

Using Information Theoretic Vector Quantization for Inverted MFCC based Speaker Verification

Sheeraz Memon, Margaret Lech and Ling He

School of Electrical and Computer Engineering, RMIT University, Melbourne AUS.
sheeraz.memon@rmit.edu.au, margaret.lech@rmit.edu.au, ling.he@student.rmit.edu.au

Abstract— Over the recent years different versions the GMM classifier combined with the MFCC features have been established as speaker verification benchmarks. Although highly efficient, these systems suffer from computational complexity and occasional convergence problems. In this study a search of alternative classification and feature extraction methods of similar classification efficiency but overcoming some of the problems of the classical methods was undertaken. Preliminary results obtained for two different classification methods: the classical GMM and the ITVQ and three different feature extraction methods: MFCC, IMFCC and the MFCC/IMFCC fusion are presented. The ITVQ did not show better results compare to the classical GMM classifier, however the EER increase in case for the ITVQ was only by 0.2%. The best feature extraction method was proven to be the MFCC/IMFCC fusion. Both the MFCC/IMFCC fusion and the IMFCC outperformed the classical MFCC method.

Keywords—ITVQ, Information Theory, MFCC and IMFCC.

I. INTRODUCTION

SPEAKER verification system identifies a person by his or her voice. A typical speaker verification system consists of a feature extractor followed by a robust speaker modeling technique for generalized representation of extracted features. Feature selection is useful in speech [4] 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) [10] and it does produce good results in most of the situations. In other studies, feature extraction based on pitch or energy contours [7], glottal waveforms [8], or formant amplitude and frequency modulation [6] are proposed, and good performance has been shown.

In their recent research Sandipan et al. [2] 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.

Feature selection is followed by a classification algorithm to generate the speaker specific data, so far the GMM have been applied in the field of speaker verification and GMM has established very good results and are being currently used in a number of applications, however in this paper we bring a new idea of using information theoretic concepts in vector quantization and apply it to the speaker verification system as a classifier. The principle objective of carrying this experiment was to test the information theoretic vector quantization ITVQ [3] for fused Cepstral coefficients (MFCC and IMFCC).

Vector quantization (VQ) based speaker verification has remained a successful method [11], [12]. The basic idea in this approach is to compress a large number of short term spectral vectors into a small set of code vectors. The successful modeling of the underlying acoustic classes allows the Vector quantization system to achieve high recognition accuracy even with very short test utterances. Vector Quantization has been applied as a tool to classify the speaker models at several times, the use of LBG vector quantization algorithm to classify the speakers best fits the speaker verification models [11] other studies address that K-means and LBG can be applied to optimize the means and covariances for a GMM based classifier [12].

II. MFCC

The primary concern of describing the MFCC algorithm here is to clearly map the working of Inverted MFCC and later in this paper their fusion as a feature extraction set for 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 [13], human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the Mel scale [14] (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,

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

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, which is elaborated in fig.1 and 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,

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

Where, $k=1,2,\dots,M$ 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).

$$\phi_i(k) = \begin{cases} 0 & \text{for } k \leq k_{b_{i-1}} \\ \frac{k - k_{b_{i-1}}}{k_{b_i} - k_{b_{i-1}}} & \text{for } k_{b_{i-1}} \leq k \leq k_{b_i} \\ \frac{k_{b_{i+1}} - k}{k_{b_{i+1}} - k_{b_i}} & \text{for } k_{b_i} \leq k \leq k_{b_{i+1}} \\ 0 & \text{for } k \geq k_{b_{i+1}} \end{cases} \quad (3)$$

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 ($i=1,2,\dots, N_F$) are calculated as equally spaced points in the Mel scale using the following formula,

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

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 $E(i)$ ($i=1,2,\dots, N_F$) 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,

$$E(i) = \sum_{k=1}^{\frac{M_s}{2}} |Y(k)|^2 \Phi_i(k) \quad (5)$$

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 $E(i)$ ($i=1,2,\dots, N_F$) is calculated yielding the final set of the MFCC coefficients C_m , given as

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

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

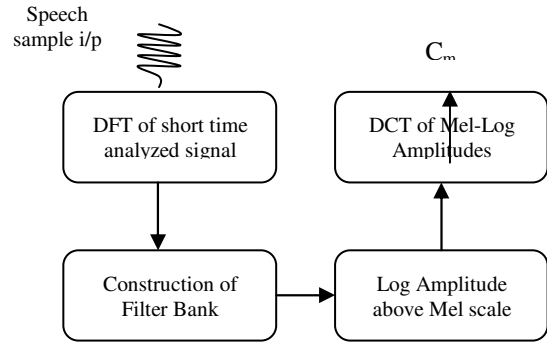


Fig. 1: Implementation structure of MFCC

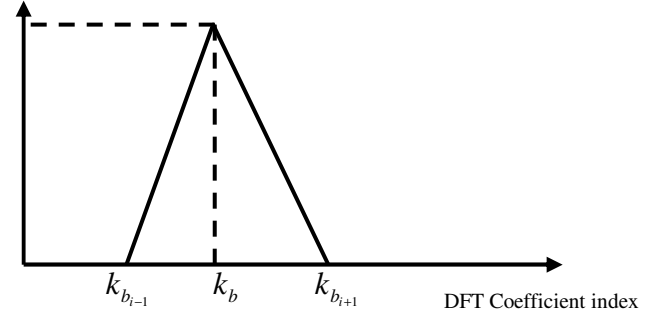


Fig.2. Response of a Mel scale Filter

III. IMFCC

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 [15]. The Inverted Mel Scale, which is shown as a dashed line in Fig.3, 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. $f_c = (f_{high} - f_{low})/2$).

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

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

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

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

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

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

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,

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

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

IV. INFORMATION THEORETIC VECTOR QUANTIZATION

The Vector Quantization methods are commonly used in the process of feature classification. The ITVQ [3] 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 [5].

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 [3].

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 [3]. The Parzen density estimator is given as,

$$p(x) = \frac{1}{N} \sum_{i=1}^N K(x - x_i) \quad (11)$$

Where $K(\cdot)$ 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 $M \ll N$.

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.

$$D_{c-s}(a(x), b(x)) = \quad (12)$$

$$\log \int a^2(x) dx - 2 \log \int a(x)b(x) dx + \log \int b^2(x) dx$$

Where, $a(x)$ and $b(x)$ denote the Parzen density estimates for the data and centroids, respectively.

The cost function in Eq. (12) 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.(12),

$\log \int a^2(x) dx$, 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 $2 \log$ 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. (12) 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 [3], a typical ITVQ algorithm makes use of an annealing procedure, which allows the algorithm to escape from local minima.

V. GAUSSIAN MIXTURE MODEL

The Gaussian Mixture Model (GMM) [1] 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 introduction of the adapted Gaussian mixture models [22] with the introduction of UBM-GMM and MAP-GMM algorithms have also proved as a milestone in the field of speaker verification. The introduction of these algorithms have increased the computational efficiency and strengthened the optimization process.

The probability density function (pdf) drawn from the GMM is a weighted sum of M component densities given as,

$$p(x|\lambda) = \sum_{i=1}^M p_i b_i(x) \quad (13)$$

Where x is a D -dimensional random vector, $b_i(x)$, $i=1,2,3,\dots,M$ are the component densities and p_i , $i=1,2,3,\dots,M$ are the mixture weights. Each component density is a D -variate Gaussian function of the form,

$$b_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\} \quad (14)$$

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

$$\lambda = \{ p_i, \mu_i, \Sigma_i \}, i = 1, \dots, M \quad (15)$$

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 optimal class models.

VI. SPEAKER VERIFICATION TESTS

The performance of the classical GMM classifier was compared with the Information Theoretic Vector Quantization (ITVQ) method. The classification was performed using three different feature extraction methods: the MFCC, the IMFCC and the fusion of the MFCC and the IMFCC (MFCC-IMFCC). The speaker verification tests results were obtained using two different data bases: YOHO and TIMIT.

A. Pre-Processing

The pre-processing included speech normalisation, pre-emphasis filtering and removal of the silence intervals.

The dynamic range of the speech amplitude was mapped into the interval from -1 to +1. The high-pass pre-emphasis filter was then applied to equalise the energy between the low and high frequency components of speech. The filter was given by the following difference equation: $y(k) = x(k) - 0.95x(k-1)$, where $x(k)$ denotes the input speech and $y(k)$ is the output

speech. The silence intervals were removed using a logarithmic technique for separating and segmenting speech from noisy background environments described in [9].

B. 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 [16], [17]. If the information supplied to the classifiers is complementary, such as the case of MFCC and IMFCC, the classification process could be largely improved [2],[23].

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,

$$D_m = \omega D_{MFCC} + (1 - \omega) D_{IMFCC} \quad (23)$$

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

$$m_{class} = \arg(\max_m D_m) \quad (24)$$

C. Test Results

The equal error rate (EER) [18] measure was used to evaluate the speaker verification performance. The verification results were obtained using three different feature extraction methods: MFCC, IMFCC and MFCC-IMFCC and two different classifiers: ITVQ and GMM. We are using 64 components for both GMM and ITVQ. We test the classification algorithm ITVQ for different feature extraction methods, it is clear from the results in Table I and Table II the fusion of MFCC and IMFCC enhances the speaker verification rate. We also compare the performance of ITVQ with the state of art GMM classification algorithm and it appears that ITVQ have comparable results to GMM.

TABLE I
THE EER VALUES FOR YOHO DATABASE

Classification Algorithm	MFCC	IMFCC	Fused MFCC and IMFCC
ITVQ	1.8%	2.0%	1.6%
GMM	1.6%	1.8%	1.4%

Table II shows the results of TIMIT database for speaker verification, it is evident from the results that the YOHO database presents a more realistic challenge to Speaker verification task than the TIMIT database. It may be because of the reason that for YOHO database the speech is recorded in a real office environment and the data is recorded for each speaker in different sessions increasing the intra-speaker variability.

The tests were performed on two databases: YOHO [20] and TIMIT [21]. The YOHO database was previously used in a

number of speaker verification tests [2],[19],[23]. It contains microphone speech of 130 speakers recorded in a real office environment during multiple sessions. The TIMIT database which contains clean speech recordings does not represent a substantial challenge for speaker verification systems and was only used to test the consistency of the classification results.

TABLE II
THE EER VALUES FOR TIMIT DATABASE

Classification Algorithm	MFCC	IMFCC	Fused MFCC and IMFCC
ITVQ	1.6%	1.8%	1.5%
GMM	1.5%	1.8%	1.4%

VII. CONCLUSION

A search for a feature extraction and classification scheme that gives results comparable with those of the GMM-MFCC was undertaken. The performance of the classical GMM classifier was compared with the recently introduced ITVQ method. The classification was performed using three different feature extraction methods: the MFCC, the IMFCC and the fusion of the MFCC and the IMFCC. The speaker verification results were obtained using two different data bases: YOHO and TIMIT. In all cases, the TIMIT data provided better classification results compare to the YOHO data, which was expected since the TIMIT data contains clean, noise-free speech made in laboratory conditions.

The fusion of MFCC an IMFCC clearly outperformed the MFCC and IMFCC feature extraction methods. ITVQ establishes comparable results with GMM, though GMM is a state-of-art-method for ASV and a number of systems are developed based on this algorithm but ITVQ could be used for the low-cost real time applications.

In general, the best classification results were obtained for the GMM classifier combined with the MFCC-IMFCC feature extraction.

These preliminary results are promising but further tests including better suited databases such as the latest NIST Speaker Recognition Evaluation (SRE) corpus and a wider range of alternative classification and feature extraction approaches are needed to draw more definite conclusions.

ACKNOWLEDGMENT

This research was partially supported by the Australian Research Council Linkage Grant LP0776235.

REFERENCES

- [1] J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.
- [2] Sandipan, C. and Ghoutam, S., "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter" *IJSP*, Vol. 5, No.1, 2008.
- [3] 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.
- [4] S. B. Davis and 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.
- [5] 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.
- [6] C. R. Jankowski jr, *et al.*, “Fine structure features for speaker identification,” in *Proc. ICASSP*, 1996, pp. 689–692.
- [7] B. Peskin *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.
- [8] M. D. Plumpe, 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.
- [9]. Lynch, Jr., Josenhans, J., Crochiere, R., “Speech/Silence segmentation for real-time coding via rule based adaptive endpoint detection”, In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp.1348-1351, 1987.
- [10]. D. A. Reynolds, “Experimental evaluation of features for robust speaker identification,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.
- [11]. H. Jialong, 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.
- [12]. G. Singh, 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.
- [13] D. O., Shaughnessy, *Speech Communication Human and Machine*, Addison-Wesley, New York, 1987.
- [14] Ben Gold and Nelson Morgan, *Speech and Audio Signal Processing*, Part- IV, Chap.14, pp. 189-203, John Willy & Sons ,2002.
- [15] B. Yegnanarayana, 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.
- [16] DJ. Mashao, M. Skosan, “Combining Classifier Decisions for Robust Speaker Identification”, *Pattern Recognition*, vol. 39, pp. 147-155, 2006.
- [17] KSR. Murty, B. Yegnanarayana, “Combining evidence from residual phase and MFCC features for speaker recognition”, *IEEE Signal Processing Letters*, vol 13, no. 1, pp. 52-55, Jan. 2006.
- [18] Cheng Min and Wang Hsiao-Chuan, “A method of estimating the equal error rate for automatic speaker verification”, *Chinese Spoken Language Processing. 2004 International Symposium on*, Publication Date: 15-18 Dec. 2004, page(s): 285- 288.
- [19] D. James, HP. Hutter and F. Bimbot, “The CAVE speaker verification project- Experiments on the YOHO and SESP corpora”, *Book Series Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, ISSN0302-9743 Volume 1206/1997.
- [20] YOHO Speaker Verification, *LDC Catalog No.:* LDC94S16, *ISBN:* 1-58563-042-X.
- [21] TIMIT Acoustic-Phonetic Continuous Speech Corpus, *LDC Catalog No.:* LDC93S1, *ISBN:* 1-58563-019-5.
- [22] D. Reynolds, T. Quatieri, and R. Dunn, ”Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol.10,pp. 19-41, 2000.
- [23] S. Chakroborty, 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.