Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization

Hossan, Md Afzal; Gregory, Mark

https://researchrepository.rmit.edu.au/esploro/outputs/9921858532701341/filesAndLinks?institution=61RMIT_INST&index=null

Hossan, M. A., & Gregory, M. (2013). Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization. *International Journal of Speech Technology*, 16, 103–113. <https://doi.org/10.1007/s10772-012-9166-0>

Published Version: <https://doi.org/10.1007/s10772-012-9166-0>

Thank you for downloading this document from the RMIT Research Repository.

The RMIT Research Repository is an open access database showcasing the research outputs of RMIT University researchers.

RMIT Research Repository: <http://researchbank.rmit.edu.au/>

Citation:

Hossan, M and Gregory, M 2013, 'Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization', International Journal of Speech Technology, vol. 16, pp. 103-113.

See this record in the RMIT Research Repository at:

<http://researchbank.rmit.edu.au/view/rmit:18569>

Version: Accepted Manuscript

Copyright Statement: © Springer Science+Business Media, LLC 2012

Link to Published Version:

<http://dx.doi.org/10.1007/s10772-012-9166-0>

Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization

M. Afzal Hossan · Mark A. Gregory

Abstract In this paper, a new and novel Automatic Speaker Recognition (ASR) system is presented. The new ASR system includes novel feature extraction and vector classification steps utilizing distributed Discrete Cosine Transform (DCT-II) based Mel Frequency Cepstral Coefficients (MFCC) and Fuzzy Vector Quantization (FVQ). The ASR algorithm utilizes an approach based on MFCC to identify dynamic features that are used for Speaker Recognition (SR). A series of experiments were performed utilizing three different feature extraction methods: (1) conventional MFCC; (2) Delta-Delta MFCC (DDMFCC); and (3) DCT-II based DDMFCC. The experiments were then expanded to include four classifiers: (1) FVQ; (2) K-means Vector Quantization (VQ); (3) Linde, Buzo and Gray VQ; and (4) Gaussian Mixed Model (GMM). The combination of DCT-II based MFCC, DMFCC and DDMFCC with FVQ was found to have the lowest Equal Error Rate for the VQ based classifiers. The results found were an improvement over previously reported non-GMM methods and approached the results achieved for the computationally expensive GMM based method. Speaker verification tests carried out highlighted the overall performance improvement for the new ASR system. The National Institute of Standards and Technology Speaker Recognition Evaluation corpora was used to provide speaker source data for the experiments.

Keywords Speaker recognition · Discrete cosine transform · Fuzzy vector quantization · K-Means,

Linde–Buzo–Gray · Mel frequency cepstral coefficients · Speech feature extraction

1 Introduction

Speaker Recognition (SR) is the identification of a speaker using properties that are extracted from an utterance. A typical Automatic Speaker Recognition (ASR) system consists of a feature extractor followed by a robust speaker modeling technique for generalized representation of the extracted features (Sahidullah and Saha 2009). Vocal tract information like formant frequency, bandwidth of formant frequency and other values may be linked to an individual person.

Conventionally SR can be classified into two categories: speaker identification and speaker verification. Speaker identification is defined as the process of determining which speaker makes an utterance. The speaker is registered into a database of speakers and utterances are added to the database that may be used at a later time during the speaker identification process. The acceptance or rejection of an identity claimed by a speaker is known as speaker verification. SR techniques can be organized into two categories: text-dependent and text-independent SR. A text-independent SR system is where the key feature of the system is speaker identification utilizing random utterance input speech. A text-dependent SR system is where recognition of the speaker's identity is based on a match with utterances made by the speaker previously and stored for later comparison. Phrases like passwords, card numbers, PIN codes, etc. made be used. In this research, the focus is on a series of text-independent speaker verification tests of a proposed feature extraction method and classification algorithm.

M.A. Hossan · M.A. Gregory (✉)
RMIT University, Melbourne, Australia
e-mail: mark.gregory@rmit.edu.au

M.A. Hossan
e-mail: s3248701@student.rmit.edu.au

The goal of a feature extraction block technique is to characterize the information (Keshet and Bengio 2009; Sahidullah and Saha 2009). A wide range of approaches may be used to parametrically represent the speech signal to be used in the SR activity (Saeidi et al. 2007; Zilca et al. 2003). Some of the techniques include: Linear Prediction Coding (LPC); Mel-Frequency Cepstral Coefficients (MFCC); Linear Predictive Cepstral Coefficients (LPCC); Perceptual Linear Prediction (PLP); and Neural Predictive Coding (NPC) (Salman et al. 2007; Paul et al. 2009; Charbuillet et al. 2007). MFCC is a popular technique because it is based on the known variation of the human ear's critical frequency bandwidth. MFCC coefficients can be obtained by de-correlating the output log energies of a filter bank which consists of triangular filters that are linearly spaced on the Mel frequency scale. Sahidullah and Saha (2009) introduced an implementation of Discrete Cosine Transform (DCT) known as Distributed DCT (DCT-II) to de-correlate speech. Sahidullah and Saha showed that DCT-II is a good approximation of the Karhunen-Loève Transform (KLT). MFCC data sets represent a melodic cepstral acoustic vector (Barbu 2009; Wang et al. 2006). The acoustic vectors can be used as feature vectors. It is possible to obtain more speech features by using a derivation of the MFCC acoustic vectors. The first order derivatives of the MFCC acoustic vectors are known as the delta MFCC (DMFCC). The second order derivatives of the MFCC values are known as the delta-delta MFCC (DDMFCC) and can be derived from the DMFCC. The research reported in this paper includes work carried out to incorporate DCT-II based MFCC with DMFCC and DDMFCC to identify speech features.

Pattern recognition is an important activity carried out in most SR systems. In the SR verification stage, a pattern recognition algorithm is applied to carry out the recognition step. Hidden Markov Method (HMM), Gaussian Mixture Model (GMM) and Vector Quantization (VQ) are well known algorithms used in the speaker verification step. In the research into SR VQ was used as a classifier, because VQ is computationally efficient and less complex than other algorithms. There are several clustering algorithms based on VQ, such as Linde, Buzo and Gray (LBG), K-means and Fuzzy C-means. Fuzzy Vector Quantization (FVQ) was selected as an improved VQ based technique that can be used for speaker modeling. Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership (Atal and Hanauer 1971). In many situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several clusters are not forced to fully belong to one of the clusters, but may be assigned membership degrees between 0 and 1 indicating partial membership of more than one cluster.

A set of SR tests were performed utilizing the NIST SR database (2004) and the results are presented. The performance of the feature extraction methods and classifiers were evaluated by comparing the SR test results.

The rest of this paper is organized as follows. Section 2 presents a review of recent research into feature extraction utilizing DCT-II and MFCC. In Sect. 3 the MFCC analysis methodology is reviewed followed by the proposed feature extraction technique. Section 4 presents the use of VQ based on different clustering algorithms. In Sect. 5, the experimental arrangements for the proposed experiments and outcomes are provided. Finally the conclusion is provided in Sect. 6.

2 Recent developments in speaker recognition

ASR is an important machine human interface process and considerable research has occurred over the past three decades. Feature extraction is a very important step in the SR process and important incremental improvements in feature extraction outcomes has occurred over time. Historically, the following spectrum-related speech feature extraction techniques have dominated the SR process: (1) Real Cepstral Coefficients (RCC) introduced by Oppenheim (1969); (2) LPC proposed by Atal and Hanauer (1971); (3) LPCC derived by Atal; and (4) MFCC introduced by Davis and Mermelstein (1980). Other speech feature extraction techniques were also developed including: (1) PLP coefficients by Hermansky (1990); (2) Adaptive Component Weighting (ACW) cepstral coefficients by Assaleh and Mammone (1994); and (3) various wavelet-based feature extraction techniques that, whilst presenting reasonable solutions for the same task, did not gain widespread practical use. The reasons why some of the approaches developed were not utilized in practice include computation requirements or the failure of an approach to provide a significant advantage when compared to MFCC (Ganchev 2005). Among the feature extraction techniques, MFCC was found to be very efficient because MFCC follows auditory perception (Wei-Guo et al. 2008). Recent contributions to the further development of the MFCC feature extraction technique include the work by Chen et al. (2008) who derived differential MFCCs, which improved the performance of SR, and Sahidullah and Saha (2009) who proposed extracting the MFCC features by using DCT-II, which significantly improved feature extraction performance. Jayanna and Prasanna (2008) and Memon et al. (2009b), used the first order and second order derivatives of MFCC to capture transitional characteristics.

Feature matching is a very important stage in speech based pattern recognition. The classification algorithm used is critical to the development of effective and efficient speaker models. VQ is a classical technique used for pattern recognition because it is computationally very efficient

and the performance results achieved are satisfactory. To derive the speaker models and carry out classification there are several clustering techniques used that were described in the previous section. Among the techniques presented Fuzzy C-means clustering is very effective and efficient (Jayanna and Prasanna 2008), yet not as good as GMM. The GMM classification algorithm is now a widely used classification technique and the performance of the GMM classification algorithm is better than other classification algorithms (Wang et al. 2009; Memon et al. 2009b). However, GMM is very complex and computationally intensive and this is the motivation for research into a less computationally intensive VQ based classification algorithm that might more closely approximate outcomes achieved when using GMM.

3 MFCC feature extraction method

3.1 Conventional method

Psychophysical studies have shown that human perception of speech signal sound frequency contents follows a non-linear scale known as the ‘Mel’ scale (Memon et al. 2009a; Sahidullah and Saha 2009). Speech signal tones are represented as a frequency measured in Hz as defined in (1).

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

where f_{mel} is the subjective pitch in Mels corresponding to a frequency in Hz.

MFCC provides a baseline acoustic feature set for speech and SR applications (Sahidullah and Saha 2009; Kim and Eriksson 2004). MFCC coefficients are a set of DCT de-correlated parameters, which are computed through a transformation of the logarithmically compressed filter-output energies (Ganchev 2005), derived through a perceptually spaced triangular filter bank that processes the Discrete Fourier Transformed (DFT) speech signal.

An N -point DFT of the discrete input signal $y(n)$ is defined in (2).

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

where $1 \leq k \leq M$. Next, the filter bank, which has linearly spaced filters in the Mel scale, is imposed on the spectrum. The filter response $\psi_i(k)$ of filter i in the filter bank is defined in (3).

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

If Q denotes the number of filters in the filter bank, then $\{k_{b_i}\}_{i=0}^{Q+1}$ are the boundary points of the filters and k denotes the coefficient index in the M s-point DFT. The boundary points for each filter i ($i = 1, 2, \dots, Q$) are calculated as equally spaced points in the Mel scale using (4).

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

where f_s is the sampling frequency in Hz and f_{low} and f_{high} are the low and high frequency boundaries of the filter bank, respectively. f_{mel}^{-1} is the inverse of the transformation shown in (1) and is defined in (5).

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

In the next step, the output energies $e(i)$ ($i = 1, 2, \dots, Q$) of the Mel-scaled band-pass filters are calculated as a sum of the signal energies $|Y(k)|^2$ falling into a given Mel frequency band weighted by the corresponding frequency response $\psi_i(k)$ as shown in (6).

$$e(i) = \sum_K^M |y(k)|^2 \psi_i(k) \quad (6)$$

Finally, the DCT-II is applied to the log filter bank energies $\{\log[e(i)]\}_{i=0}^Q$ to de-correlate the energies and the final MFCC coefficients C_m are provided in (7).

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

where $m = 0, 1, 2, \dots, R - 1$, and R is the desired number of MFCCs.

3.2 Dynamic speech features

The speech features, which are the time derivatives of the spectrum-based speech features, are known as dynamic speech features. Memon et al. (2009b) showed that system performance may be enhanced by adding time derivatives to the static speech parameters (Jayanna and Prasanna 2008). The first order derivatives, referred to as delta features, can be calculated as shown in (8).

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{i+\theta} - c_{i-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (8)$$

where d_t is the delta coefficient at time t , computed in terms of the corresponding static coefficients $c_{t-\theta}$ to $c_{t+\theta}$ and Θ is the delta window size. The delta and delta-delta cepstra are evaluated based on MFCC (Chen et al. 2001a, 2001b, 2008).

3.3 DCT-II

In Sect. 3.1, DCT is used which is an optimal transformation for de-correlating the speech features (Sahidullah and Saha 2009). This transformation is an approximation of KLT for the first order Markov process.

The correlation matrix for a first order Markov source is defined in (9).

$$C = \begin{bmatrix} 1 & \rho & \rho^2 & \cdot & \cdot & \rho^{N-1} \\ \rho & 1 & \rho & \rho^2 & \cdot & \cdot \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \cdot \\ \cdot & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \cdot & \cdot & \rho^2 & \rho & 1 & \cdot \\ \rho^{N-1} & \cdot & \cdot & \rho^2 & \rho & 1 \end{bmatrix} \quad (9)$$

where ρ is the inter element correlation ($0 \leq \rho \leq 1$). Sahidullah and Saha showed that for the limiting case where $\rho \rightarrow 1$, the Eigen vector of (8) can be approximated as shown in (10) (Sahidullah and Saha 2009).

$$k(nt) = \sqrt{\frac{2}{N}} \cos \left[\frac{n\pi(2t+1)}{N} \right] \quad (10)$$

where $0 \leq t \leq N-1$ and $0 \leq n \leq N-1$. Equation (9) is the DCT Eigen function. DCT is used rather than the signal dependent optimal KLT transformation because it is possible to closely approximate the DCT Eigen function. But in reality the value of ρ is not 1. In the filter bank structure of the MFCC, filters are placed along the Mel-frequency scale. As the adjacent filters have an overlapping region, the neighboring filters contain more correlated information than filters further away. Filter energies have various degrees of correlation (not holding to a first order Markov correlation). Applying a DCT to the entire log-energy vector is not suitable as there is non-uniform correlation among the filter bank outputs (Kanade and Hall 2007). It is proposed to use DCT in a distributed manner to follow the Markov property more closely. The array $\{\log[e(i)]\}_{i=1}^Q$ is subdivided into two parts (analytically this is optimum) which are SEG#1 $\{\log[e(i)]\}_{i=1}^{\lfloor Q/2 \rfloor}$ and SEG#2 $\{\log[e(i)]\}_{i=\lfloor Q/2 \rfloor+1}^Q$.

3.4 Algorithm for DCT-II based DDMFCC

Unlike conventional methods, during this research DCT was applied in a distributed manner to compute MFCCs, which are then concatenated to present the final set of MFCCs. Based on the final set of MFCCs the dynamic features were computed. By utilizing this novel and new feature extraction technique it is possible to identify de-correlated features that provide a more distinct set of features for a given speech utterance and the result is a significant improvement in the SR system performance.

The algorithm steps include:

1: **if** $Q = \text{EVEN}$ **then**

2: $P = Q/2$;

3: **PERFORM** DCT of $\{\log[e(i)]\}_{i=1}^{\lfloor Q/2 \rfloor}$ to get $\{C_m\}_{m=0}^{P-1}$;

4: **PERFORM** DCT of $\{\log[e(i)]\}_{i=P+1}^Q$ to get $\{C_m\}_{m=P}^{Q-1}$;

5: **else**

6: $P = \lfloor \frac{Q}{2} \rfloor$;

7: **PERFORM** DCT of $\{\log[e(i)]\}_{i=1}^{\lfloor Q/2 \rfloor}$ to get $\{C_m\}_{m=0}^{P-1}$;

8: **PERFORM** DC of $\{\log[e(i)]\}_{i=P+1}^Q$ to get $\{C_m\}_{m=P}^{Q-1}$;

9: **end if**

10: **DISCARD** C_0 and C_P ;

11: **CONCATENATE** $\{C_m\}_{m=1}^{P-1}$ & $\{C_m\}_{m=P+1}^{Q-1}$ to form feature vector $\{Cep(i)\}_{i=1}^{P-2}$;

12: **CALCULATE** DMFCC;

13: **CALCULATE** DDMFCC;

14: **CONCATENATE** $\{Cep(i)\}_{i=1}^{P-2}$, DMFCC and DDMFCC to form final feature vector.

In the proposed algorithm the number of feature vectors is reduced by 1 for the same number of filters when compared to conventional MFCC. For example if 20 filters are used to extract MFCC features, then 19 coefficients are used and the first coefficient is discarded. In the proposed method 18 coefficients are sufficient to represent the 20 filter bank energies as the other two coefficients represent the signal energy. A window size of 12 is used and the delta and acceleration (double delta) features are evaluated utilizing the MFCC.

4 Classifiers

4.1 Vector quantization

In brief, VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by the cluster's center. The cluster's center may be represented by the coordinates of the center in the vector space and this representation is known as a codeword. The collection of all codewords is called a codebook and each codeword has an index in the codebook. Clustering applications cover several fields such as audio and video data compression, pattern recognition, computer vision, medical image recognition, etc.

The objective of VQ is the representation of a set of feature vectors $x \in X \subseteq \mathbb{R}^k$ by a set $y = \{y_1, y_2, \dots, y_{N_c}\}$, of N_c reference vectors in $\mathbb{R}^k$. Y is called the codebook and its elements codewords. The vectors of X are called also input patterns or input vectors. So, VQ can be represented as a function: $q : X \rightarrow Y$. The knowledge of q permits us to obtain a partition S of X constituted by the N_c subsets S_i (called cells) as shown in (11).

$$S_i = \{x \in X : q(x) = y_i\}, \quad \text{where } i = 1, 2, \dots, N_c \quad (11)$$

4.2 Clustering techniques

4.2.1 Linde, Buzo and Gray clustering technique

The acoustic vectors extracted from a speaker utterance provide a set of training vectors. The next step is to build a speaker-specific VQ codebook using the training vectors. The LBG algorithm is a suitable starting point for this activity, clustering a set of L training vectors into a set of M codebook vectors. The LBG VQ algorithm is an iterative algorithm which alternatively solves the two optimality criteria. The algorithm requires an initial codebook $c^{(0)}$. This initial codebook is obtained by the *splitting* method. In this method, an initial codevector is set as the average of the entire training sequence. This codevector is then split into two codevectors. The iterative algorithm is run with these two vectors as the initial codebook. The two codevectors are split into four and the process is repeated until the desired number of codevectors is obtained.

4.2.2 K-Means clustering technique

The standard K-means algorithm is a clustering algorithm used in data mining and which is widely used for clustering large sets of data. In 1967, MacQueen proposed the K-means algorithm and it was a simple, non-supervised learning algorithm, which was applied to solve the problem of the well-known cluster (Shi et al. 2010). It is a partitioning clustering algorithm and this method is used to classify the given data objects into k different clusters iteratively, converging to a local minimum. So the results of generated clusters are compact and independent.

The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance. The next phase is to arrange each data object with the nearest center. Euclidean distance is generally used to determine the distance between each data object and the cluster centers. When all the data objects are included in a cluster, the first step is completed and an early grouping is done. This process is repeated continuously until the criterion function becomes the minimum. Supposing that the target object is x , x_i indicates the average of cluster C_i . The criterion function is defined in (12).

$$E = \sum_{i=1}^K \sum_{x \in C_i} |x - x_i|^2 \quad (12)$$

E is the sum of the squared error of all objects in database. The distance of the criterion function is Euclidean distance, which is used for determining the nearest distance between each data object and cluster center. The Euclidean distance between one vector $x = (x_1, x_2, \dots, x_n)$ and another vec-

tor $y = (y_1, y_2, \dots, y_n)$. The Euclidean distance $d = (x_i, y_i)$ can be obtained as shown in (13).

$$d(x_i, y_i) = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2} \quad (13)$$

4.2.3 Fuzzy C-means clustering

In speech-based pattern recognition, VQ is a widely used feature modeling and classification algorithm, since it is simple and computationally very efficient technique. FVQ reduces disadvantages of classical VQ. Unlike LBG and K-means algorithms, the FVQ technique follows the principle that a feature vector located between the clusters should not be assigned to only one cluster. Therefore, in FVQ each feature vector has an association with all clusters (Jayanna and Prasanna 2008). The discrete nature of hard partitioning also causes analytical and algorithmic intractability of algorithms based on analytic function values, since the function values are not differentiable (Atal and Hanauer 1971).

Fuzzy c -means is a clustering technique that permits one piece of data to belong to more than one cluster at the same time. It aims at minimizing the objective function defined by (14) (Abida 2007).

$$J = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m (\|x_i^{(j)} - c_j\|)^2, \quad 1 < m < \infty \quad (14)$$

u_{ij} where C is the number of clusters, N is the number of data elements, x_i is a column vector of X , and is defined as the centroid of the i th cluster. u_{ij} is an element of U , and denotes the membership of data element j to cluster i is subject to the constraints $u_{ij} \in [0, 1]$ and $\sum_{i=1}^C u_{ij} = 1$ for all j . m is a free parameter which plays a central role in adjusting the blending degree of different clusters. If m is set to 0, J is a sum-of-squared error criterion, and u_{ij} becomes a Boolean membership value (either 0 or 1). $\|\cdot\|$ can be any norm expressing similarity (Wang et al. 2009).

Fuzzy partitioning is carried out using an iterative optimization of the objective function shown in (11), with the update of membership function u_{ij} , an element of U , which denotes the membership of data element j to cluster i . The cluster center c_j is derived using (15) and (16).

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i^{(j)} - c_j\|}{\|x_i^{(j)} - c_{jk}\|} \right)^{\frac{2}{m-1}}} \quad (15)$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m x_i}{\sum_{i=1}^N u_{ij}^m} \quad (16)$$

This iteration will stop when $\max_{ij} \{|u_{ij}^{k+1} - u_{ij}^k|\} < \epsilon$, where ϵ is the termination criterion.

The algorithm for Fuzzy c -means clustering includes the steps (Abida 2007):

1. **Initialize** C, N, m, U
2. **repeat**
3. **minimize** j , by computing:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i^{(j)} - c_j\|}{\|x_i^{(j)} - c_{jk}\|} \right)^{\frac{2}{m-1}}}$$

4. **normalize** u_{ij} by $\sum_{i=1}^C u_{ij} = 1$
5. **compute** centroid c_j

$$c_j = \frac{\sum_i^N u_{ij}^m x_i}{\sum_i^N u_{ij}^m}$$

6. **until** slightly change in U and V
7. **end**

4.3 Gaussian mixture model (GMM)

The GMM 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 (Memon et al. 2009a; Chen et al. 2001a, 2001b). Adapted GMMs known as UBM-GMM and MAP-GMM have further enhanced speaker verification outcomes (Memon et al. 2009b; Wang et al. 2009). The introduction of the adapted GMM algorithms has increased computational efficiency and strengthened the speaker verification optimization process.

The Probability Density Function (PDF) drawn from the GMM is a weighted sum of M component densities as described in (17).

$$p(x|y) = \sum_{i=1}^M p_i b_i(x) \quad (17)$$

where x is a D-dimensional random vector, $b_i(x)$, the component densities are $i = 1, 2, 3, \dots, M$ and the mixture weights are p_i , for $i = 1, 2, 3, \dots, M$. Each component density is a D-variate Gaussian function of the form shown in (18).

$$b_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i) \Sigma_i^{-1} (x - \mu_i) \right\} \quad (18)$$

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 as shown in (19).

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, \dots, M \quad (19)$$

Each class is represented by a mixture model and is referred to by the class model λ . The Expectation Maximization (EM) algorithm is most commonly used to iteratively derive optimal class models.

5 Proposed feature extraction method

In this SR method, a VQ algorithm along with Fuzzy C-means clustering technique was used and the performance found was compared with GMM, the state of the art technique for SR systems. The principle reason to select VQ and Fuzzy C-means was to identify a mathematically simple and computationally efficient algorithm when compared to traditional approaches that include the computationally expensive GMM. The research objective was to identify a new, novel and computationally efficient SR method that approached the performance achieved when utilizing GMM.

5.1 Pre-processing

The pre-processing stage includes speech normalization, pre-emphasis filtering and removal of silence intervals. The dynamic range of the speech amplitude is mapped into the interval from -1 to $+1$. A high-pass pre-emphasis filter can then be applied to equalize the energy between the low and high frequency speech components. The filter is given by the 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 can be removed using a logarithmic technique for separating and segmenting speech from noisy background environments.

5.2 Experiment speech databases

The annually produced NIST speaker recognition evaluation (SRE) database has become the state of the art corpora for evaluating methods used or proposed for use in the field of speaker recognition. VQ-based systems have been widely tested on NIST SRE. The research evaluated the proposed algorithm utilizing two datasets: a subset of the NIST SRE-02 (Switchboard-II phase 2 and 3) data set and the NIST SRE-04 data (Linguistic data consortium's Mixer project). The background training set consisted of 1225 conversation sides from Switchboard-II and Fisher. For the purpose of the research the data in the background model did not occur in the test sets and did not share speakers with any of the test sets. Data sets that have duplicate speakers have been removed. The speaker recognition experiment test setup also included a check to ensure that the test or background data sets were used in training or tuning the speaker recognition system.

5.3 System architecture

The SR system used in the research is presented in Fig. 1. The training and testing steps are shown and how the classifier utilizes the trained data set. The training and testing steps include signal pre-processing to remove noise and

Fig. 1 Speaker Recognition System

clean up the signal prior to the next stage of training and testing. The next stage includes the classifier which involves the feature extraction and classification utilizing VQ. The resulting vectors are models within the training steps and in the test steps the resulting vector is compared with the trained codebook to identify if there is a match or not.

5.4 Speaker recognition experiments and results

5.4.1 Equal Error Rate (EER)

The error rates for a SR system were initially measured using Receiver Operating Characteristic (ROC) curves. However in the more recent studies of SR systems, the nonlinear ROC curves are replaced by Detection Error Trade-off (DET) plots, which are believed to provide a more efficient representation of the system performance because of their linear behavior in the logarithmic coordinate system. In this research DET plots were used to evaluate the performance of SR systems. The DET plots are related to the EER parameter representing a normalized measure of the system error rates (Hossan 2011).

The DET plot is a curve representing the percentage of the false rejection probability as a function of the percentage of the false acceptance probability. An example of a DET plot is shown in Fig. 2. As illustrated in Fig. 2, the false

Fig. 2 An example of the Detection Error Trade off curve and the Process of determining the Equal Error Rates

rejection probability is an inverse proportion to the false acceptance probability. Which means that, by decreasing the false rejection probability the false acceptance probability will be increased and vice versa. Since the ultimate goal of all speaker verification is to simultaneously minimize both errors (false rejection and false acceptance), the best compromise can be achieved when both errors are equal. The value of the percentage of the false rejection (or false acceptance) at the point when these two errors are equal is the EER. As illustrated in Fig. 2, the EER can be determined graphically as the percentage of false rejection (or false acceptance) at the intersection point between a 45° line and the DET curve. The smaller the EER for a given speaker verification system, the better is the overall system performance. So small change in the EER value carries a significant value.

In an ASV system, the EER is a measure calculated to evaluate the system performance. The EER was defined for each system using the Detection Error Trade-off (DET) curve. Usually a large number of test samples are required to calculate the EER accurately. The ASV stage is the task of verifying the speaker's identity from the speaker's voice. In a conventional ASV system the decision rule for acceptance or rejection is based on the score of a test utterance and a predefined threshold (Cheng and Wang 2004). There are two phases for an ASV system to accomplish this task. In the training phase, the ASV learns each client's voice features from the training utterances to generate a statistical model. Also, a single speaker independent model called the world model or the Universal Background Model (UBM) is generated. In the test phase, the ASV system analyzes an incoming utterance, and then uses the claimed speaker model and the UBM to compute a LLR score. The score is then compared with a pre-set threshold to determine whether the speaker is accepted or rejected. Cheng and Wang (2004), proposed an EER estimation method to manipulate the speaker models of the client speakers and the imposters so that the distribution of the computed likelihood scores is closer to the distribution of likelihood scores

Fig. 3 Classifier performance

obtained from test samples. A more reliable EER can then be calculated by the speaker models. In this research the method presented by Cheng and Wang was used to compute the EER for the GMM classifier.

5.4.2 Experimental results

Matlab R2008b (2012) was used to simulate the new ASV system. The NIST SRE (2004) provided a data set that could be used to evaluate the proposed ASV system performance. In the new algorithm presented in this paper the number of feature vectors is reduced by 1 for the same number of filters when compared to conventional MFCC. For example if 20 filters are used to extract MFCC features, then 19 coefficients are used and the first coefficient is discarded. In the proposed method 18 coefficients are sufficient to represent the 20 filter bank energies as the other two coefficients represent the signal energy (Hossan et al. 2010). A window size of 12 is used and the delta and acceleration (double delta) features are evaluated utilizing the MFCC methodology. The proposed method has concatenated DCT-II based MFCC (12), DMFCC (12) and DDMFCC (12). In this research, VQ is used for verification and classification. To perform the verification task the VQ needs to be trained. In this experiment 120 and 60 second utterances of 128 speakers were used for training and testing respectively. For FVQ, K-means VQ and LBG VQ, a codebook of size 256 was used for MFCCs, DCT-II based MFCCs, DMFCCs and DDMFCCs to build speaker models. The feature vectors of the test speech data were then compared with the codebooks for the different speakers to identify the most likely speaker of the test speech signal.

The EER experimental results for five feature extraction methods and four classifiers are shown in Fig. 3. The result of proposed feature extraction method, with different classifiers, is presented in Fig. 4. The results show the overall performance for the new and novel DCT-II based MFCC (12) + DMFCC (12) + DDMFCC (12) feature extraction method is an improvement over the methods reported in the

literature and used within this study. Figure 4 provides DET plots for the five feature extraction methods studied and used with the four classifier approaches. The overall results highlight that the performance of the new algorithm presented in this research, DCT-II based MFCC (12) + DMFCC (12) + DDMFCC (12) coupled with a FVQ classifier provides better results than that found for the other VQ classifiers and the performance results found for the new algorithm coupled with a GMM classifier were the best overall.

In the experiments carried out, the performance of SR is evaluated based on different MFCC feature extraction methods using GMM and VQ classifiers based on different clustering algorithms. For the experiments 128 speakers were used with 120 and 60 second utterances of each speaker used for training and testing respectively. For FVQ, K-means VQ and LBG VQ, a codebook of size 256 was used for each of MFCC, DCT-II based MFCC, DMFCC and DDMFCC to build speaker models. The feature vectors of the test speech data were then compared with the codebooks for the different speakers to identify the most likely speaker of the test speech signal. GMM models with 128 mixtures and features of various sizes and types including individual and fused were used (Shi et al. 2010). The delta and double delta features are derived based on MFCC coefficients since it was found that they have a performance supremacy over MFCC. In the proposed algorithm concatenated DCT-II based MFCC (12), DMFCC (12) and DDMFCC (12) was used and this stage was incorporated into a complete ASR system. The results based on various combinations of feature extraction techniques and classifiers are listed in Table 1.

The results obtained can be summarized as follows:

- The performance of the MFCC feature extraction method improved when used with DCT-II.
- The fusion of MFCC and DMFCC didn't enhance the performance significantly.
- The fusion of DCT-II based DDMFCC outperformed the other feature extraction techniques.
- The performance of FVQ is superior to other VQ techniques.
- It is proved that the performance of widely used GMM is better than any other classification algorithm.

From Fig. 4, the performance of the proposed SR system is shown to be an improvement over the other systems studied. The proposed feature extraction algorithm outperformed the feature extraction methods reported in the literature.

The proposed technique used the DCT in distributed manner to get more de-correlated features and derived dynamic features to capture the transitional characteristics. This is a new and novel step and the performance results

Table 1 The EER (%) values of different feature extraction methods based on different VQ techniques

No.	Feature extraction method	K-Means VQ	LBG VQ	FVQ
1	MFCC	12.68	12.17	11.28
2	DCT-II based MFCC	12.11	11.70	10.76
3	MFCC + DMFCC	11.18	10.97	10.24
4	MFCC + DMFCC + DDMFCC	10.94	10.57	9.96
5	DCT-II based MFCC + DMFCC + DDMFCC	10.35	9.97	8.94

highlight the value of the research carried out. The concatenation of DCT-II based MFCCs and dynamic feature extraction has enhanced the overall performance outcomes because the algorithm captured more specific information from the speech utterance than other feature extraction techniques currently do.

In this research, fuzzy *c*-means clustering algorithm is used. The working principle of FVQ is different from K-means VQ and LBG VQ, in the sense that the soft decision making process is used while designing the codebooks in FVQ (Jayanna and Prasanna 2008), whereas in K-means VQ and LBG VQ the hard decision process is used. Moreover, in K-means VQ and LBG VQ each feature vector has an association with only one of the clusters. It may be difficult to come to a conclusion that the feature vector belongs to a particular cluster. Whereas, in FVQ each feature vector has an association with all of the clusters with certain degrees of association, dictated by the membership function. In FVQ all of the feature vectors are associated with all of the clusters and therefore there are relatively more feature vectors within each cluster and hence the representative vectors i.e. the code vectors may be more reliable than for the other VQ techniques. Therefore, clustering may be found to be improved when using FVQ in some situations and the use of FVQ in ASR systems has been shown to be an improvement over other VQ techniques. Even though the GMM approach performed better than FVQ at the verification stage, FVQ is a simple and computationally very efficient technique and therefore has advantages over using GMM in a commercial system (Fig. 4).

6 Conclusion

The research objective was to carry out a detailed study of current ASR implementations and to identify an approach that could be implemented in one of the ASR steps that would provide an overall improvement in the system performance without significantly increasing complexity. The research carried out identified a new and novel algorithm for the feature extraction step utilizing DCT-II and DDMFCC and an improvement in the classifier operation by utilizing FVQ rather than other VQ techniques. The use of FVQ was

Fig. 4 DET plots of (a) MFCC (12) (b) DCT-II based MFCC (12) (c) MFCC (12) + DMFCC (12) (d) MFCC (12) + DMFCC (12) + DDMFCC (12) (e) DCT-II based MFCC (12) + DMFCC (12) + DDMFCC (12)

compared with GMM and FVQ was found to be an improvement over other VQ techniques when compared to the use of GMM. The research objectives were successfully achieved.

Fig. 4 (Continued)

In this paper, a new approach for speech feature extraction and classification utilizing DCT-II based DDMFCC and FVQ has been presented. The outcomes achieved by utilizing the ASR system with the NIST speech database were an improved SR performance for VQ based classification systems. The research outcomes were compared with previous MFCC feature extraction techniques and the results presented.

Initially the research concentrated on comparison with the performance of conventional MFCC. The test system was refined with the use of DCT-II in the de-correlation process and additional simulation results have been presented. The research highlighted that correlation among the filter bank outputs can't effectively be removed by applying conventional DCT to all of the signal energies at the same time. It was found that the use of DCT-II and DDMFCC in the new approach improved performance in terms of identification accuracy together with a lower number of features being used and as a consequence the new approach reduced computational time.

The MFCC feature vectors that were extracted did not accurately capture the transitional characteristics of the speech signal which contains the speaker specific information (Jayanna and Prasanna 2008). The transitional speech characteristics were found by computing DMFCC and DDMFCC, respectively the first-order and second-order time-derivative of the MFCC.

The new approach presented in this paper includes feature extraction utilizing DCT-II for de-correlation prior to utilizing the FVQ technique for the classification step in the speaker modeling. The results presented for the FVQ technique were an improvement over previous VQ techniques found in the literature and the results showed the proposed system performance was approaching that achieved by utilizing the computationally expensive GMM in the speaker modeling.

This new approach for feature extraction and classification utilizing DCT-II and DDMFCC with a FVQ classifier is promising and provided an overall improvement in SR performance for VQ based speaker modeling. In future research, evaluating the performance of DCT-II based dynamic feature extraction should be carried out and compared with other speech feature classifiers at the SR stage.

In this paper, the new and novel feature extraction algorithm was presented and simulated. A comparative analysis was carried out by comparing the new algorithm with algorithms mentioned in the literature. The simulation results highlighted the performance advantage achieved using the new algorithm and the option of using FVQ rather than GMM in the data clustering classifier stage was discussed.

The experimental methodology was presented and discussed. The steps in the ASR system presented in this thesis are consistent with the standard approach presented in the

literature. The innovative and new algorithm that utilizes DCT-II and DDMFCC to carry out the feature extraction step provides an improvement in overall performance.

References

- Abida, M. K. (2007). *Fuzzy gmm-based confidence measure towards keyword spotting application*. A thesis presented to the University of Waterloo in fulfilment of the thesis requirement for the degree of master of applied science, electrical and computer engineering, University of Waterloo, Ontario, Canada.
- Assaleh, K. T., & Mammone, R. J. (1994). Robust cepstral features for speaker identification. In *IEEE international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 129–132).
- Atal, B. S. & Hanauer, S. L. (1971). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6), 1304–1312.
- Barbu, T. (2009). Comparing various voice recognition techniques. In *Proceedings of the 5-th conference on speech technology and human-computer dialogue* (pp. 1–6).
- Charbuillet, C., Gas, B., Chetouani, M., & Zarader, J. L. (2007). Complementary features for speaker verification based on genetic algorithms. In *IEEE international conference on acoustics, speech and signal processing* (Vol. 4, pp. IV-285–IV-288).
- Chen, J., Paliwal, K. K., Mizumachi, M., & Nakamura, S. (2001a). Robust MFCCs derived from differentiated power spectrum. In *Proc. intern. conf. on speech processing*, TaeJen, Korea (Vol. 2, pp. 577–582).
- Chen, W., Zhenjiang, M., & Xiao, M. (2001b). Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16(16), 582–589.
- Chen, W., Zhenjiang, M., & Xiao, M. (2008). Differential mfcc and vector quantization used for real-time speaker recognition system. In *Congress on image and signal processing* (pp. 319–323).
- Cheng, J., & Wang, H. C. (2004). A method of estimating the equal error rate for automatic speaker verification. In *International symposium on Chinese spoken language processing* (pp. 285–288).
- Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28, 357–366.
- Ganchev, T. D. (2005). *Speaker recognition*. A dissertation submitted to the University of Patras in partial fulfilment of the requirements for the degree doctor of philosophy.
- Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4), 1738–1752.
- Hossan, M. A. (2011). *Automatic speaker recognition dynamic feature identification and classification using distributed discrete cosine transform based Mel frequency cepstral coefficients and fuzzy vector quantization*. A thesis presented to the RMIT University in fulfilment of the thesis requirement for the degree of master of engineering, electrical and computer engineering, RMIT University, Melbourne, Australia.
- Hossan, M. A., Memon, S., & Gregory, M. A. (2010). A novel approach for MFCC feature extraction. In *4th international conference on signal processing and communication systems (ICSPCS)*, Dec. 2010 (pp. 1–5), 13–15.
- Jayanna, H. S., & Prasanna, S. R. M. (2008). Fuzzy vector quantization for speaker recognition under limited data conditions. In *IEEE region 10 conference (TENCON 2008)* (pp. 1–4).
- Kanade, P. M., & Hall, L. O. (2007). Fuzzy ants and clustering. *IEEE Transactions on Systems, Man and Cybernetics. Part A. Systems and Humans*, 37(5), 758–769.
- Keshet, J., & Bengio, S. (2009). *Automatic speech and speaker recognition: large margin and kernel methods*. New York: Wiley.
- Kim, S., & Eriksson, T. (2004). A pitch synchronous feature extraction method for speaker recognition. In *IEEE international conference on acoustics, speech, and signal processing, 2004* (Vol. 1, pp. 405–408).
- MATLAB (2012). MATLAB & Simulink, Mathworks, USA. <http://www.mathworks.com.au/products/matlab/>. Accessed on 1 March 2012.
- Memon, S., Lech, M., & He, L. (2009a). Using information theoretic vector quantization for inverted mfcc based speaker verification. In *2nd international conference on computer, control and communication* (pp. 1–5).
- Memon, S., Lech, M., & Maddage, N. (2009b). Speaker verification based on different vector quantization techniques with Gaussian mixture models. In *Third international conference on network and system security* (pp. 403–408).
- National Institute of Standards and Technology speaker recognition evaluation (2004). <http://www.itl.nist.gov/iad/mig/tests/spk/2004/>. Accessed online 20/9/2010.
- Oppenheim, A. V. (1969). A speech analysis-synthesis system based on homomorphic filtering. *The Journal of the Acoustical Society of America*, 45, 458–465.
- Paul, A. K., Das, D., & Kamal, M. (2009). Bangla speech recognition system using lpc and ann. In *Seventh international conference on advances in pattern recognition* (pp. 171–174).
- Saeidi, R., Mohammadi, H. R. S., Rodman, R. D., & Kinnunen, T. (2007). A new segmentation algorithm combined with transient frames power for text independent speaker verification. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (Vol. 4, pp. 305–308).
- Sahidullah, M., & Saha, G. (2009). On the use of distributed DCT in speaker identification. In *2009 annual IEEE India conference (INDICON)* (pp. 1–4).
- Salman, A., Muhammad, E., & Khurshid, K. (2007). Speaker verification using boosted cepstral features with Gaussian distributions. In *IEEE international multitopic conference* (pp. 1–5).
- Shi, N., Liu, X., & Guan, Y. (2010). Research on *k*-means clustering algorithm: an improved *k*-means clustering algorithm. In *Third international symposium on intelligent information technology and security informatics (IITSI)*, 2–4 April 2010 (pp. 63–67).
- Wang, W., Zhang, Y., Li, Y., & Zhang, X. (2006). The global fuzzy C-means clustering algorithm. In *The sixth world congress on intelligent control and automation (WCICA 2006)* (Vol. 1, pp. 3604–3607).
- Wang, H., Zhang, X., Suo, H., Zhao, Q., & Yan, Y. (2009). A novel fuzzy-based automatic speaker clustering algorithm. In *6th international symposium on neural networks (ISNN 2009)*, Wuhan, China, 26–29 May 2009 (pp. 639–646), Part II.
- Wei-Guo, G., Li-Ping, Y., & Di, C. (2008). Pitch synchronous based feature extraction for noise-robust speaker verification. In *Congress on image and signal processing (CISP'08)* (Vol. 5, pp. 295–298).
- Zilca, R. D., Navratil, J., & Ramaswamy, G. N. (2003). Depitch and the role of fundamental frequency in speaker recognition. In *IEEE international conference on acoustics, speech, and signal processing, 2003* (Vol. 2, pp. 81–84).