academicJournals

Vol. 8(35), pp. 1746-1752, 23 September, 2013
DOI: 10.5897/IJPS2013.3955
ISSN 1992 - 1950 © 2013 Academic Journals
http://www.academicjournals.org/IJPS

International Journal of Physical Sciences

Full Length Research Paper

Wavelet based dynamic Mel Frequency Cepstral Coefficients (MFCC) and block truncation techniques for efficient speaker identification under narrowband noise conditions

S. Selva Nidhyananthan*, R. Shantha Selva Kumara and D. S. Roland

Department of ECE, Mepco Schlenk Engineering College, Sivakasi-626005, India.

Accepted 09 September, 2013

Speaker identification strategies are well convincing in their performance when clean speeches are scrutinized. But the performance degrades when speech samples are corrupted by narrowband noise. Block truncation of the cepstral coefficients ensures that not all the features are affected by narrowband noise but it cannot reduce the extent of degradation. This work is focused towards improving the performance of speaker identification systems by block truncating the features which are subjected to wavelet processing. Wavelet decomposition divides the entire energy spectrum of the speech signal into bands corresponding to the number of levels of decomposition performed in the wavelet transformation thereby segregating the noise affected bands from other bands. In addition to that, wavelet filters provide the smoothening of the noisy speech signals which enhances the identification of the correct speaker. Dynamic Mel filtering of these wavelet coefficients followed by block truncation provides better identification, taking advantage of the fact that some filter bank coefficients remain unaffected by narrowband noise. The features are modeled by Gaussian mixture model - Universal background model (GMM-UBM) that serves as a generic one timed trained model. Speaker identification efficiency of 97.23% is achieved through this wavelet based dynamic MFCC technique which exhibits 7.58% improvement in speaker identification accuracy when compared with non wavelet based block truncation method.

Key words: Wavelet decomposition, block truncation, Dynamic Mel Filtering Cepstral Coefficients (DMFCC), Gaussian mixture model - Universal background model (GMM-UBM), speaker identification.

INTRODUCTION

Speaker identification is a biometric process (ZoranCirovi et al., 2010) of identifying a person by comparing the features extracted from the person to be identified, with the features extracted from the speakers enrolled in the database. The success of speaker identification depends upon the feature extracted and its modeling method. The feature extracted during testing phase will match the feature extracted during enrollment phase perfectly, if the conditions under which it is tested are ideal. But under practical circumstances the testing conditions will include noise disturbances such as car noise, train noise, narrowband noise, etc. The car and train noises can be thought of disturbances that degrade the performance effectively depending upon the signal to noise ratio.

*Corresponding author. E-mail: nidhyan@mepcoeng.ac.in, Tel.:+91-4562-235409, Fax: +91-4562-235111.

Abbreviations: MFCC, Mel Frequency Cepstral Coefficients; **UBM**, Universal Background Models; **DCT**, Discrete Cosine Transform; **MFLE**, Mel Filter Log Energies.

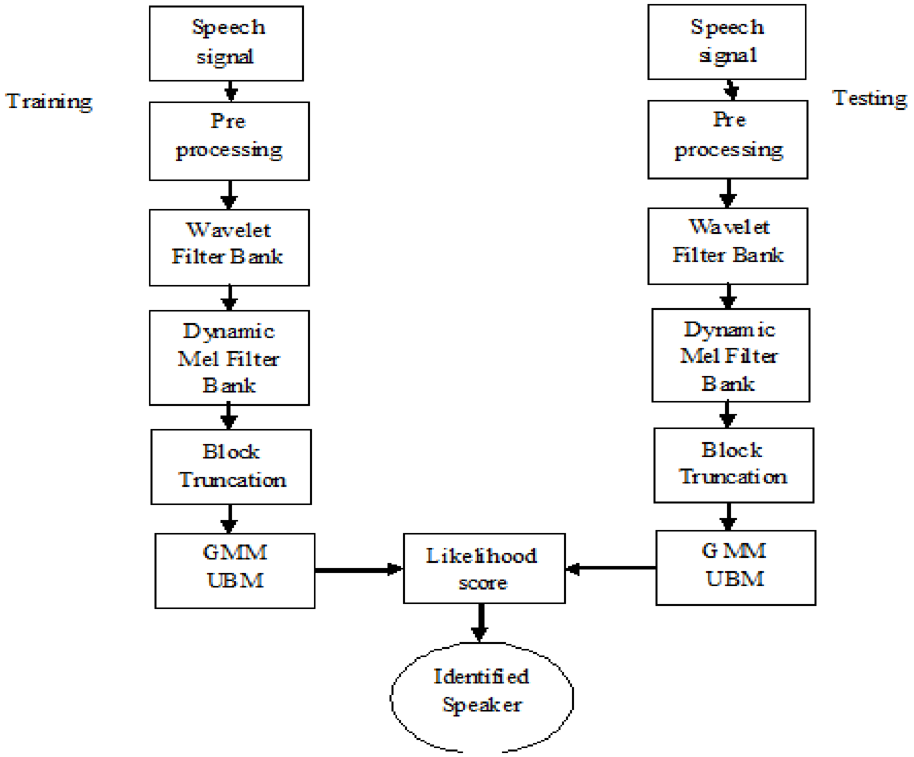

Figure 1. Overall block diagram of the proposed speaker identification system.

Moreover, the testing environments with these conditions are usually avoided in the practical speaker identification circumstances. But the narrowband noises are short time disturbances that are present in almost all the testing environments which lead to performance degradation of the speaker identification system. This problem can be overcome by a strategy called block based transformation, which is detailed in Sahidullah and Saha (2011).

The most common feature used for speaker identification is Mel Frequency Cepstral Coefficients (MFCC). As the human auditory system is most sensitive to the pitch frequency of the speaker, a feature that is based on the pitch frequency will model the speaker much more efficiently than the features that are not based on the pitch frequency of the speaker. In this work, Dynamic Mel Frequency Cepstral Coefficients (DMFCC) are used as features, which are formed by imparting the pitch frequency into the MFCC thereby producing dynamic features. The feature extracted is then used to model the speaker by using Universal Background Models (UBM). But the problem in modeling is that whether the model adapts itself for all the speakers enrolled in the database or not. This is called the bias/ variance dilemma problem (Utpal and Kshirod, 2012).

The Discrete Cosine Transform (DCT) is popularly used for MFCC and DMFCC computation, because the correlation matrix of Mel Frequency Log Energy data is similar to the correlation matrix of first order Markov process and DCT provides better energy compaction (Kekre and Vaishali, 2011) than any other linear transform. In this work, DCT is carried out in blocks to mitigate the effects of narrowband noise since the effect of a noise corrupted Mel Filter Log Energies (MFLE) will not pronounce in the MFCC obtained through other DCT blocks. Therefore by combining the narrowband noise overcoming strategies (Qi and Yan, 2011) of Block Truncated DCT and sub band processing, together with the added advantage of eliminating the bias / variance dilemma by sub band concept, enhanced speaker identification can be obtained. TIMIT database has been used in this work. The speeches in TIMIT database (John et al., 2013) was recorded at TI, transcribed at MIT and produced by the National Institute of Standards and Technology (NIST). The TIMIT database consists of 630 speakers.

This paper focuses on segregating the noise affected portions of speech, by making blocks of speech frames to confine the noise spread using wavelet transform, which reduces the effects of noise degradation when higher levels of decomposition is performed.

PROPOSED SYSTEM

The overall block diagram representing the speaker identification system is shown in Figure 1. The speech signal is first pre-processed and the energy spectrum of the speech wave is

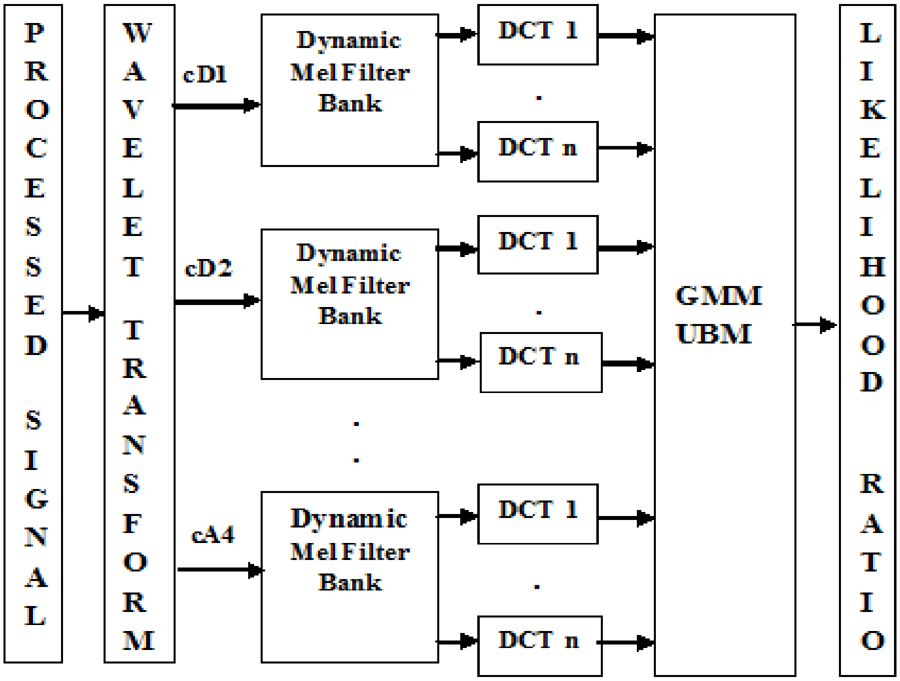

Figure 2. Block diagram of wavelet processed DMFCC.

computed using multilevel wavelet decomposition. The wavelet transformation divides the spectrum into appropriate number of bands corresponding to the number of levels of decomposition performed in the transformation. The wavelet filter outputs are robust to noise degradation due to the smoothening effect of the low pas filters while performing the discrete wavelet transformation on the speech signal. Moreover, the wavelet filter bank acts as dividing the entire bandwidth of the speech signal into bands corresponding to the number of levels used in the wavelet transformation. Moreover, the noise affected band is confined to one or few bands rather than being available at the whole spectrum. Hence it is easy to process the noise affected sub bands uniquely using Block truncation strategy. The methodology of the proposed work is enhancing the speaker identification under narrowband noise condition which makes the speech signal at 10 dB signal to noise ratio (SNR), by the features extracted through wavelet processing. Wavelet filters provide smoothening of the noise affected speech signal which provides reduction of the degradation caused by noise when higher levels of decomposition is performed on wavelet transformation. Further, the block truncation of the cepstral coefficients helps in restricting the effect of narrowband noise in affecting all the energy coefficients. Hence when the features are modeled using GMM-UBM, improvement in speaker identification accuracy can be achieved. The central schematic of this work lies in the sub band and block truncation techniques which is depicted in Figure 2. Each of these bands of wavelet coefficients is then provided as inputs to individual Dynamic Mel Filter Bank to obtain the Wavelet Dynamic Mel Frequency Cepstral Coefficients (WDMFCC). These features are more robust because the noise affected speech signal is smoothened by the low pass filters of the wavelet transformation process.

The features are then extracted by decorrelating the Dynamic Mel Filter Log Energies using block truncated DCT. Here, instead of taking DCT for the entire log energies of the sub band, DCT is performed in a block truncated manner because when DCT is performed on individual blocks narrowband noise in a block will not spread to the other blocks in the sub band. The features are then used to model the speaker by means of Gaussian Mixture Models-Universal Background Model (GMM-UBM).

Speaker identification steps

Speaker identification process is a kind of pattern classification. In pattern classification problem, the first step is evaluating representation of input pattern. In speaker identification, this step is evaluation of power spectrum. These acoustic representations are extracted within successive analysis windows of 20-30 ms overlapped by 10 ms size. As vocal tract is a slowly varying system, speech signal is nearly stationary over this analysis window. Other pre processing stages are briefly outlined here for the sake of completeness.

Pre-emphasis

Pre-emphasis (Tomi and Haizhou, 2010) is performed to boost the higher frequencies of the signal. It is performed with a pre-emphasis factor of 0.97 according to the equation given by:

$$y(n) = x(n) - \alpha x(n-1) \tag{1}$$

Pre-emphasis offsets the negative spectral slope of 20 dB per decade that is naturally present in the speech signals.

Windowing

The pre-emphasized signal is then segmented into smaller frames for the stationary property to be satisfied in taking DFT. Hamming window (Rabiner and Biing-Hwang, 2007) is used in this work. It is given by:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \quad ,0 \le n \le N-1 \tag{2}$$

The windowed signal is given by:

$$S_w(n) = y(n) * w(n) \tag{3}$$

where $y(n)$ is the pre-emphasized signal and $w(n)$ is the window used.

Energy spectrum

The energy spectrum of the windowed frames is computed by taking wavelet transform. Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks, and for perfect deconstructing and reconstructing finite, non-periodic and/or non-stationary signals. The transformation is given by:

$$X(k) = \left|DWT(S_w(n))\right|^2 \tag{4}$$

Wavelet bands

Successful speaker identification is critically dependent on obtaining good speaker models from training data. Data modeling is subjected to the bias/variance dilemma (Vibha and Jyoti, 2011). According to this, models with too many adjustable parameters will tend to overfit the data, exhibiting high variance and hence the model will generalise poorly. On the other hand, few parameters will make the model biased. Wavelet processing (Pawar and Badave, 2011) helps to solve this problem by dividing the entire energy spectrum of the speech signal into bands corresponding to the number of levels performed in the wavelet transformation. Also, the low pass filters of the wavelet transformation acts as smoothening filters thereby reducing the degradation of signal by noise. The wavelet energy coefficients are then fed to the Mel scale filters for extracting speaker specific features.

DMFCC feature extraction

Features are the representatives of the speech signal in speaker identification task. Feature extraction is the estimation of variables, called a feature vector, from another set of variables called speech samples. The feature extraction will transform the speech signal into feature vectors which present the specific properties of each speaker. Raw speech signals cannot be used as such for speaker identification because of two reasons: (i) direct comparison and identification are complex and unreliable and (ii) requires large storage capability.

The human auditory system can sensitively perceive the changes in pitch. The pitch frequency is calculated by taking the autocorrelation of the signal and then taking maximum value for the autocorrelation function. Therefore by incorporating the pitch information into the MFCC feature, dynamic mel frequency cepstral coefficients can be extracted, which proves to provide strong robustness to background noise compared other features (Wang, et al., 2009) thus increasing the identification rate.

$$Mel(f_{i\,p}) = 2595\log\left(1 + f_{i\,p}/700\right) \tag{5}$$

where $f_{i\,p}$ is the pitch frequency of i^{th} frame. The Mel frequency energy spectrum is then passed through the Gaussian Mel filter bank (Sandipan and Goutam, 2009) followed by cosine transformation to obtain DMFCC features.

Block truncation transformation

Discrete cosine transform is performed on Dynamic Mel Filter Log Energies (DMFLE) to decorrelate them and so to make the extracted feature suitable for modeling. When such DCT is applied to a narrowband noise affected speech signal's log energies, all the features will be affected by the noise and hence will make it unsuitable for speaker identification. To alleviate this problem, block based transformation is performed. The filter log energies are divided into blocks and DCT is performed on them. This will ensure that a narrowband noise affected block will not pronounce its effect in the feature extracted from the other DMFLE blocks of the speech signal and hence will facilitate correct identification of the speaker. Narrow band noise (Ming et al., 2007) is synthetically generated by adding four frequency components (that is, sinusoidal tones) of 2000, 2100, 2200 and 2300Hz. The amplitudes of the sinusoids are chosen randomly.

The filter bank log energies are decomposed into several blocks unlike standard full band based DCT technique. In this work, the whole signal is divided into non-overlapping blocks (Jingdong et al., 2000) and individual blocks are processed independently. Therefore, the presence of narrowband noise in one block will not affect the other blocks because of truncation. The transformation matrix can be given as:

$$L = \Phi_1 \oplus \Phi_2 \oplus \cdots \oplus \Phi_N = \begin{bmatrix} \Phi_1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \Phi_2 & 0 & \cdot & \cdot & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & \cdot & \cdot & \Phi_N \\ & & & & & \end{bmatrix} \tag{6}$$

Where $\Phi_1, \Phi_2, \Phi_3 .... \Phi_N$ are orthogonal discrete cosine transformation matrices applied to individual blocks of Dynamic Mel Filter Log Energies (DMFLE).

Suppose two blocks of same sizes *q* are considered then the DCT matrix of size $q \times (q-1)$ is given by $\phi_1 = \sqrt{\frac{2}{q}}\cos\left[\frac{\pi i(2j+1)}{2q}\right]$ and $\phi_2 = \sqrt{\frac{2}{p-q}}\cos\left[\frac{\pi(i-q)(2j+1)}{2(p-q)}\right]$.

GMM-UBM

A Universal Background Model (UBM) is a model used in a biometric identification system to represent general, person independent feature characteristics to be compared against a model of person-specific feature characteristics. The likelihood ratio statistic is given by:

$$LR(X) = \frac{p(X / \lambda_p)}{p(X / \lambda_{\bar{p}})} \tag{7}$$

where $p(X/\lambda_p)$ is the probability that the feature models the speaker correctly, $p(X/\lambda_{\bar{p}})$ is the probability that the feature belongs to the alternate hypothesis.

$\lambda = \{w_i, \mu_{i,} \sigma_i\}$, i=1,2,…*M*

M is the number of Gaussian components, w_i is the mixture weights, μ_i is the means and σ_i is the variance.

The alternate hypothesis that gives the probability of speaker belonging to the false category is modeled by means of UBM. It is a speaker independent one time trained model. Since UBM is a large GMM (Reynolds, 1995; Reynolds and Rose, 1995) trained to represent the speaker independent features, its idea is to capture the general characteristics of a population and then adapting it to the individual speaker by means of EM algorithm. With training vectors from the hypothesized speaker, *X*={ $x_1, x_2 \ldots x_T$ } and for mixture '*i*', the probability distribution is given by:

$$\Pr(i/x_i) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{8}$$

Then with the distribution known, the statistics for the weight, mean, and variance parameters are initialized as given as follows:

$$n_i = \sum_{i=1}^{T} \Pr(i/x_t)$$

$$E_i(x) = \frac{1}{n_i}\sum_{t=1}^{T} \Pr(i/x_t)x_t \tag{9}$$

$$E_i(x^2) = \frac{1}{n_i}\sum_{t=1}^{T} \Pr(i/x_t)x_t^2$$

The updated coefficients are given by:

$$\begin{aligned} w_i^{new} &= [\alpha n_i / T + (1-\alpha) w_i]\gamma \\ \mu_i^{new} &= \alpha E_i(x^2) + (1-\alpha)\mu_i \\ \sigma_i^{2new} &= \alpha E_i(x^2) + (1-\alpha(\sigma_i^2 + \mu_i^2)) - \mu_i^{2new} \end{aligned} \tag{10}$$

where $\alpha = \frac{n_i}{n_i + r}$ with relevance factor r = 16

RESULTS AND DISCUSSION

TIMIT database is used for the analysis in this speaker identification work. Each speaker record contains ten speech signals, from which six signals are used for training and the remaining four signals are used for testing. The database consists of clean speech recorded at 16 kHz sampling frequency. The maximum frequency content in the speech waveform is 8 kHz. The TIMIT database speech signals are subjected to Narrow band noise generated by adding four frequency components of 2000, 2100, 2200 and 2300 Hz. The narrow band noise affected speech signal is first pre-processed and the energy spectrum of the speech wave is computed using four level decomposed wavelet transformation. Thus energy spectrum is obtained into bands corresponding to the number of decomposition levels. Now the noise affected band is confined to one or few bands rather than being available at the whole spectrum. Hence, it is easy to process the noise affected sub bands uniquely using Block truncation strategy. Since the band processing is done well ahead of decorrelarion step, the narrowband noise affected portion can be segregated from the rest of the bands that are not affected by the noise and better decorrelation is achieved at the Discrete Cosine Transform stage of Dynamic Mel Frequency Cepstral Coefficient feature extraction.

To improve the speaker identification efficiency, Dynamic Mel scale filter bank is constructed using Gaussian shaped filters in contrary to the triangular filters used in conventional systems. A triangular filter provides crisp partitions in an energy spectrum by providing non-zero weights to the portion covered by it while giving zero weight outside it. This phenomena cause loss of correlations between a sub band output and the adjacent spectral components that are present in the other sub band, whereas Gaussian shaped filters can provide much smoother transition from one sub band to other preserving most of the correlation between them. The DMFCC feature thus extracted is modeled using GMM-BUM modeling with 2048 mixture components.

During testing, the percentage of correct identification is calculated by using the formula:

$$\textit{Percentage of correct identifica tion} = \frac{\textit{No. of utterances correctly identified}}{\textit{Total no. of utterances under test}} * 100$$

The DMFCC features results have been observed for frame sizes of 1024, 512, and 256 without wavelet decomposition. The identification performance is tabulated in Table 1. The performance deteriorates in full band DCT since the effect of narrowband noise will be spread out to all the filter bank coefficients. But under the pro block based DCT system, the identification rate enhances to 89.65% for a frame size of 256. This proves the advantage of block based transformation under narrowband noise conditions. This accuracy can be further enhanced by the proposed wavelet based DMFCC feature. The WDMFCC features results have been observed for frame sizes of 1024, 512, and 256 with two level wavelet decomposition. The identification performance is tabulated in Table 2.

The percentage of correct identification improves to 94.13% for the frame size of 256. The wavelet filters provides coefficients that represent the smoothened version of the noise affected speech signal which when dynamic mel filtered followed by block truncation, reduces the noise degradation and provides improved identification percentage. Now the WDMFCC features results have been observed for frame sizes of 1024, 512, and 256 with three level wavelet decomposition. The identification performance is tabulated in Table 3. It is inferred from Table 3 that the performance of speaker

Table 1. Identification performance for DMFCC with and without block truncation

S/N	Total number of speakers	GMM-UBM		
		Frame size	Identification accuracy in presence of narrowband noise (%)	
			Without block truncation	With block truncation
1		1024	61.57	73.05
2	630	512	75.25	81.97
3		256	83.73	89.65

Table 2. Identification performance for WDMFCC with and without block truncation for two-level wavelet decomposition

S/N	Total number of speakers	GMM-UBM		
		Frame size	Identification accuracy in presence of narrowband noise (%)	
			Without block truncation	With block truncation
1		1024	71.13	82.88
2	630	512	80.50	87.25
3		256	87.21	94.13

Table 3. Identification performance for WDMFCC with and without block truncation for three level wavelet decomposition

S/N	Total number of speakers	GMM-UBM		
		Frame size	Identification accuracy in presence of narrowband noise (%)	
			Without block truncation	With block truncation
1		1024	72.75	83.5
2	630	512	81.13	88.38
3		256	88.25	96.37

identification system is improved to 96.37% corresponding to frame size of 256. The improvement corresponds to the higher level decomposition of the wavelet transformation which exhibit better noise smoothening compared to the results obtained through two level wavelet processed DMFCC results.

The identification performance for WDMFCC features with four level wavelet decomposition for frame sizes of 1024, 512, and 256 is tabulated in Table 4. From Table 4, it is observed that the identification accuracy is improved to 97.23% for the Dynamic Mel features obtained through four-level wavelet decomposition. The accuracy is enhanced by 7.58%. The results are compared with (Ramaligeswararao et al., 2011) the work on text-independent speaker identification model is developed by integrating MFCC’s with Independent component analysis (ICA) for obtaining feature independency and to achieve low dimensionality in feature vector extraction. The work by Ramaligeswararao et al. (2011) evaluated the speaker identification performance for a database of 50 speakers under 0dB, 10dB and 20 dB SNR conditions and obtained a maximum identification performance of 72.34% for 10 dB SNR and 88.45% for 20 dB SNR. But our proposed work achieves 97.23% for 630 speakers even at 10 dB SNR.

Conclusion

In this paper, the speaker identification rate under

Table 4. Identification performance for WDMFCC with and without block truncation for four level wavelet decomposition

S/N	Total number of speakers	GMM-UBM		
		Frame size	Identification accuracy in presence of narrowband noise (%)	
			Without block truncation	With block truncation
1	630	1024	73.62	84.69
2		512	81.63	88.87
3		256	88.50	97.23

narrowband noise conditions is found to be enhanced by the features extracted through wavelet processing. Wavelet filters provide smoothening of the noise affected speech signal which provides reduction of the degradation caused by noise when higher levels of decomposition is performed on wavelet transformation. The block truncation of the cepstral coefficients helps in restricting the effect of narrowband noise in affecting all the energy coefficients. The identification performance stands at 97.23% for the four level wavelet decomposed WDMFCC for a frame size of 256. Further developments such as fusion of several other features with adaptive weights can improve the narrowband noise performance to significant levels of successful identification accuracies.

REFERENCES

Kekre HB, Vaishali K (2011). Speaker Identification using Row Mean of DCT and Walsh Hadamard Transform. Int. J. Comput. Sci. Eng. 3(3):1295–1301.

Jingdong C, Paliwal K, Nakamura S (2000). A block cosine transform and its application in speech identification. In: Proc Int. Conf. Spoken Language Processing (INTERSPEECH 2000 – ICSLP) IV:117–120.

John HL, Hansen L, Jun-Won S, Matthew RL (2013). In-set/out-of-set speaker recognition in sustained acoustic scenarios using sparse data. Speech Commun. 55(6):769–781.

Ming J, Timothy JH, James RG, Senior M, Douglas AR, Senior M (2007). Robust speaker identification in noisy conditions. IEEE Trans. Audio Speech Lang. Process. 15(5):1711–1723.

Ramaligeswararao NM, Sailaja V, Srinivasa R (2011). Text Independent Speaker Identification using Integrated Independent Component Analysis with Generalized Gaussian Mixture Model. Int. J. Adv. Comput. Sci. Appl. 2:12.

Pawar MD, Badave SM (2011). Speaker Identification System Using Wavelet Transformation on Neural Network. Int. J. Comput. Appl. Eng. Sci. I Special Issue on Cns ,July 2011.

Rabiner LR, Biing-Hwang J (2007). Fundamentals of speech Identification. Pearson Education Book.

Reynolds D, Rose R (1995). Robust text-independent speaker Identification using Gaussian mixture speaker models." Speech Audio Process. IEEE Trans. 3(1):72–83.

Reynolds DA (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communication.

Sahidullah Md, Saha G (2011). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker identification, IEEE.

Sandipan C, Goutam S (2009). Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter. Int. J. Info. Commun. Eng. 5(1).

Tomi K, Haizhou L (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech Commun. 52:12–40.

Wang Y, Li B, Jiang X, Liu F, Wang L (2009). IEEE April 2009 Speaker Identification based on Dynamic MFCC parameters, pp. 406-409.

Qi L, Yan H (2011). An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions. IEEE Trans. Audio, Speech Lang. Process. 19(6).

Utpal B, Kshirod S (2012). GMM-UBM Based Speaker Verification in Multilingual Environments. Int. J. Comput. Sci. 9(6):2.

Vibha T, Jyoti S (2011). Wavelet Based Noise Robust Features for Speaker Recognition. Sig. Process. An Int. J. (SPIJ) 5(2).

ZoranCirovi C, Milan M, Zoran B (2010). Multimodal Speaker Verification Based on Electroglottograph Signal and Glottal Activity Detection. EURASIP J. Adv. Sig. Process. Article ID 930376, p. 8.