A novel approach for MFCC feature extraction

Conference Paper · January 2011

DOI: 10.1109/ICSPCS.2010.5709752 · Source: IEEE Xplore

CITATIONS
87

READS
2,341

3 authors, including:

Sheeraz Memon
Mehran University of Engineering and Technology
23 PUBLICATIONS **192** CITATIONS

SEE PROFILE

Mark A Gregory
RMIT University
144 PUBLICATIONS **510** CITATIONS

SEE PROFILE

A Novel Approach for MFCC Feature Extraction

Md. Afzal Hossan, Sheeraz Memon, Mark A Gregory
School Electrical and Computer Engineering
RMIT University
Melbourne, Australia
mdafzal.hossan@student.rmit.edu.au, sheeraz.memon@rmit.edu.au, mark.gregory@rmit.edu.au

***Abstract*—The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction method is a leading approach for speech feature extraction and current research aims to identify performance enhancements. One of the recent MFCC implementations is the Delta-Delta MFCC, which improves speaker verification. In this paper, a new MFCC feature extraction method based on distributed Discrete Cosine Transform (DCT-II) is presented. Speaker verification tests are proposed based on three different feature extraction methods including: conventional MFCC, Delta-Delta MFCC and distributed DCT-II based Delta-Delta MFCC with a Gaussian Mixture Model (GMM) classifier.**

Keywords-Speech Feature Extraction, Mel Frequency Cepstral Coefficients (MFCC), Discrete Cosine Transform (DCT-II), Delta MFCC (DMFCC), Delta-Delta MFCC (DDMFCC), Gaussian Mixture Model (GMM).

I. INTRODUCTION

Speaker verification systems identify a person by analyzing and characterizing a person's voice [17]. A typical speaker verification system consists of a feature extractor followed by a robust speaker modeling technique for generalized representation of extracted features. Vocal tract information like formant frequency, bandwidth of formant frequency and other values may be linked to an individual person. The goal of a feature extraction block technique is to characterize the information [12], [11]. A wide range of possibilities exists for parametrically representing the speech signal to be used in the speaker verification activity [14], [15]. Some of the techniques used are: Linear Prediction Coding (LPC); Mel-Frequency Cepstral Coefficients (MFCC); Linear Predictive Cepstral Coefficients (LPCC); Perceptual Linear Prediction (PLP); and Neural Predictive Coding (NPC) [1], [2], [4]. MFCC is a popular technique because it is based on the known variation of the human ear's critical frequency bandwidth. MFCC coefficients are obtained by de-correlating the output log energies of a filter bank which consists of triangular filters, linearly spaced on the Mel frequency scale. Conventionally an implementation of discrete cosine transform (DCT) known as distributed DCT (DCT-II) is used to de-correlate the speech as it is the best available approximation of the Karhunen-Lo`eve Transform (KLT) [12]. Sahidullah used the DCT in distributed manner. MFCC data sets represent a melodic cepstral acoustic vector [3], [22]. The acoustic vectors can be used as feature vectors. It is possible to obtain more detailed speech features by using a derivation on the MFCC acoustic vectors. This approach permits the computation of the delta MFCC (DMFCCs), as the first order derivatives of the MFCC. Then, the delta-delta MFCC (DDMFCCs) are derived from DMFCC, being the second order derivatives of MFCCs.

Feature selection is followed by a classification algorithm to generate speaker specific data and the Gaussian mixture model (GMM) is currently being applied in the field of speaker verification and the use of GMM has been found to produce high quality results [17], [12].

In this paper, a speaker verification test is proposed based on DCT-II based DDMFCC speech features with a GMM classifier.

The rest of this paper is organised as follows. In Section II the basics of MFCC analysis is reviewed followed by the proposed feature extraction technique. In Section III the experimental arrangements for the proposed experiments are provided and the description of the existing DDMFCC and distributed DCT based DDMFCC approaches are summarised. Finally the conclusion is provided in Section IV.

II. MFCC FEATURE EXTRACTION METHOD

A. *Conventional Method*

Psychophysical studies have shown that human perception of the sound frequency contents for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale [17], [12] (1).

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

Where f_{mel} is the subjective pitch in Mels corresponding to a frequency in Hz. This leads to the definition of MFCC, a baseline acoustic feature set for speech and speaker recognition applications [12], [16].

MFCC coefficients are a set of DCT decorrelated parameters, which are computed through a transformation of the logarithmically compressed filter-output energies [6], 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).

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

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

$$\psi_i(k) = \begin{cases} 0 & \text{For } k < k_{b_{i-1}} \\ \dfrac{k - k_{b_{i-1}}}{k_{b_i} - k_{b_{i-1}}} & \text{For } k_{b_{i-1}} \leq k \leq k_{b_b} \\ \dfrac{k_{b_{i+1}} - k}{k_{b_{i+1}} - k_{b_i}} & \text{For } k_{b_{i-1}} \leq k \leq k_{b_{i+1}} \\ 0 & \text{For } k \leq k_{b_{i+1}} \end{cases} \tag{3}$$

If Q denotes the number of filters in the filter bank, then

$$\left\{ k_{b_i} \right\}_{i=0}^{Q+1}$$

are the boundary points of the filters and *k* denotes the coefficient index in the *Ms*-point DFT. The boundary points for each filter *i (i=1,2,...,Q)* are calculated as equally spaced points in the Mel scale using (4).

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

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).

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

In the next step, the output energies $e(i)(i = 1,2,....,Q)$ 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 $\psi_i(k)$ (6).

$$e(i) = \sum_{K}^{M} |y(k)|^2 \psi_i(k) \tag{6}$$

Finally, the DCT-II is applied to the log filter bank energies $\{\log[e(i)]\}_{i=1}^{Q}$ to de-correlate the energies and the final MFCC coefficients *Cm* are provided in (7).

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

Where, *m=0, 1, 2,..., R-1*, and *R* is the desired number of MFCCs.

B. Dynamic Speech Features

The speech features which are the time derivatives of the spectrum-based speech features are known as dynamic speech features. Memon and Maddage showed that system performance may be enhanced by adding time derivatives to the static speech parameters [18], [8]. The first order derivatives are referred to as delta features may be calculated as shown in (8).

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

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 size of delta window. The delta and delta-delta cepstra are evaluated based on MFCC [10], [21].

C. Distributed DCT

In Section II-A, DCT is used which is an optimal transformation for de-correlating the speech features [12]. This transformation is an approximation of KLT for the first order Markov process.

The correlation matrix for a first order markov source is given *by*

$$C = \begin{bmatrix} 1 & \rho & \rho^2 & . & . & \rho^{N-1} \\ \rho & 1 & \rho & \rho^2 & . & . \\ \rho^2 & \rho & 1 & \rho & \rho^2 & . \\ . & \rho^2 & \rho & 1 & \rho & \rho^2 \\ . & . & \rho^2 & \rho & 1 & . \\ \rho^{N-1} & . & . & \rho^2 & \rho & 1 \end{bmatrix} \tag{9}$$

Where ρ is the inter element correlation ($0 \leq \rho \leq 1$). Sahidullah showed that for the limiting case where $\rho \to 1$, the Eigen vector of (8) can be approximated as shown in (10) [12].

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

Where $0 \leq t \leq N-1$ and $0 \leq n \leq N-1$. Clearly Eqn. (9) is the Eigen function of the DCT. This is the reason behind the usage of DCT in the place of signal dependent optimal KLT transformation.

But in reality the value of ρ is not 1. In the filter bank structure of the MFCC, filters have placed on the Mel-frequency scale. As the adjacent filters have an overlapping

region, the neighbouring 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 [13]. 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 $\{\log[e(i)]\}_{i=1}^{[Q/2]}$ and SEG#2 $\{\log[e(i)]\}_{i=[Q/2]+1}^{Q}$.

Algorithm for Distributed DCT Based DDMFCC

1: **if** Q = EVEN **then**
2: $P = Q/2$;
3: **PERFORM** DCT of $\{\log[e(i)]\}_{i=1}^{[Q/2]}$ to get $\{C_m\}_{m=0}^{P-1}$;
4: **PERFORM** DCT of $\{\log[e(i)]\}_{i=P+1}^{[Q/2]}$ to get $\{C_m\}_{m=P}^{Q-1}$;
5: **else**
6: $P = \left\lceil \frac{Q}{2} \right\rceil$;
7: **PERFORM** DCT of $\{\log[e(i)]\}_{i=1}^{[Q/2]}$ to get $\{C_m\}_{m=0}^{P-1}$;
8: **PERFORM** DCT of $\{\log[e(i)]\}_{i=P+1}^{[Q/2]}$ 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 final feature vector $\{Cep(i)\}_{i=1}^{P-2}$;
12: **CALCULATE** *DDMFCC coefficients* d_t;

In this process the number of feature vectors is reduced by 1 for same number of filters compared to conventional MFCC. For example if 20 filters are used to extract MFCC features 19 coefficients are used and the 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 utilising the MFCC.

III. SPEAKER VERIFICATION EXPERIMENT

In this section speaker verification experiment setup is described and the speaker verification test results obtained based on DDMFCC [19] and distributed DCT based MFCC [12], [20] are discussed.

A. *Pre-processing*

The pre-processing stage includes speech normalisation, pre-emphasis filtering and removal of silence intervals [17], [18]. The dynamic range of the speech amplitude is mapped into the interval from -1 to +1. The high-pass pre-emphasis filter can then be applied to equalise the energy between the low and high frequency components of speech. 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 [19].

B. *Classification & Verification stage*

The GMM with expectation maximization is a feature modelling and classification algorithm widely used in the speech based pattern recognition, since it can smoothly approximate a wide variety of density distributions [17], [5]. Adapted GMMs known as UBM-GMM and MAP-GMM have further enhanced speaker verification outcomes [18], [9]. 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 (11).

$$p(x|y) = \sum_{i=1}^{M} p_i b_i(x) \tag{11}$$

Where x is a D-dimensional random vector, $b_i(x)$, the component densities are *i=1,2,3,…,M* and the mixture weights are p_i, for *i=1,2,3,…,M*. Each component density is a D-variate Gaussian function of the form shown in (12).

$$b_i(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)\sum_i^{-1}(x-\mu_i)\right\} \tag{12}$$

Where μ_i is the mean vector and $\sum$ 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,

$$\lambda = \{p_i, \mu_i, \sum_i\}, \; i = 1, 2 \ldots \mathrm{M} \tag{13}$$

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.

C. *Experiment Speech Databases*

The annually produced NIST speaker recognition evaluation (SRE) has become the state of the art corpora for evaluating methods used or proposed for use in the field of speaker recognition [18]. GMM-based systems have been widely tested on NIST SRE. The research will evaluate the proposed method utilising 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 verification 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.

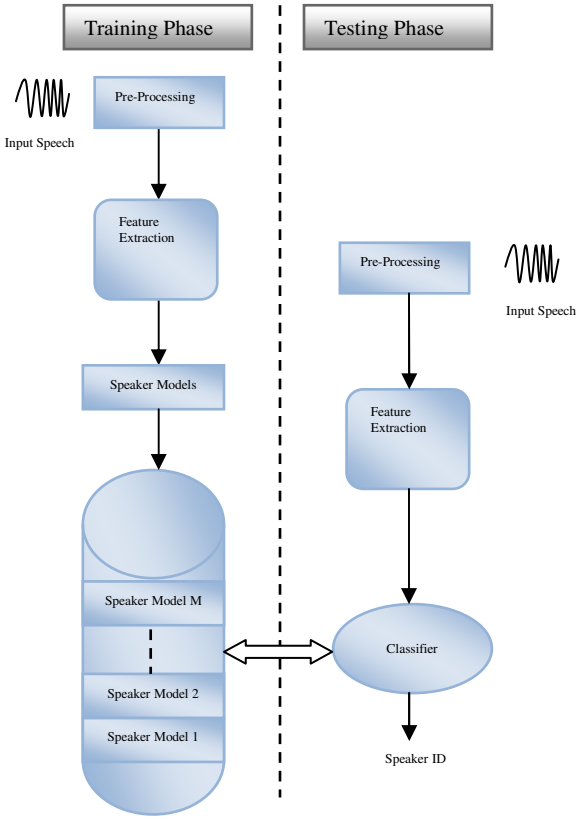

Figure 1. Speaker Verification System

D. *Speaker verification Experiments and Results*

In this research, the performance of speaker verification based on different MFCC feature extraction methods was evaluated. The percentage of Identification (100%-Equal Error Rate) [18] for the conventional MFCC is (90.36%) lower than the other feature extraction methods. A small improvement in the performance is seen by using DMFCC, while DDMFCC improved the performance significantly to 91.35%.

TABLE I
Percentage of Identification Accuracy

Classification Algorithm	Conventional MFCC	Delta MFCC	Delta-Delta MFCC	Distributed DCT based MFCC
GMM_EM	90.36%	90.68%	91.35%	96.72%

Sahidullah evaluated the performance of a speaker identification system based on distributed DCT based MFCC [12] and found that distributed DCT based MFCC outperformed the other feature extraction methods by improving the identification accuracy extensively to 96.72%.

The results found in the literature and from the initial research outcomes highlight that when using distributed DCT it is possible to find more de-correlated MFCC features than when DCT is used. The result is an improved speaker recognition outcome. As well as providing a dynamic speech feature, the use of DDMFCC, was found to significantly improve the speaker verification system performance.

The next step in the research is to combine DCT-II with DDMFCC feature extraction and to identify possible process refinements that will improve accuracy, performance and overall speaker recognition outcomes.

The research methodology has included a step by step analysis of previous speaker recognition systems based upon MFCC feature extraction and analysis of the positive attributes of the different approaches. The research has highlighted the positive use of DCT-II and DDMFCC in previous studies and the opportunity exists to combine the techniques and to refine the complete speaker verification system.

IV. CONCLUSION

In this research a new approach for the feature extraction to improve the performance of speaker verification systems has been identified and initial research outcomes comparing previous MFCC feature extraction approaches has been presented. Initially the performance of conventional MFCC was evaluated. The test system was refined with the use of DCT-II in the de-correlation process and results have been presented. The correlation among the filter bank output 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 improved performance in terms of identification accuracy with a lower number of features used and therefore 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 [8]. Improvements in the transitional characteristic capture was found by computing DMFCC and DDMFCC which were obtained respectively from the first-order and second-order time-derivative of the MFCC.

A new approach which applies DCT-II for de-correlation, incorporated with DDMFCC speech feature extraction and evaluation for speaker verification with a GMM classifier could establish better results based upon the results of the analysis carried out. Further work to complete a detailed analysis and system optimisation is currently being carried out.

This new approach for feature extraction is promising and may provide an improvement in results achieved in other studies. This novel approach will be implemented, results achieved and a comparative analysis with the results found using the test system that has been developed that permits previous approaches to be used and results computed.

REFERENCES

[1] Ahmed Salman, Ejaz Muhammad and Khawar Khurshid, "Speaker verification using boosted cepstral features with gaussian distributions," *IEEE International Multitopic Conference, 2007. INMIC 2007*, 2007 pp. 1 – 5.

[2] Anup Kumar Paul, Dipankar Das and Md. Mustafa Kamal, "Bangla speech recognition system using lpc and ann," *Seventh International Conference on Advances in Pattern Recognition*, 2009, pp. 171 – 174.

[3] T. Barbu, "Comparing various voice recognition techniques," *Proceedings of the 5-th Conference on Speech Technology and Human-Computer Dialogue*, 2009 pp. 1 – 6.

[4] C. Charbuillet, B. Gas, M. Chetouani and J. L. Zarader, "Complementary features for speaker verification based on genetic algorithms," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4 2007 pp. IV-285 - IV-288

[5] Wang Chen, Miao Zhenjiang and Meng Xiao, "Comparison of different implementations of mfcc," *J. Computer Science & Technology*, 2001, pp. 16(16): 582-589.

[6] T. D. Ganchev, "Speaker recognition," *A dissertation submitted to the University of Patras in partial fulfilment of the requirements for the degree Doctor of Philosophy*, 2005.

[7] Gong Wei-Guo, Yang Li-Ping and Chen Di, "Pitch synchronous based feature extraction for noise-robust speaker verification," *Congress on Image and Signal Processing (CISP '08)*, vol. 5 2008, pp. 295 - 298

[8] H. S. Jayanna and S. R. M. Prasanna, "Fuzzy vector quantization for speaker recognition under limited data conditions," *TENCON 2008 - IEEE Region 10 Conference* ,2008, pp. 1 - 4.

[9] Haipeng Wang, Xiang Zhang, Hongbin Suo, Qingwei Zhao and Y. Yan, "A novel fuzzy-based automatic speaker clustering algorithm," *ISNN*, 2009, pp. 639–646.

[10] J. Chen , K. K. Paliwal, M. Mizumachi and S. Nakamura, "Robust mfccs derived from differentiated power spectrum " *Eurospeech 2001, Scandinavia*, 2001.

[11] Joseph Keshet and Samy Bengio, *Automatic speech and speaker recognition : Large margin and kernel methods* wiley, west Sussex, United Kingdom, 2009.

[12] Md. Sahidullah and Goutam Saha, "On the use of distributed dct in speaker identification," *2009 Annual IEEE India Conference (INDICON)*,2009,pp.1-4.

[13] Parag M. Kanade and Lawrence O. Hall, *Fuzzy ants and clustering*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS VOL. 37**,** NO. 5 (2007).

[14] R. Saeidi, H. R. Sadegh Mohammadi, R. D. Rodman and T Kinnunen, "A new segmentation algorithm combined with transient frames power for text independent speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP,* , vol. 4, 2007, pp. IV-305 - IV-308.

[15] Ran D. Zilca, Jiri Navratil and Ganesh N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, vol. 2, 2003 pp. II - 81-84.

[16] Samuel Kim and Thomas Eriksson, "A pitch synchronous feature extraction method for speaker recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, vol. vol.1 pp. I - 405-408.

[17] Sheeraz Memon, Margaret Lech and Ling He, "Using information theoretic vector quantization for inverted mfcc based speaker verification," *2nd International Conference on Computer, Control and Communication, 2009. IC4 2009*, pp. 1 – 5.

[18] Sheeraz Memon, Margaret Lech and Namunu Maddage, "Speaker verification based on different vector quantization techniques with gaussian mixture models," *Third International Conference on Network and System Security*, 2009, pp. 403 - 408

[19] Shigeru Ono and Kazunori Ozawa, "2.4kbps pitch prediction multi-pulse speech coding," *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-88*, vol. 1, pp. 175 – 178.

[20] Syed Abdul Rahman Al-Haddad, Khairul Anuar Ishak, Salina Abdul Samad, Ali O. Abid and Aini Hussain Noor, "Robust digit recognition with dynamic time warping and recursive least squares," *International Symposium on Information Technology, 2008. ITSim 2008*, vol. 2 2008, pp. 1 - 8 .

[21] Wang Chen, Miao Zhenjiang and Meng Xiao, "Differential mfcc and vector quantization used for real-time speaker recognition system," *Congress on Image and Signal Processing*, 2008, pp. 319 - 323.

[22] Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang, "The global fuzzy c-means clustering algorithm," *Proceedings of the 6th World Congress on Intelligent Control and Automation*, 2006.