quantization (ITVQ) with EM-GMM setup in the speaker recognition.
The study includes speaker verification tests performed on the NIST2004 Speaker Recognition evaluation Corpus. NIST2004 SRE consists of conversational telephone speech. Thus performance evaluation of proposed methods using this corpus allows us to analyse and validate the results with high confidence. The results are presented using detection error trade-off (DET) plots showing the miss probability against the false alarm probability; a number of tables are also presented to compare the recognition rates based on different combination of these techniques.

2. Speaker Recognition

Speaker Recognition is a biometric based identity process where a person's identity is verified by the voice of a person. Biometrics based verification has received much attention in the recent times as such characteristics come natural to each individual and they are not required to be memorised, like passwords and personal identification numbers.
The speaker recogniton can be further classified in speaker identification and verification. Identification deals when a person is needed to verify from a group of people, however in verification task a person is accepted or rejected based on a claimant's identity.
In text-independent speaker verification the speaker is not bound to say a specific phrase to be identified but he/she is free to utter any sentence. However when we are dealing with text-dependent speaker recogntion the person is bound to utter a pre-defined phrase.
The speaker verification system comprises of three stages (see Fig. 1), in the first stage pre-processing and feature extraction is performed over a database of speakers. The second step addresses establishment of speaker models; where vectors representing speakers distinguishing characteristics are generated this corresponds to finding the distributions of feature vectors. The third step is of decision, which confirms or rejects the claimed identity of a speaker. In this stage the test set is also performed which includes the pre-processing and feature extraction of the test speaker and inputs to the classifier.
The introduction of the adapted Gaussian mixture models (Reynolds et al.,2000) with the introduction of UBM-GMM with MAP adaptation has established very good results on NIST evaluations. The use of expectation maximization (EM) optimization procedure is widely adapted to obtain the iterative updates for gaussian distributions. However EM encounters a number of problems, such as local convergence, mean adaptations etc. A number of EM variants are also proposed recently (Ueda, N. & R. Nakano, 1998), (Hedelin, P. Skoglund, J., 2000), (Ververidis, D. Kotropoulos, C.,2008) and (Ethem A.,1998).

3. Vector Quantization and EM based GMM

The study in (Hedelin, P. Skoglund, J., 2000) proposes how vector quantization based on GMM enhances the performance. A number of statistical tests are conducted in (Ververidis, D. Kotropoulos, C.,2008), it suggests around seven EM variants which under enhanced

methods improve the GMM performance. The relation between number of vector quantization methods and EM is established in (Ethem A.,1998). To overcome the problem of local maxima caused by EM algorithm with an annealing approach is suggested in (Ueda, N. & R. Nakano, 1998).

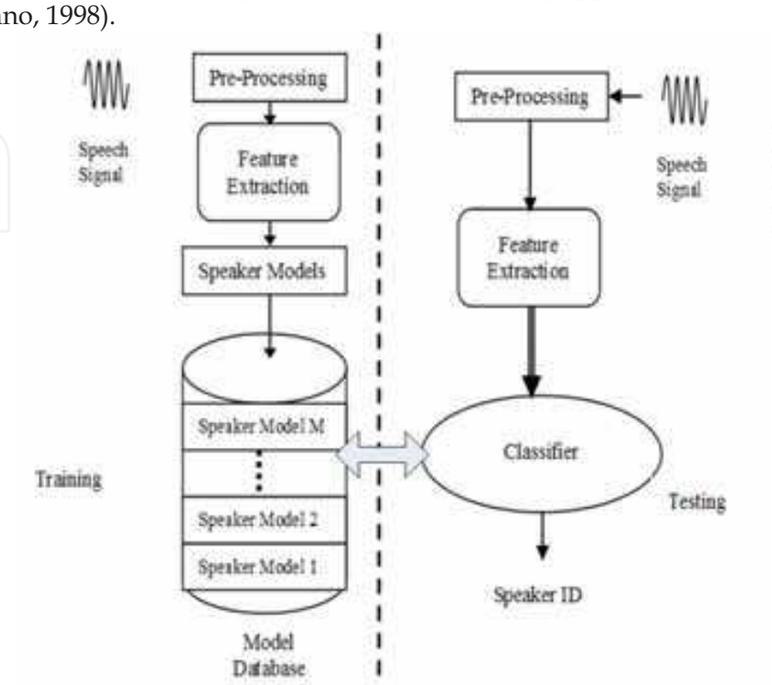


Fig. 1. overview of speaker verification systems

Vector Quantization (VQ) based speaker verification has been recognized as a successful method in the field of speaker recognition systems. A number of attempts have been made to use VQ methods with the GMM to optimize the performance of a speaker recognition system (Jialong et. Al, 1997) and (Singh et. al, 2003). The basic idea of VQ is to compress a large number of short term spectral vectors into a smaller set of code vectors. Until the development of GMM, vector quantization techniques were the most often applied methods in the field of speaker verification.
In this chapter we apply ITVQ algorithm (Tue et al.,2005), beside K-means and LBG VQ processes to estimate EM parameters. The ITVQ algorithm, which incorporates the Information Theoretic principles into the VQ process, was found to be the most efficient VQ algorithm (Sheeraz M. & Margaret L, 2008).

4. Feature Extraction Methods

Feature extraction is useful in speech (Davis, S. B. & P. Mermelstein,1980) and speaker recognition and the study of feature extraction has remained a core of research. A number of studies best support Mel-frequency cepstrum coefficients (MFCCs) (Reynolds, D. A. , 1994) and it does produce good results in most of the situations. In other studies, feature extraction based on pitch or energy contours (Peskin B. *et al.*,2003), glottal waveforms

(Plumpe, M. D. Et. al, 1999), or formant amplitude and frequency modulation (Jankowski C. R. jr. *et al.*,1996) are proposed, and good performance has been shown.
In their recent research (Sandipan, C. & Ghoutam, S.,2008) suggested, that the classification results can be significantly improved when the MFCC method is fused with the Inverse MFCC (IMFCC). This is because the IMFCC helps to capture the speaker specific information lying in the higher frequency range of the spectrum, which is largely ignored by the MFCC feature extraction method.

4.1 Mel-frequency cepstral coefficients (MFCC)

The primary concern of describing the MFCC algorithm here is to clearly map the working of Inverted MFCC and later in this chapter their fusion as a feature extraction set for GMM based on EM, K-means, LBG and ITVQ classifier. MFCC algorithm has been widely used for both the speech and speaker recognition in the recent years as it is designed keeping the human perception of listening as the core concern. According to psychophysical studies (Shaughnessy, D. O.,1987), human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale (Gold, B. & Morgan, N.,2002) (Fig. 2). Mel scale is defined as a logarithmic scale of frequency based on human pitch perception. Equal intervals in Mel units correspond to equal pitch intervals. It is given by,

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Where f_{mel} is the subjective pitch in Mels corresponding to f which is the actual frequency in Hz. This leads to the definition of MFCC, a baseline acoustic feature for Speech and Speaker Recognition applications, can be calculated by following steps.
Step.1*:* Let $\{x(n)\}_{n=1}^{M}$ represent a time-domain frame of pre-processed speech. The speech samples *x(n)* are first transformed to the frequency domain by the M-point Discrete Fourier Transform (DFT) and then the signal energy is calculated as,

$$|X(k)|^2 = \left| \sum_{n=1}^{M} x(n).e^{\left(\frac{-j2\pi nx}{M}\right)} \right|^2 \tag{2}$$

Where, *k=1,2,...,M* and *X(k) = DFT(x(n))*.
Step.2*:* This is followed by the construction of a filter bank with triangular frequency responses centered at equally spaced points on the Mel scale. Fig. 2 shows the frequency response of the i^{th} filter. The frequency response $\Phi_i(k)$ of this filter is calculated using Eq.(3).

$$\phi_i(k) = \begin{cases} 0 & for \quad k \le k_{b_{i-1}} \\ \dfrac{k - k_{b_{i-1}}}{k_{b_i} - k_{b_{i-1}}} & for \quad k_{b_{i-1}} \le k \le k_{b_i} \\ \dfrac{k_{b_{i+1}} - k}{k_{b_{i+1}} - k_{b_i}} & for \quad k_{b_i} \le k \le k_{b_{i+1}} \\ 0 & for \quad k \ge k_{b_{i+1}} \end{cases} \tag{3}$$

If N_F denotes the number of filters in the filter bank, then $\{k_{b_i}\}_{i=0}^{N_F+1}$ are the boundary points of the filters. The boundary points for each filter *i (i=1,.2,..., N_F)* are calculated as equally spaced points in the Mel scale using the following formula,

$$k_{b_i} = \left(\frac{M}{f_s}\right) f_{mel}\left[f_{mel}(f_{low}) + \frac{i\left\{f_{mel}(f_{high}) - f_{mel}(f_{low})\right\}}{N_F + 1} \right] \tag{4}$$

Where, f_s is the sampling frequency in Hz and $f_{low}=f_s/M$ and $f_{high} = S_F/2$ are the low and high frequency boundaries of the filter bank, respectively.

Step.3*:* In the next step, the output energies *E(i) (i=1,.2,..., N_F)* of the Mel-scaled band-pass filters are calculated as a sum of the signal energies $|X(k)|^2$ falling into a given Mel frequency band weighted by the corresponding frequency response $\Phi_i(k)$. This is given as,

$$E(i) = \sum_{x=1}^{\frac{M_S}{2}} |X(k)|^2 \Phi_i(k) \tag{5}$$

Where M_s is the number of DFT bins falling into the i^{th} filter.

Step.4*:* Finally, the Discrete Cosine Transform *(DCT)* of the log of the filter bank output energies *E(i) (i=1,.2,..., N_F)* is calculated yielding the final set of the MFCC coefficients C_m , given as

$$C_m = \sqrt{\frac{2}{N_F}} \sum_{l=0}^{N_F-1} \log[E(i+1)].\cos\left[m.\left(\frac{2l-1}{2}\right).\frac{\pi}{N_F} \right] \tag{6}$$

Where, *m=0,1,2,...,R-1*, and *R* is the desired number of the Mel Frequency Cepstral Coefficients.

4.2 Inverted Mel-frequency cepstral coefficients (MFCC)

The MFCC represent the information perceived by the human auditory system while the Inverse Mel Frequency Cepstral Coefficients capture the information which could have been missed by the MFCC (Yegnanarayana, B. et. al, 2005). The Inverted Mel Scale, which is shown as a dashed line in Fig.4, is defined by a filter bank structure that follows the opposite path to that of MFCC. The inverted filter bank structure can be generated by flipping the original filter bank around the mid frequency point f_c, of the filter bank frequency range (i.e. $f_c = (f_{high} - f_{low})/2$).

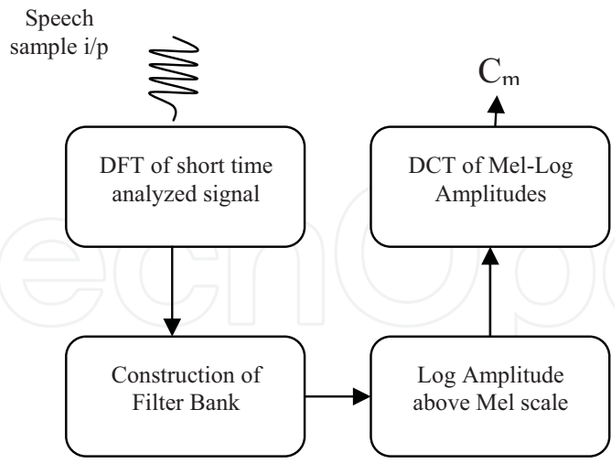


Fig. 2. Implementation structure of MFCC

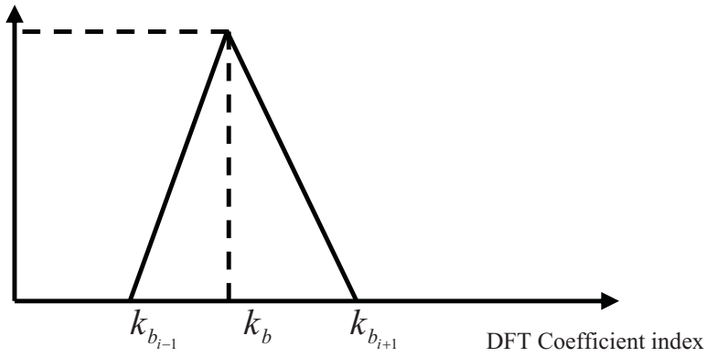


Fig.3. Response of a Mel scale Filter

The frequency responses $\hat{\Phi}_i(k)$ *(i=1,2,..., N_F)* for the inverted filter bank are given as,

$$\hat{\Phi}_i(k) = \Phi_{N_F+1-i}\left(\frac{M}{2}+1-k\right) \tag{7}$$

For a given frequency *f* in *Hz*, the corresponding inverted Mel-scale frequency $\hat{f}_{mel}(f)$ can be calculated as,

$$\hat{f}_{mel}(f) = 2195.2860 - 2595\log_{10}\left(1+\frac{4031.25-f}{700}\right) \tag{8}$$

The energies of the inverted filters outputs can be determined in the same way as for the non-inverted filters, i.e.,

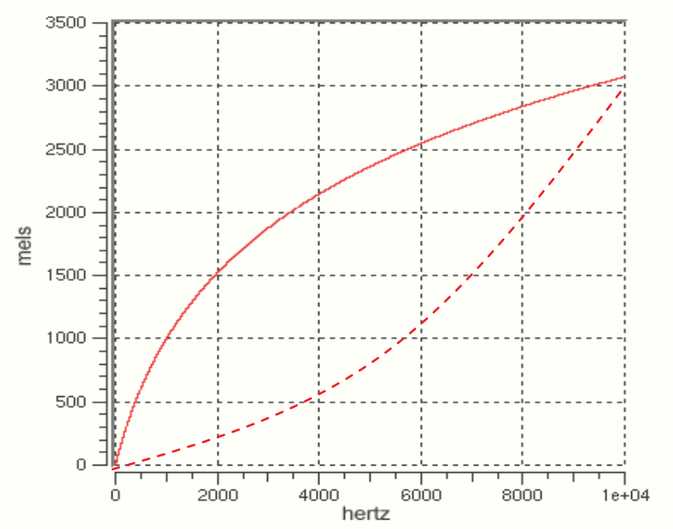


Fig. 4. Subjective pitch in Mels vs. frequency in Hz.

$$\hat{E}(i) = \sum_{x=1}^{\frac{M_S}{2}} |Y(k)|^2 \hat{\Phi}_i(k) \tag{9}$$

Finally, the *DCT* of the log filter bank energies is calculated, and the final Inverted Mel Frequency Cepstral Coefficients $\hat{C}_m$ are given as,

$$\hat{C}_m = \sqrt{\frac{2}{N_F}} \sum_{l=0}^{N_F-1} \log[\hat{E}(i+1)].\cos\left[m\left(\frac{2l-1}{2}\right)\frac{\pi}{N_F}\right] \tag{10}$$

Where, $m=0,1,2,...,R-1$, and R is the number of the Inverted Mel Frequency Cepstral Coefficients.

4.3 Fusion of MFCC and IMFCC

The idea of combining the classifiers to optimize the decision making process has been successfully applied in the fields of pattern recognition and classification (Mashao, DJ. & Skosan, M, 2006), (Murty, KSR. & Yegnanarayana, B.,2006). If the information supplied to the classifiers is complementary, such as the case of MFCC and IMFCC, the classification process could be largely improved (Sandipan, C. & Ghoutam, S.,2008) , (Chakroborty, S et. al, 2006).

The MFCC and the IMFCC feature vectors, containing complimentary information about the speakers, were supplied to a given classifier independently and the classification results for the MFCC features and for the IMFCC were fused in order to obtain optimal decisions in the process of speaker verification. A uniform weighted sum rule was adopted to fuse the scores from the two classifiers. If D_{MFCC} denotes the classification score based on the MFCC, and D_{IMFCC} denotes the classification score based on the IMFCC, then the combined score for the m^{th} speaker was given as,

$$D_m = \omega D_{MFCC} + (1-\omega) D_{IMFCC} \tag{11}$$

The constant value of ω = 0.5 was used in all cases. The speaker class was determined as,

$$m_{class} = \arg\left(\max_m D_m\right) \tag{12}$$

5. Vector Quantization (VQ) Methods

In this section of the chapter a number of VQ procedures are described, which have been used to optimize the EM parameters for GMM modelling.

5.1 K-means Method

It is an algorithm to classify or to group data based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data. K-means algorithm (Furui, S., 1989) was developed for vector quantization codebook generation. It represents each cluster by the mean of the cluster. Assume a set of vectors $X=\{x_1,x_2,x_3,\ldots..,x_T\}$ is to be divided into M clusters represented by their mean vectors $\{\mu_1, \mu_2, \mu_3,\ldots, \mu_M\}$ the objective of K-means algorithm is to minimize the total distortion given by,

$$total\_distortion = \sum_{i=1}^{M} \sum_{t=1}^{T} \left\| x_t - \mu_i \right\| \tag{13}$$

K-means is an iterative approach; in each successive iteration it redistributes the vectors in order to minimize the distortion. The procedure is outlined below:

(a) Initialize the randomized centroids as the means of M clusters.
(b) Data points are associated with the nearest centroid.
(c) The centroids are moved to the centre of their respective clusters.
(d) Steps b & c are repeated until a suitable level of convergence has been reached, i.e. the distortion is minimized.

When the distortion is minimized, redistribution does not result in any movement of vectors among the clusters. This could be used as an indicator to terminate the algorithm. The total distortion can also be used as an indicator of convergence of the algorithm. Upon convergence, the total distortion does not change as a result of redistribution. It **is** to be noted that in each iteration, K-means estimates the means of all the *M* clusters.

5.2 LBG Method

The LBG algorithm is a finite sequence of steps in which, at every step, a new quantizer, with a total distortion less or equal to the previous one, is produced. We can distinguish two phases, the initialization of the codebook and its optimization. The codebook optimization starts from an initial codebook and, after some iterations, generates a final codebook with a distortion corresponding to a local minimum. The following are the steps for LBG algorithm.

a. Initialization. The following values are fixed:

- N_C: number of codewords;
- $\varepsilon \geq 0$: precision of the optimization process;
- Y_0: initial codebook;
- $X = \{\mathbf{x}_j ; j = 1 , \ldots,N_P\}$: input patterns;

Further, the following assignments are made:

- $m = 0$; where m is the iteration number.
- $D_{-1} = +\infty$; where D is the minimum quantization error calculated at every m^{th} iteration.

b. Partition calculation. Given the codebook Y_m, the partition $P(Y_m)$ is calculated according to the *nearest neighbour condition*, given by

$$S_i = \{x \varepsilon X : d(x, y_i) \geq d(x, y_j), \quad i=1,2,\ldots,N_C. \\ j = 1,2,....,N_C, j \neq i\} \tag{14}$$

c. Termination condition check. The quantizer distortion ($D_m = D(\{Y_m,P(Ym)\})$ is calculated according to following equation.

$$MQE \equiv D(\{Y,S\}) = \frac{1}{N_P}\sum_{p=1}^{N_P} d(x_p, q(x_p)) = \frac{1}{N_P}\sum_{i=1}^{N_C} D_i \tag{15}$$

Where D_i indicates the total distortion of i^{th} cell.

If $\left|(D_{m-1} - D_m)\right| / D_m \leq \varepsilon$ then the optimization ends and Y_m is the final returned codebook.

d. New codebook calculation. Given the partition $P(Y_m)$, the new codebook is calculated according to the Centroid condition. In symbols:

$$Y_{m+1} = X\,(P(Y_m)) \tag{16}$$

After, the counter *m* is increased by one and the procedure follows from step b.

5.3 Information Theoretic VQ

The Vector Quantization methods are commonly used in the process of feature classification. The ITVQ (Tue, L. et. al, 2005) algorithm uses a new set of concepts from information theory and provides a computationally very efficient technique, which eliminates many disadvantages of classical vector quantization algorithms. Unlike LBG, this algorithm relies on minimization of a well defined cost function. The cost function used in LBG and K-means algorithms is defined as an average distortion (or distance), and as such, it is very complex and may contain discontinuities making the application of traditional optimization procedures very difficult (Erwin, E. et. al, 1991).

According to the information theory a distance minimization is equivalent to the minimization of the divergence between distribution of data and distribution of code vectors. Both distributions can be estimated using the Parzen density estimator method (Tue, L. et. al, 2005).

The ITVQ algorithm is based on the principle of minimizing the divergence between Parzen estimator of the code vectors density distributions and a Parzen estimator of the data distribution. The Parzen density estimator is given as,

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i) \tag{17}$$

Where *K(.)* is the Gaussian Kernel, x is the independent variable for which we seek the estimate and x_i represents the data points. The Parzen estimate of the data has N kernels, where N is the number of data points, and the Parzen estimator of the code vectors has M kernels, where M is the number of code vectors and M<<N.

The density estimation is followed by minimization of the divergence between data points and centroids. In order to minimize the divergence between the data points distribution *a(x)* and the centroids distribution *b(x),* the following expression is minimized.

$$D_{C-S}(a(x), b(x)) = \\ \log \int a^2(x)dx - 2\log \int a(x)b(x)dx + \log \int b^2(x)dx \tag{18}$$

Where, *a(x)* and *b(x)* denote the Parzen density estimates for the data and centroids, respectively.
The cost function in Eq. (18) is minimized through a gradient descent search, which iteratively changes the positions of centroids until the decrease rate of the cost value becomes sufficiently small. The first term in Eq.(18), $\log \int a^2(x)dx$, represents the Renyi's quadratic entropy of data points, the third term, $\log \int b^2(x)dx$, represents the Renyi's quadratic entropy of centroids, and the second term, $-2\log \int a(x)b(x)dx$, is the *2log* of the cross information potential between the densities of the centroids and the data. Since the entropy of the data points remains constant during the iterations, the minimization of the cost function in Eq. (18) is equivalent to the maximization of the sum of the entropy of the centroids and the cross information potential between the densities of the centroids and the data.
As explained in more detail in (Tue, L. et. al, 2005), a typical ITVQ algorithm makes use of an annealing procedure, which allows the algorithm to escape from local minima.

6. Gaussian Mixture Models

In this section of the chapter we describe the modelling methods. GMM use EM procedure for the optimization however the use of VQ methods is proposed here.

6.1 GMM with EM

The Gaussian Mixture Model (GMM) (Douglas, A.R., 1995) with Expectation maximization is a feature modeling and classification algorithm widely used in the speech-based pattern recognition, since it can smoothly approximate a wide variety of density distributions.
The probability density function (pdf) drawn from the GMM is a weighted sum of M component densities given as,

$$p(x \mid \lambda) = \sum_{k=1}^{M} p_k b_k(x) \tag{19}$$

Where x is a D-dimensional random vector, $b_k(x)$, k =1… M are the component densities and p_k, k =1… M are the mixture weights. Each component density is a D-variate Gaussian function of the form

$$b_k(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_k\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)'\sum_i^{-1}(x-\mu_k)\right\} \tag{20}$$

Where μ_i is the mean vector and $\sum$ is the covariance matrix. The mixture weights satisfy the constraint that $\sum_{k=1}^{M} p_k = 1$ .The complete Gaussian mixture density is the collection of the mean vectors, covariance matrices and mixture weights from all components densities,

$$\lambda = \{p_k, \mu_k, \sum_k\}, k = 1,....., M \tag{21}$$

Each class is represented by a mixture model and is referred by the class model λ.
The Expectation Maximization (EM) algorithm is most commonly used to iteratively derive class models. The EM algorithm initialized with a speaker model λ and estimates at each iteration a new model $\bar{\lambda}$ such that $p(X \mid \bar{\lambda}) \geq p(X \mid \lambda)$.

6.2 GMM with VQ

Although EM algorithm performs well but the literature has suggested that it suffers with some of the problems which can enhance its performance for pattern recognition applications such as speaker recognition (Ueda, N. & R. Nakano, 1998). The areas where the performance improvement can be achieved are listed below.

1. The number of mixtures is mostly set *a priori.*
2. The initialization procedure applied to set the parameters affects the final result.
3. EM converges to local optima instead of global optima.

Thus investigation of alternative training algorithms is unavoidable. However this may include either modifying the standard EM steps or by proposing enhanced optimization procedures. We in this paper propose the use of several VQ methods to replace the maximization step of EM algorithm. At each EM iteration expectation is set which is given by,

$$h_{kj} = \frac{g_j \left|\sum_j\right|^{-1/2} \exp\left[-(1/2)(x_k-\mu_j)^T \sum_j^{-1}(x_k-\mu_j)\right]}{\sum_l g_l \left|\sum_j\right|^{-1/2} \exp\left[-(1/2)(x_k-\mu_l)^T \sum_j^{-1}(x_k-\mu_l)\right]} \tag{22}$$

The above equation is the evaluation of a speaker model at each EM iteration. The numerator is the pdf of a target model and the denominator is the sum of all the pdf's.

However the next part of EM based GMM is to obtain the iterative updates where we propose to use the cost function of VQ methods. We apply the clustering techniques such as K-means, LBG and ITVQ to optimize the means. However the covariances are computed as evaluated in the initialization procedure, however based on the new clusters/distribution of the speaker data. The iterative weights are the updates from the new expectation h_{kj} as evaluated in the EM procedure (see equation 22).

$$g_j^{(n+1)} = \frac{1}{n}\sum_k h_{kj}^{(n)} \tag{23}$$

The K-means algorithm has been applied for finding a robust model approximation to the GMM in (Singh et. al, ,2003) and (Pelecanos et. al, 2000). Hence we are using a number of vector quantization algorithms including K-means, LBG and recently designed ITVQ to investigate its suitability to avoid local convergence when using EM algorithm. We also compare the performance of ITVQ over other vector quantization approaches.

How the cost minimization procedure is implemented for each clustering technique is described in section 2 and the distortion function for each the clustering techniques are listed in equations (1), (3) and (11) respectively.

A multi-dimensional Gaussian is calculated using the mean and variance statistics from the test vectors in each code vector region, with the training vectors already grouped into their code books. An approximation of the GMM is determined by estimating the mixture weights p_k, means μ_k, and covariances $\sum_k$. Each mean μ_k is assigned to its corresponding code vector, $\vec{c}_k$. The covariance matrix Σ_k for each GMM is calculated from the variances of the vector observations in each code vector region. To achieve the optimal approximation the feature vectors need to be well clustered and the VQ based GMM also need to have the features uncorrelated, for many applications including SV it is difficult to satisfy this condition, however by attempting to match these requirements, model estimation errors could be minimised. Normalization techniques (Mariethoz, J. & S. Bengio, 2005), (Barras, C. & J. Gauvain, 2003) are also applied for this purpose to reduce the mismatch of features.

7. Experiments

In this section of the chapter we describe the speaker verification tests. First we describe the speech corpora used for the experiments, secondly pre-processing and feature extraction settings are summarised. Finally the DET plots and EER scores are provided.

7.1 Speech Corpus

Recently the annual NIST speaker recognition evaluation (SRE) has become the state of the art corpora for evaluating the methods in the field of speaker recognition. GMM-based systems have been widely tested on NIST SRE. We evaluated our methods on NIST SRE 04. The 2004 evaluation uses conversational speech data collected in the Mixer Project using the Linguistic Data Consortium's new "Fishboard" platform. The data is not used by the previous evaluations. This data is mostly conversational telephone speech in English but it also includes some speech in languages other than English such as Spanish and Arabic. The evaluation includes twenty-eight different speaker detection tests defined by the duration and type of both the training and the test segments of the individual trials of which these tests are composed (NIST, 2004).

The performance of the system is based on the detection error tradeoffs DET function. This detection cost function is defined as a weighted sum of miss and false alarm error probabilities. The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{FalseAlarm}$ and the *a priori* probability of the specified target speaker, P_{Target}.

$$\begin{aligned} C_{Det} &= C_{Miss} \times P_{Miss|T\arg et} \times P_{T\arg et} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|NonT\arg et} \times (1 - P_{T\arg et}) \end{aligned} \tag{24}$$

7.2 Pre-processing and feature extraction

The performance of a modeling method depends intensively on how well the features are derived. The speaker speech is passed through a number of stages to avoid a number of drawbacks. First, the speech is segmented into frames by a 20ms sliding hamming window with a 15ms skip rate is used to obtain a sequence of frames for which speech feature vectors are estimated.

The frames are then passed through a speech activity detector (SAD) to remove the silence–noise frames. It is a self-normalizing and energy based detector that detects the noise floor of the speech signal frames. SAD can adapt to changing noise conditions (Erwin, E. et. al, 1991) and it removes 20–25% of the signal from conversational telephone recordings such as that in the fisher database from which the NIST SRE 04 corpus is derived.

After removing the silence-noise frames mel-scale cepstral feature vectors are derived from the speech frames. The mel-scale cepstrum is the discrete cosine transform of the log spectral energies of the speech segment. The spectral energies are calculated over logarithmically spaced filters with increasing bandwidths (mel-filters). The inverted MFCC are also computed, a detailed description of the feature extraction steps can be found in section 4. For band limited telephone speech, cepstral analysis is performed only over the Mel filters in the telephone pass band (300–3400 Hz). All cepstral coefficients except its zeroth value (the DC level of the log-spectral energies) are retained in the processing.

7.3 Experimental results

The experiments are designed to compare the performance of EM, K-means, LBG and ITVQ. All systems use the same test, train, GMM settings such as initialization procedure and number of mixtures and feature configurations. This point is important not only for direct comparisons of the modeling methods but also in assessing the fusion strategies. The GMM baseline is used; GMM models have *128* mixtures and features of various sizes individually and in fusion are used.

A standard GMM-UBM system was developed using NIST 04 SRE. The standard train and test conditions for NIST 04 were used to investigate the proposed protocols. We train the coefficient data by using the different vector quantization versions along with GMM to optimize the speaker models as described in section 5.1. Thre feature combinations are used to conduct the experiments. The feature sets include 13 MFCC, and 13 IMFCC.

Examination of the speaker verification results are based on equal error rate (EER) values which are listed in fig3 to fig 5. EER is the most commonly used criteria to compare the recognition rates. The results obtained can be summarized as follows:

1. In all cases the MFCC-IMFCC features give better results than the MFCC features.
2. The GMM-K-means and the GMM-LBG algorithms show very similar performance but both yield results that are clearly below the performance level of the GMM-EM method.
3. The results given by the GMM-ITVQ, on the other hand, are very close to the results given by the GMM-EM algorithm.

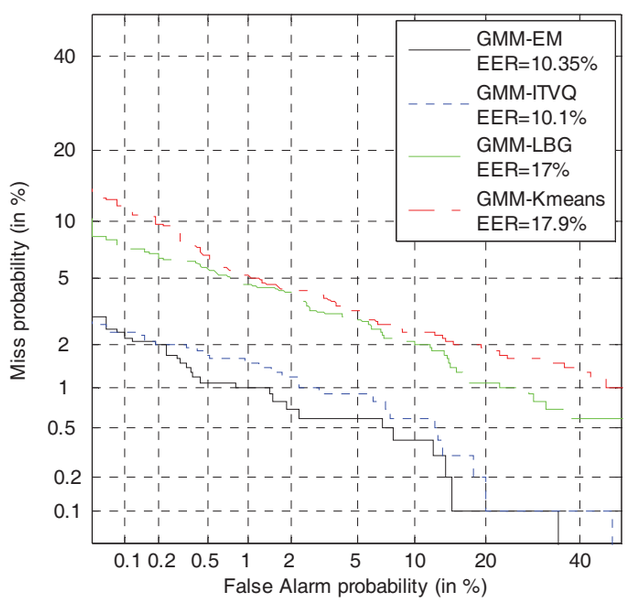


Fig. 3. DET plot for MFCC features

The experiments with different implementations of vector quantization were carried out such as K-means, LBG and ITVQ to make a comparative analysis and we observed that ITVQ behaves a better vector quantization approach than the other VQ implementations. The VQ methods were used as part of the standard GMM.

8. Conclusion

A number of tasks were focused in the experiments. One, to test the vector quantization algorithms for the speaker verification when used along with the GMM, Second to compare the performance of different VQ techniques, and third is to apply the recently established VQ technique called ITVQ for some experimental data.

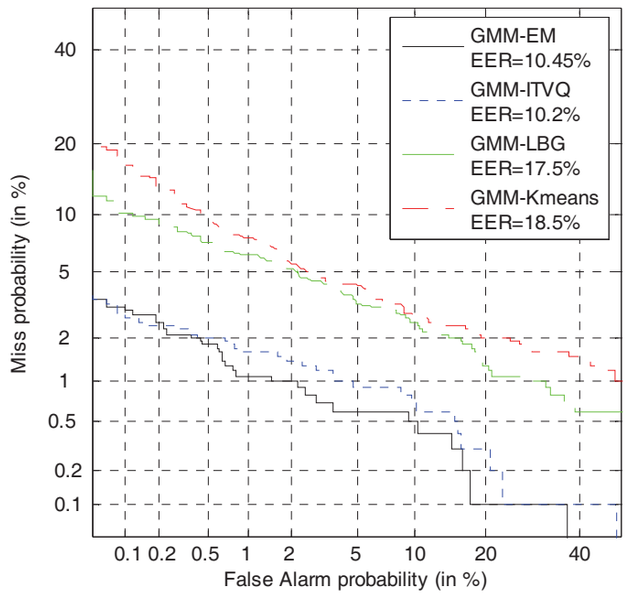


Fig. 4. DET plot for IMFCC features

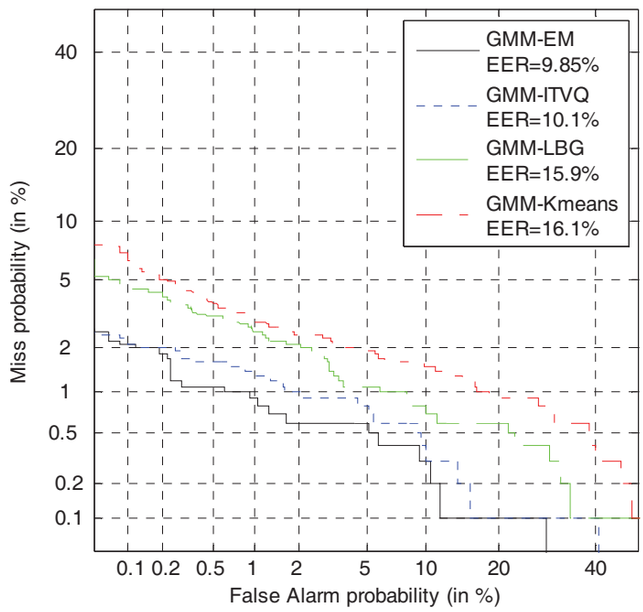


Fig. 5. DET plot for MFCC-IMFCC features

The GMM-EM speaker verification approach was compared with the GMM-Kmeans, GMM-LBG and GMM-ITVQ methods. The speaker verification tests were performed using combinations of the feature extraction methods such as MFCC, IMFCC and their fusion.
The ITVQ algorithm, which incorporates the Information Theoretic principles into the VQ process, was found to be the most efficient alternative for the EM algorithm. It gives correct classification rates at a similar level to that of EM.
In some applications the small degradation of performance in case of the GMM-ITVQ compare to the GMM-EM can be compensated by ITVQ advantages, such as the computational simplicity and ability to escape local minima, which provides a potential for better performance in case of irregular and complex potential functions.

9. References

Barras, C. & J. Gauvain, "Feature and score normalisation for speaker verification of cellular data", IEEE International conference on acoustics speech and signal processing, Vol.2, pp.49-52, 2003.

Chakroborty, S.; A. Roy, G. Saha "Fusion of a Complementary Feature Set with MFCC for Improved Closed Set Text-Independent Speaker Identification", IEEE International Conference on Industrial Technology, pp387- 390, Dec.2006.

Davis, S. B. & P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

Douglas, A.R.; "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, vol. 3, No. 1, pp. 72-83, January 1995.

Erwin, E., Obermayer, K., Schulten, K.: Selg organizing maps, ordering, convergence properties and energy functions. In: Biological Cybernetics. vol. 67, No.1, pp. 47-55, 1991.

Ethem A."Soft vector quantization and EM algorithm", Neural networks, Vol 11, issue 3, , Pages 467-477, April 1998

Furui, S.: Digital Speech Processing, Synthesis and Recognition, Marcel Dekker Inc., New York, (1989).

Gold, B. & Morgan, N. " *Speech and Audio Signal Processing"*, Part- IV, Chap.14, pp. 189-203, John Willy & Sons ,2002.

Hedelin, P. Skoglund, J. "Vector quantization based on Gaussian Mixture Models", IEEE transactions on **speech** and audio processing, Vol 8, issue 4, pp. 385-401, July 2000.

Jialong, H.; L. Liu, and P. Gunther, "A new codebook training algorithm For VQ-based speaker recognition", IEEE international conference on acoustics, speech and signal processing, Vol. 2, pp.1091-1094, 1997.

Jankowski C. R. jr. *et al.*, "Fine structure features for speaker identification," in *Proc. ICASSP*, 1996, pp. 689–692.

Mashao, DJ. & Skosan, M., "Combining Classifier Decisions for Robust Speaker Identification", *Pattern Recognition*, vol. 39, pp. 147-155, 2006.

Murty, KSR. & Yegnanarayana, B., "Combining evidence from residual phase and MFCC features for speaker recognition", *IEEE Signal Processing Letters*, vol 13, no. 1, pp. 52-55, Jan. 2006.

Mariethoz, J. & S. Bengio, "A unified framework for score normalisation techniques applied to Text independent Speaker verification", IEEE signal processing letters, Vol.12, No.7,July 2005.

NIST "The NIST 2004 Speaker Recognition Evaluation Plan" http://www.itl.nist.gov/iad/mig/tests/sre/2004/index.html

Peskin B. *et al.*, "Using prosodic and conversational features for high performance speaker recognition: Report from JHU WS02," in *Proc. ICASSP*, vol. 4, 2003, pp. 792–795.

Plumpe, M. D. ; T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

Pelecanos, J.; S. Myers, S. Sridharan, and V. Chandran, "Vector Quantization Based Gaussian Modelling for Speaker Verification", In: International conference on pattern recognition", Vol. 3, pp. 294-297, 2000.

Reynolds, D. A. "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994

Reynolds, D. ; Quatieri, T. & Dunn, R. "Speaker verification using adapted Gaussian mixture models", *Digital Signal Process.*, vol.10,pp. 19-41, 2000.

Singh, G.; A. Panda, S. Bhattacharyya, and T. Srikanthan, " Vector quantization techniques for GMM based speaker verification", IEEE international conference on acoustics, speech and signal processing, Vol. 2, pp. II65-II68, 2003.

Sheeraz M. & Margaret L, "Speaker Verification Based on Information Theoretic Vector Quantization", Wireless Networks, Information Processing and Systems, Springer Berlin Heidelberg, vol. 20, pp. 391-399, November 14, 2008.

Sandipan, C. & Ghoutam, S., "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter" IJSP, Vol. 5, No.1, 2008.

Shaughnessy, D. O., "*Speech Communication Human and Machine*", Addison-Wesley, New York, 1987.

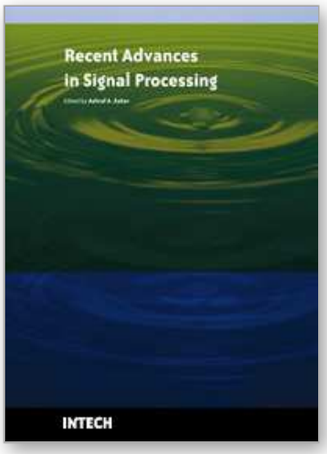
Tue, L.; Anant, H.; Deniz, E. & Jose, C. "Vector Quantization using information theoretic concepts", *Natural Computing*, vol . 4, Issue. 1, pp. 39 - 51, January 2005.

Tue, L., Anant, H., Deniz, E., Jose, C., "Vector quantization using information theoretic concepts", In: Natural Computing: an international journal, vol. 4, Issue. 1, pp. 39 – 51. 2005.

Ueda, N. & R. Nakano, "Deterministic annealing EM algorithm," *Neural Netw.*, no. 11, pp. 271–282, 1998.

Ververidis, D. Kotropoulos, C. "Gaussian mixture modeling by exploiting the mahalanobis distance", IEEE transactions on signal processing, Vol 56, issue 7, pp. 2797-2811 , July 2008.

Yegnanarayana, B.; Prasanna S.R.M., Zachariah J.M. and Gupta C. S., "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 4, pp. 575-582, July 2005.



Recent Advances in Signal Processing

Edited by Ashraf A Zaher

ISBN 978-953-307-002-5

Hard cover, 544 pages

Publisher InTech

Published online 01, November, 2009

Published in print edition November, 2009

The signal processing task is a very critical issue in the majority of new technological inventions and challenges in a variety of applications in both science and engineering fields. Classical signal processing techniques have largely worked with mathematical models that are linear, local, stationary, and Gaussian. They have always favored closed-form tractability over real-world accuracy. These constraints were imposed by the lack of powerful computing tools. During the last few decades, signal processing theories, developments, and applications have matured rapidly and now include tools from many areas of mathematics, computer science, physics, and engineering. This book is targeted primarily toward both students and researchers who want to be exposed to a wide variety of signal processing techniques and algorithms. It includes 27 chapters that can be categorized into five different areas depending on the application at hand. These five categories are ordered to address image processing, speech processing, communication systems, time-series analysis, and educational packages respectively. The book has the advantage of providing a collection of applications that are completely independent and self-contained; thus, the interested reader can choose any chapter and skip to another without losing continuity.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen