Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC

Satyanand Singh
Associate Professor
Dept of Electronics and Comm Engineering
St Peter's Engineering College, Near Forest Academy, Dhoolapally, Hyderabad

Dr. E.G. Rajan
Professional Member, ACM
Founder President and Director Pentagram Research Center (P) Limited
1073, Road No. 44, Jubilee Hills, Hyderabad

ABSTRACT
Front-end or feature extractor is the first component in an automatic speaker recognition system. Feature extraction transforms the raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal. Since the front-end is the first component in the chain, the quality of the later components (speaker modeling and pattern matching) is strongly determined by the quality of the front-end. In other words, classification can be at most as accurate as the features. Over the years, Mel-Frequency Cepstral Coefficients (MFCC) modeled on the human auditory system has been used as a standard acoustic feature set for speech related applications. In this paper it has been shown that the inverted Mel-Frequency Cepstral Coefficients is one of the performance enhancement parameters for speaker recognition, which contains high frequency region complementary information in it. This paper introduces the Gaussian shaped filter (GF) while calculation MFCC and inverted MFCC in place of traditional triangular shaped bins. The main idea is to introduce a higher amount of correlation between subband outputs. The performance of both MFCC and inverted MFCC improve with GF over traditional triangular filter (TF) based implementation, individually as well as in combination. In this study the Vector Quantization (VQ) feature matching technique was used, due to high accuracy and its simplicity. The proposed investigation achieved 98.57% of efficiency with a very short test voice sample 2 seconds.

Keywords
GF, Triangular Filter, Subbands, Correlation, MFCC, inverted MFCC, Vector Quantization

1. INTRODUCTION
A speaker recognition system mainly consists of two main module, speaker specific feature extractor as a front end followed by a speaker modeling technique for generalized representation of extracted features [1, 2]. Since long time MFCC is considered as a reliable front end for a speaker recognition application because it has coefficients that represents audio, based on perception [3, 4]. In MFCC the frequency bands are positioned logarithmically (on the mel-scale) which approximated the human auditory systems response more closely than the linear spaced frequency bands of FFT or DCT. This allows for better processing of data. An illustrative speaker recognition system is shown in fig. 1.The state of the art speaker recognition research primarily investigates speaker specific complementary information relative to MFCC. It has been observed that the performance of speaker recognition improved significantly when complementary information is Ex-Ored with MFCC in feature level either by simple concatenation or by combining models scores. The main complementary information is pitch [5], residual phase [6], prosody [7], dialectical features [8] etc. These features are related with vocal chord vibration and it is very difficult to extract speaker specific information. It has been shown that complementary information can be captured easily from the high frequency part of the energy spectrum of a speech frame via reversed filter bank methodology [9]. There are some features of speaker which used to present at high frequency part of the spectrum and generally ignored by MFCC that can be captured by inverted MFCC is proposed in this paper. The complementary information captured by inverted MFCC is modeled by VQ [10] technique.
So the present study was undertaken with the objective of to improve the efficiency of speaker recognition by using inverted MFCC.

2. METHODOLOGY
In the present investigation GF were used as the averaging bins instead of triangular for calculating MFCC as well as inverted MFCC in a typical speaker recognition application [11, 12]. There are three main inspiration of using GF. First inspiration is GF can provide much smoother transition from one subbands to other preserving most of the correlation between them. Second inspiring point is the means and variances of these GFs can be independently chosen in order to have control over the amount of overlap with neighboring subbands. Third inspiring point is the filter design parameters for GF can be calculated very easily from mid as well as end-points located at the base of the Original TF used for MFCC and inverted MFCC. In this investigation both MFCC and inverted MFCC filter bank are realized using a moderate variance where a GF's coverage for a subbands and the correlation is to be balanced. Results show that GF based MFCC and inverted MFCC perform better than the conventional TF based MFCC and inverted MFCC individually. Results are also better when GF based MFCC & inverted MFCC is combine together by modulo two adder (Ex-OR) their model scores in comparison to the results that are obtained by combining MFCC and inverted MFCC feature sets realized using traditional TF [13]. All the implementations have been done with VQ-Linde Buzo Gray (LBG) algorithm as speaker modeling paradigm [14].

3. MEL FREQUENCY AND THEIR CALCULATION BY USING GAUSSIAN FILTERS

3.1 Mel-Frequency Cepstral Coefficients using triangular filters
According to psychophysical studies human perception of the frequency content of sounds follows a subjectively defined

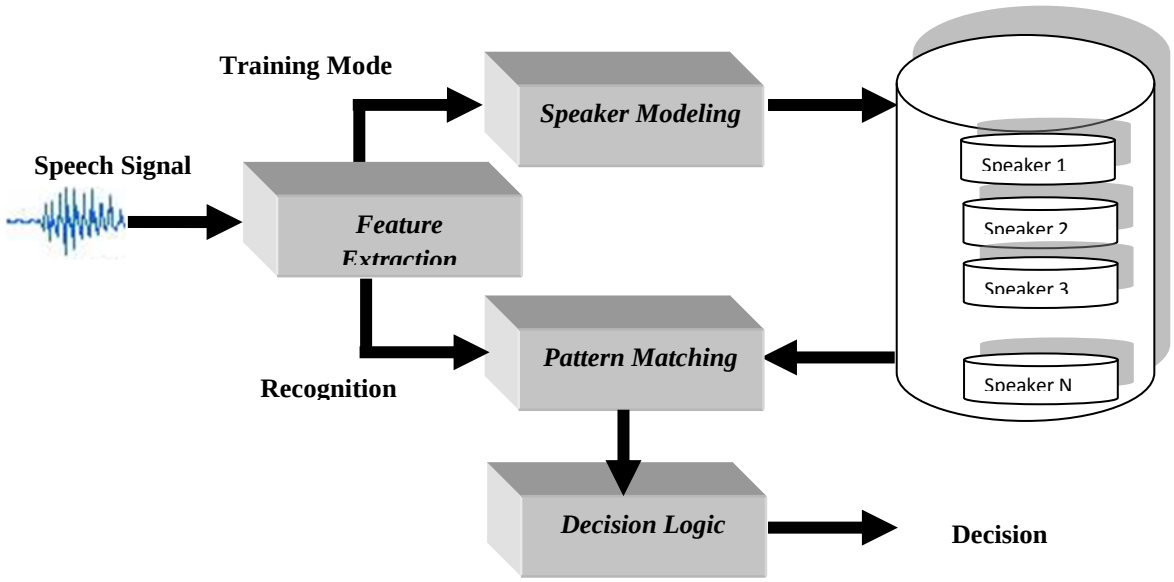

Fig 1: Speaker recognition system

nonlinear scale called the Mel scale [15, 16]. MFCC is the most commonly used acoustic features for speech/speaker recognition. MFCC is the only acoustic approach that takes human perception (Physiology and behavioral aspects of the voice production organs) sensitivity with respect to frequencies into consideration, and therefore is best for speaker recognition. The acoustic model is defined as,

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

Where f_{mel} is the subjective pitch in Mels corresponding transfer function and the actual frequency in Hz. This leads to the definition of MFCC, a baseline acoustic feature for speech and speaker recognition applications, which can be calculated as follows [17]

Let $\{y(n)\}_{n=1}^{Ns}$ represent a frame of speech that is preemphasized and Hamming-windowed. First, y(n) is converted to the frequency domain by an Ms-point DFT which leads to the energy spectrum,

$$Y|k|^2 = \left| \sum_{n=1}^{Ms} y(n).e^{\left(-\frac{j2\pi nk}{Ms}\right)} \right|^2 \quad (2)$$

Where $1 \le k \le Ms$. this is followed by the construction of a filter bank with Q unity height TFs, uniformly spaced in the Mel scale eqn. (1). The filter response Ψi (k) of the ith filter in the bank fig. 2 is defined as,

$$\Psi_i(k) = \begin{cases} 0 & for\ k \le \hat{k}_{b\,i-1} \\ \dfrac{k - k_{bi-1}}{k_{bi} - k_{bi-1}} & for k_{b\,i-1} \le k \le k_{bi} \\ \dfrac{k_{bi+1} - k}{k_{bi+1} - k_{bi}} & for\ k_{bi} \le k \le k_{b+1} \\ 0 & for\ k \ge k_{b\,i+1} \end{cases} \quad (3)$$

Where $1 \le i \le Q$, Q is the number of filters in the bank, $\{k_{bi}\}_{i=0}^{Q+1}$ are the boundary points of the filters and k denotes the coefficients index in the Ms point DFT. The filter bank boundary points are equally spaced in the Mel scale which is satisfying the definition,

$$k_{bi} = \left(\frac{M_s}{F_s}\right) f_{mel}^{-1}\left[f_{mel}(f_{low}) + \frac{i\{f_{mel}(f_{high}) - f_{mel}(f_{low})\}}{Q+1} \right] \quad (4)$$

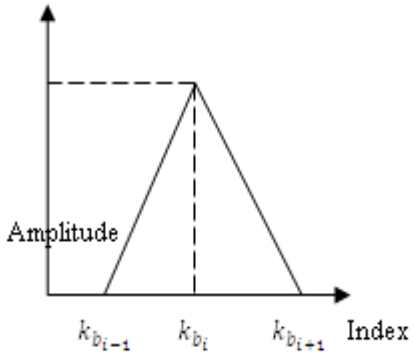

Fig 2: Response I (k) of a typical Mel scale filter

where the function fmel (•) is defined in eqn. (1), Ms is the number of points in the DFT eqn. (2), F_s is the sampling frequency, f_{low}, and f_{high} are the low and high frequency boundaries of the filter bank and f_{mel}^{-1} is the inverse of the transformation in eqn.(1) defined as,

$$f_{mel}^{-1}(f_{mel}) = 700.\left[10^{\frac{f_{mel}}{2595}} - 1\right] \quad (5)$$

The sampling frequency F_s and f_{low}, f_{high} frequencies are in Hz while f_{mel} is in Mels. In this work, F_s is 8 kHz. Ms is taken as 256 f_{low} = Fs/Ms = 31.25 Hz while f_{high} = F_s /2 = 4 kHz. Next, this filter bank is imposed on the spectrum calculated in eqn. (2). The outputs $e(i)_{i=1}$ Q of the Mel-scaled band-pass filters can be calculated by a weighted summation

between respective filter response Ψi(k) and the energy spectrum $|Y(k)|^2$ as

$$e(i) = \sum_{k=1}^{\frac{M_s}{2}} |Y(K)|^2 \cdot \Psi_i(k) \quad (6)$$

Finally DCT is taken on the log filter bank energies $\{\log[e(i)]\}_{i=1}$ and the final MFCC coefficients C_m can be written as.

$$C_m = \sqrt{\frac{2}{Q} \sum_{l=0}^{Q-1} log[e(i+1)].cos\left[m\left(\frac{2l-1}{2}\right).\frac{\pi}{Q}\right]} \quad (7)$$

Where $0 \le m \le R-1, R$ is the desired number of cepstral features.

3.2 Mel-Frequency Cepstrum Coefficients using Gaussian filters.

The transfer function of any filter is asymmetric, tapered and filter does not provide any weight outside the subband that it covers. As a result, the correlation between a subband and its nearby spectral components from adjacent subbands is lost. In this investigation a GF is proposed, which produced gradually decaying weights at its both ends and symmetric for compensating possible loss of correlation. Referring to eqn. (3), the expression for GF can be written as [18]

$$\Psi_i^{g_{MFCC}} = e^{\frac{(k-k_{bi})^2}{2\sigma_i^2}} \quad (8)$$

Where k_{bi} is a point between the ith transfer boundaries located at it base and it is considered here as a mean of the ith GF while the σ_i is the standard deviation and can be defined as,

$$\sigma_i = \frac{k_{b+1} - k_{bi}}{\alpha} \quad (9)$$

Where α is the variance controlling parameter. In fig. 3 shows the transfer function for different values of α, k_{bi} is the centre point of all above filters and after that transfer functions are gradually decaying. However the Gaussian with higher variance shows larger correlation with nearby frequency component. Thus selection of α is a critical part for setting the variances of GF. In the present study the value of α = 5 then eqn. (9) can be written as,

$$5\sigma = k_{bi+1} - k_{bi} \quad (10)$$

98% of subband is covered once α = 5, is selected. Probability $[5\sigma \ge (k_{bi+1} - k_{bi})]$ = 0.98. Therefore, α = 5 can provide better correlation with nearby subbands in comparison to α = 6. In this study, we have chosen α = 4 to design filters for the MFCC filter bank. Thus, a balance is achieved where significant coverage of a particular subband is ensured while allowing moderate correlation between that subband and neighboring ones. fig. 4. shows MFCC filter bank structure using triangular and Gaussian bins. The cepstral vector using GFs can be calculated from the filter's response eqn. (8) which is as follows

$$e^{g_{MFCC}}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(K)|^2 \ \Psi^{g_{MFCC}} i(k) \quad (11)$$

and,

$$C^{g_{MFCC}} m = \sqrt{\frac{2}{Q}} \sum_{i=1}^{Q-1} log\ [e^{g_{MFCC}}(i+1)]. cos\left[m\left(\frac{2l-1}{2}\right).\frac{\pi}{Q}\right] \quad (12)$$

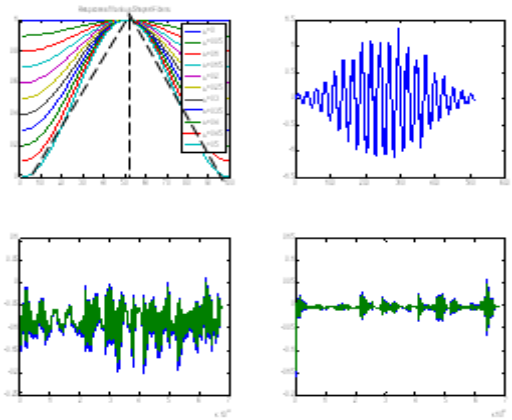

Fig 3: Response of various shaped filters

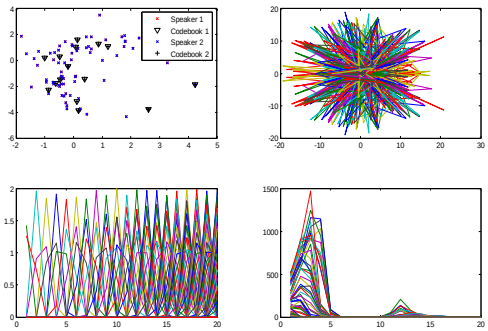

Fig 4: Filter bank structure for canonical MFCC and normalized MFCC filter bank

Here last 20 coefficient from both models are used and the value of Q=22 and R = 25 are taken.

4. INVERTED MEL FREQUENCY CEPSTRAL COEFFICIENTS CALCULATION BY GAUSSIAN FILTERS

4.1 Inverted Mel-Frequency Cepstral Coefficients using triangular filters

The main objective is to capture that information which has been missed by original MFCC [19]. In this study the new filter bank structure is obtained simply by flipping the original filter bank around the point f = 2 kHz which is precisely the mid-point of the frequency range considered for speaker recognition applications. This flip-over is expressed mathematically as,

$$\hat{\Psi}_i(k) = \Psi_{Q+1-i}\left(\frac{M_s}{2} + 1 - k\right) \quad (13)$$

Where $\hat{\Psi}_i(k)$ is the inverted Mel Scale filter response while $\Psi_i(k)$ is the response of the original MFCC filter bank $1 \le i \le Q$ and Q is the number of filters in the bank. From eqn. (13) we can derive an expression for $\hat{\Psi}_i(k)$ with analogous to eqn. (3) $\hat{\Psi}_i(k)$for the original MFCC filter bank.

$$\hat{\Psi}_i(k) = \begin{cases} 0 & for\ k \le \hat{k}_{b\,i-1} \\ \dfrac{k - \hat{k}_{b\,i-1}}{k_{bi} - k_{bi-1}} & for\ \hat{k}_{b\,i-1} \le k \le \hat{k}_{bi} \\ \dfrac{\hat{k}_{b_{i+1}} - k}{\hat{k}_{b_{i+1}} - \hat{k}_{bi}} & for\ \hat{k}_{bi} \le k \le \hat{k}_{b+1} \\ 0 & for\ k \ge \hat{k}_{b\,i+1} \end{cases} \quad (14)$$

Where $l \le k \le M_s$ and $\{\hat{k}_{bi}\}_{i=0}^{Q+1}$
Here inverted mel-scale is defined as follows

$$\hat{f}_{mel}(f) = 2195.2860 - 2595\ log_{10}\left(1 + \frac{4031.25 - f}{700}\right) \quad (15)$$

Where $\hat{f}_{mel}(f)$ is subjective pitch in the new scale corresponding to f, the actual frequency in Hz .
The filter outputs$\{\hat{e}(i)\}_{i=1}^{Q}$ in the same way as MFCC from the same energy spectrum$|Y(K)|^2$ as

$$\hat{e}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(K)|^2 . \hat{\Psi}_i(k) \quad (16)$$

DCT is taken on the log filter bank energies $\{log_{10}[\hat{e}(i)]\}_{i=1}^{Q}$ and the final inverted MFCC coefficient $\{\hat{C}_m\}_{m=1}^{R}$ can be written as

$$\hat{C}_m = \sqrt{\frac{2}{Q}} . \sum_{l=0}^{Q-1} log[\hat{e}(i+1)]\, cos\left[m.\left(\frac{2l-1}{2}\right)\frac{\pi}{Q}\right] \quad (17)$$

4.2 Inverted Mel-Frequency Cepstral Coefficients using Gaussian filters

It is expected that introduction of correlation between subband outputs in inverted mel-scaled filter bank makes it more complementary than what was realized using TF.
Flipping the original triangular filter bank, around 2 KHz inverts also the relation mentioned in eqn. (10) that gives

$$\left(\hat{k}_{bi} - \hat{k}_{bi-1}\right) > \left(\hat{k}_{bi+1} - \hat{k}_{bi}\right) \quad (18)$$

Here $\hat{k}_{bi}$ is the mean of the ith GF and standard deviation can be calculated as

$$\hat{\sigma}_i = \frac{\hat{k}_{bi} - \hat{k}_{bi-1}}{\alpha} \quad (19)$$

Here α value is chosen 2. The response of the GF for inverted MFCC filter bank and corresponding cepstral parameters can be calculated as follows;

$$\hat{\Psi}_i^{g_{IMFCC}} = e^{-\frac{(k-\hat{k}_{bi})^2}{2\hat{\sigma}_i^2}} \quad (20$$

$$\hat{e}^{g_{IMFCC}}(i) = \sum_{k=1}^{\frac{Ms}{2}} |Y(k)|^2 . \hat{\Psi}_i^{g_{IMFCC}}(k) \quad (21)$$

And

$$\hat{C}_m^{g_{IMFCC}} = \sqrt{\frac{2}{Q}} \sum_{l=0}^{Q-1} log[\hat{e}^{g_{IMFCC}}(i+1)]. cos\left[m.\left(\frac{2l-1}{2}\right).\frac{\pi}{Q}\right] \quad (22)$$

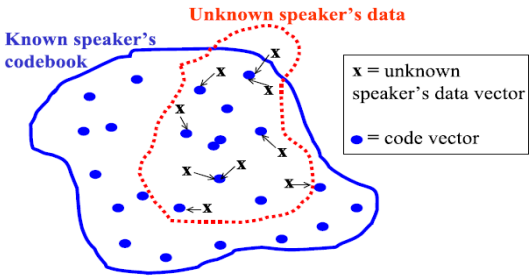

Fig 5: Illustration of VQ match score computation.

5. THEORETICAL BACKGROUND OF VQ

In VQ-based approach the speaker models are formed by clustering the speaker's feature vectors in K non-overlapping clusters. Each cluster is represented by a code vector ci, which is the centroid [20]. The resulting set of code vectors $\{C_1, C_2, C_3, C_4, , \ldots\ldots\ldots\ldots\ldots C_k\}$is called a codebook, and it serves as the model of the speaker. The model size (number of code vectors) is significantly smaller than the training set. The distribution of the code vectors follows the same underlying distribution as the training vectors. Thus, the codebook effectively reduces the amount of data by preserving the essential information of the original distribution. In LBG algorithms user must require the desired codebook size K, then it starts from an initial codebook of size K (usually, randomly selected vectors from the training set), which it iteratively refines in two successive steps until the codebook does not change [21]. The matching function in VQ-based speaker recognition is typically defined as the quantization distortion between two vector sets $X = \{x_1, \ldots\ldots\ldots\ldots, x_T\}$ and $C = \{c_1, \ldots\ldots\ldots\ldots, c_k\}$.Consider a feature vector xi generated by the unknown speaker, and a codebook C. The quantization distortion $d_q$of the vector $x_i$with respect to C is given by

$$d_q = (x_i, C) = \min_{cj \in C} d(x_i, c_j) \quad (23)$$

where $d(.,.)$ is a distance measure defined for the feature vectors. The code vectorc_{j*} for which $d(x_i, c_{j*})$ is minimum and the nearest neighbor of x_i in the codebook C. Most often, Euclidean or Euclidean squared distance measure is used due to the straightforward implementation and intuitive notion [22]. The average quantization distortion D_Q is defined as the average of the individual distortions:

$$D_Q(X,C) = \frac{1}{T}\sum_{i=1}^{T} d_q(x_i, C) \quad (24)$$

The eqn. (24) is not symmetrical, i.e. $D_Q(X,C) \neq D_Q(C,X)$.In this work we made an assumption that the first argument of D_Q is the sequence of the unknown speaker's feature vectors, and the second argument is a known speaker's codebook. The computation of the distortion is illustrated and shown in fig .5.

6. LOGICALLY MODULO-2 ADDITION (EX-OR) OF SPEAKER MODELS

As per the literature survey on speaker recognition, the logical combination of two or more classifiers would perform better if they would be processed with the information that is complementary in nature [23]. In this investigation we have done modulo 2 addition (EX-OR) operation on the MFCC and inverted MFCC which are complementary in nature. Here idea is to adapt the sheep and goat concept which are not in nature of sheep and wolf.

Two separate models have been developed for training phase that is MFCC and inverted MFCC, by using VQ technique [24]. During the test phase, MFCC and inverted MFCC features were extracted in a similar way from an incoming speech utterance as done in the training phase and were sent to their respective models. For each speaker, two scores were generated, one each from the MFCC and inverted MFCC models. Since modulo two adders rule outperforms other combination strategies due to its lesser sensitivity to estimation errors, a uniform weighted sum rule was adopted to combine the scores from the two classifiers [25].

If S_{MFCC} and S_{iMFCC} are the scores generated by the two models shown in fig. 6 for the ith speaker then the output score of modulo 2 adder out score S_{com} is expressed as [26]

$$S_{com} = wS_{MFCC} + (1-w)S_{IMFCC} \quad (25)$$

Modulo 2 output of parallel classifiers methodology via weighted sum rule the equation is expressed as below.

$$S_{com} = w\sum_{t=1}^{T} \log p(x_{tMFCC}|\lambda_{sMFCC}) + (1-w)\sum_{t=1}^{T} \log p\,(x_{tIMFCC}|\lambda_{sIMFCC}) \quad (26)$$

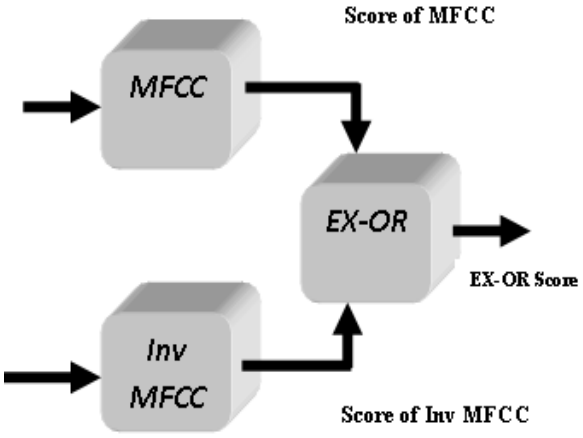

Fig. 6: Modulo 2 additions of MFCC and Inverted MFCC.

Where w=0.5, the matching of an unknown speaker is then performed by measuring the similarity/dissimilarity between the feature vectors of the unknown speaker to the models (codebooks) of the known speakers in the database.

Thus, the best matching codebook is now defined as the codebook that maximizes the similarity measure of the mapping $s: X \times C \rightarrow R$, i.e.:

$$C_{best} = \arg\max\{s(X, C_i)\} \quad (27)$$

Here the similarity measure is defined as the average of the inverse distance values:

$$s(X, C_i) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{d(x_t, c_{min}^{i,t})}, \quad (28)$$

Where $c_{min}^{i,t}$ denotes the nearest code vector to x_t in the code book C_i and $d: R^P \times R^P \rightarrow R$ is a given distance function in the feature space, whose selection depends of the properties of the feature vectors. If the distance function d satisfies $0 < d < \infty$ then s is a well-defined and $0 < s < \infty$. In the rest of the paper, we use Euclidean distance for simplicity. Note that in practice, we limit the distance values to the range $1 < d < \infty$ and, thus, the effective values of the similarity measure are$0 < s < 1$.

7. SPEAKER DISCRIMINATIVE MATCHING

Consider the example shown in fig. 7, in which the code vectors of two different speakers are marked by cross red and blue colors. There is also a set of vectors from an unknown speaker marked by stars. The region at the top rightmost corner cannot distinct the speakers from each other since it contains code vectors from all speakers. The region at the top leftmost corner is somewhat better in this sense because samples there indicate that the unknown speaker is not "triangle". The rest of the code vectors, on the other hand, have much higher discrimination power because they are isolated from the other code vectors [27]. The situation is not so evident if we use the unweighted similarity score of the eqn. (28). It gives equal weight to all sample vectors despite the fact that they do not have the same significance in the matching. Instead, the similarity value should depend on two separate factors: the distance to the nearest code vector, and the discrimination power of the code vector. Outliers and noise vectors that do not match well to any code vector should have small impact, but also vectors that match to code vectors of many speakers should have smaller impact on the matching score.

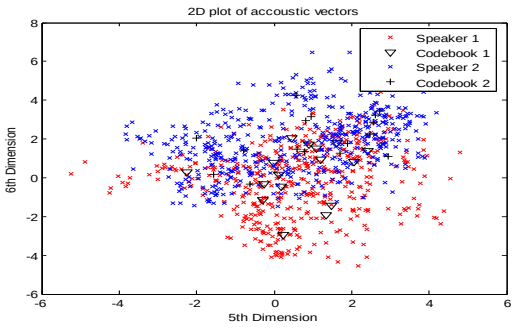

Fig 7: Illustration of code vectors having different discriminating power.

8. EXPERIMENTAL RESULTS

The voice corpus was collected in unhealthy environment by using Microsoft sound recorder. A good quality head phone was used to record the voice corpus. There were 70 speakers (61 males + 9 females) of twenty utterance of each using sampling rate of 8.0 kHz with 16 bits/sample. The average duration of the training samples was 6 seconds per speaker and out of twenty utterances one is used for training. For matching purposes remaining 19 voice corpus of the length 6 seconds, which was further divided into three different subsequences of the lengths 6 s (100%), 3 s (50%), 2s (33%) 1s (16 %) and 0.5s(8%) . Therefore, for 70 speakers we put 70X19X5 = 6650 utterance under test and evaluated the identification efficiency.

The identification rates are summarized through table 1.for the three different subsequences by varying the codebook sizes from K=1 to 256. It reaches 98.57% identification rates by taking training and testing voice corpus of 2 seconds. Even with a very short test sequence of 0.5 second the proposed method achieved identification rate of 91.42%.

Table 1. Summary of identification rate

Voice Sample	No Of Utterances	Correct Identification	Efficiency %
6 sec	6650	6650	100
3 sec	6650	6555	98.57
2 sec	6650	6550	98.49
1 sec	6650	6270	94.28
0.5 sec	6650	6080	91.42

9. CONCLUSION

We have proposed and evaluated a vector quantization approach for speaker recognition using MFCC and inverted MFCC for text independent speaker recognition. Experiments show that the method gives tremendous improvement and it can detect the correct speaker from much shorter (16% of training length, even 0.5sec of duration) speech samples. It is therefore well applicable in real-time systems. Furthermore, the method can be generalized to any other pattern recognition tasks because it is not designed for any particular features or distance metric.

10. REFERENCES

[1] D. Gatica-Perez, G. Lathoud, J.-M. Odobez and I. Mc Cowan. 2007 Audiovisual probabilistic tracking of multiple speakers in meetings, IEEE Transactions on Speech and Audio Processing, 15(2), pp. 601–616.

[2] J. P. Cambell, Jr. 1997 Speaker Recognition A Tutorial Proceedings of the IEEE, 85(9), pp. 1437-1462.

[3] Faundez-Zanuy M. and Monte-Moreno E. 2005 State-of-the-art in speaker recognition , Aerospace and Electronic Systems Magazine, IEEE, 20(5), pp. 7-12.

[4] K. Saeed and M. K. Nammous. 2005 Heuristic method of Arabic speech recognition, in *Proc. IEEE 7th Int. Conf. DSPA*, Moscow, Russia, pp. 528–530

[5] D. Olguin, P.A.Goor, and A. Pentland. 2009 Capturing individual and group behavior with wearable sensors, in Proceedings of AAAI Spring Symposium on Human Behavior Modeling.

[6] S. B. Davis and P. Mermelstein. 1980 Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Trans. On ASSP, 28(4), pp. 357-365.

[7] R. Vergin, B, O Shaughnessy and A. Farhat. 1999 Generalized Mel frequency Cepstral coefficients for large-vocabulary speaker independent continuous-speech recognition, IEEE Trans. On ASSP,7(5), pp. 525-532.

[8] Chakroborty, S., Roy, A. and Saha, G. 2007 Improved Closed set Text- Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks , International Journal of Signal Processing, 4(2), pp. 114-122.

[9] S.Singh and Dr. E.G Rajan. 2007 A Vector Quantization approach Using MFCC for Speaker Recognition, International conference Systemic, Cybernatics and Informatics ICSCI under the Aegis of Pentagram Research Centre Hyderabad, pp. 786-790.

[10] K. Sri Rama Murty and B. Yegnanarayana. 2006 Combining evidence from residual phase and MFCC features for speaker recognition, IEEE Signal Processing Letters, 13(1), pp. 52-55.

[11] Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and Gupta C. S. 2005 Combining evidence from source suprasegmental and spectral features for a fixed-text speaker verification system , IEEE Trans. Speech and Audio Processing, 13(4), pp. 575-582.

[12] J. Kittler, M. Hatef, R. Duin, J. Mataz. 1998 On combining classifiers, IEEE Trans, Pattern Anal. Mach. Intell, 20(3), pp. 226-239.

[13] He, J., Liu, L., Palm, G. 1999 A Discriminative Training Algorithm for VQ-based Speaker Identification , IEEE Transactions on Speech and Audio Processing, 7(3), pp. 353-356.

[14] Laurent Besacier and Jean-Francois Bonastre. 2000 Subband architecture for automatic speaker recognition, Signal Processing, 80, pp. 1245-1259.

[15] Zheng F., Zhang, G. and Song, Z. 2001 Comparison of different implementations of MFCC, J. Computer Science & Technology 16(6), pp. 582-589.

[16] Ganchev, T., Fakotakis, N., and Kokkinakis, G. 2005 Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, Proc. of SPECOM Patras, Greece, pp. 1191-194.

[17] Zhen B., Wu X., Liu Z., Chi H. 2000 On the use of band pass filtering in speaker recognition, Proc. 6th Int. Conf. of Spoken Lang. Processing (ICSLP), Beijing, China.

[18] S. Singh, Dr. E.G Rajan, P.Sivakumar, M.Bhoopathy and V.Subha. 2008 Text Dependent Speaker Recognition System in Presence Monitoring, International conference Systemic, Cybernatics and Informatics ICSCI -under the Aegis of Pentagram Research Centre Hyderabad, pp. 550-554.

[19] Kyung Y.J., Lee H.S. 1999 Bootstrap and aggregating VQ classifier for speaker recognition, Electronics Letters, 35(12), pp. 973-974.

[20] Y. Linde, A. Buzo, and R. M. Gray. 1980 An algorithm for vector quantizer design, IEEE Trans. Commun, 28(1), pp. 84-95.

[21] S.R. Mahadeva Prasanna, Cheedella S. Gupta, B. Yegnanarayana. 2006 Extraction of speaker-specific excitation information from linear prediction residual of speech, Speech Communication, 48(10), pp. 1243- 1261.

[22] Daniel J. Mashao, Marshalleno Skosan. 2006 Combining Classifier Decisions for Robust Speaker Identification, Pattern Recog,, 39, pp. 147-155.

[23] Ben Gold and Nelson Morgan. 2002 Speech and Audio Signal Processing, John Willy & Sons, Chap.14, pp. 189-203.

[24] A. Papoulis and S. U. Pillai. 2002 Probability, Random variables and Stochastic Processes, Tata McGraw-Hill Edition, Fourth Edition, Chap. 4, pp.72-122.

[25] Daniel Garcia-Romero, Julian Fierrez-Aguilar, Joaquin Gonzalez- Rodriguez, Javier Ortega-Garcia. 2006 Using quality measures for multilevel speaker recognition, Computer Speech and Language, 20(2), pp. 192-209.

[26] He J., Liu L., Palm G. 2008 A discriminative training algorithms for VQ based speaker identification, IEEE Transactions on Speech and Audio Processing, 7(3), pp. 353-356.

[27] Tomi Kinnunen and Pasi Franti. 2005 Speaker Discriminative Weighting Method for VQ-based Speaker identification, Macmillan Publishing Company, New York.