TELKOMNIKA, Vol. 11, No. 5, May 2013, pp. 2731 ~ 2738
ISSN: 2302-4046

Wavelet Cesptral Coefficients for Isolated Speech Recognition

T. B. Adam*[1], M. S. Salam[1], T. S. Gunawan[2]
[1]School of Computing
Universiti Teknologi Malaysia, 81300 Skudai, Johor, Malaysia
[2]Department of Electrical and Computer Engineering
International Islamic University Malaysia
*Corresponding author, e-mail: tarmizi_adam2005@yahoo.com, sah@utm.my, tsgunawan@iium.edu.my

Abstract

The study proposes an improved feature extraction method that is called Wavelet Cepstral Coefficients (WCC). In traditional cepstral analysis, the cepstrums are calculated with the use of the Discrete Fourier Transform (DFT). Owing to the fact that the DFT calculation assumes signal stationary between frames which in practice is not quite true, the WCC replaces the DFT block in the traditional cepstrum calculation with the Discrete Wavelet Transform (DWT) hence producing the WCC. To evaluate the proposed WCC, speech recognition task of recognizing the 26 English alphabets were conducted. Comparisons with the traditional Mel-Frequency Cepstral Coefficients (MFCC) are done to further analyze the effectiveness of the WCCs. It is found that the WCCs showed some comparable results when compared to the MFCCs considering the WCCs small vector dimension when compared to the MFCCs. The best recognition was found from WCCs at level 5 of the DWT decomposition with a small difference of 1.19% and 3.21% when compared to the MFCCs for speaker independent and speaker dependent tasks respectively.

Keywords*: Speech recognition, Speech processing, Cepstral analysis, Wavelet Transform, Feature Extraction*

1. Introduction

Feature extraction (FE) could be seen as one of the most significant phases in Speech Recognition (SR) systems. The FE phase in SR systems plays a major role in the accuracy of the SR system. In other words, in order to obtain reliable accuracy for SR systems the feature extraction phase should yield speech features that are easy to discriminate and classify between different classes.

Traditionally, the most dominant and popular feature extraction technique is the Mel-Frequency Cepstral Coefficient (MFCC) [1, 2]. It was shown by Davis and Mermelstein that MFCCs outperformed several other speech features thus making it the most widely used feature for SR systems [3].

However, MFCCs suffer from several problems [4]. Experiments showed that the MFCCs behaved poorly under noisy conditions. MFCC features extracted from noisy speech signal showed reduced accuracy [5]. Another issue regarding MFCCs is related to the fixed window or frame used when computing the MFCCs [6]. With fixed frame size, speech samples lying between the frames are assumed to be stationary. Unfortunately, this assumption is not true as speech signals tend to be non-stationary in nature. Thus, information such as plosive sounds is difficult to extract [7, 8].

With fixed frame size, abrupt changes and localized events such as sharp transitions in speech signals cannot be analyzed or extracted with the use of MFCCs. These localized events may contain significant information that may be important to further increase the speech recognition system accuracy [9, 10]. As an example, more information from the speech signal must be retained from acoustic confusable words.

To address these issues, wavelets has been of particular interests. The use of Discrete Wavelet Transform (DWT) or Wavelet Packet Transform (WPT) for feature extraction has been shown in several works. In this paper, we propose a set of new features called Wavelet Cepstral

Received December 18, 2012; Revised March 19, 2013; Accepted March 26, 2013

Coefficients (WCC) for isolated spoken English alphabet recognition. Here, the WCC are proposed to remedy the issues possessed by MFCCs.

The rest of the paper is structured as follows, section 2 review some related works that uses wavelets and wavelet cepstrum. Section 3 reviews some theoretical and mathematical background used in this paper. Section 4 and 5 explains the method and the proposed feature extraction. Section 6 explains the experimental verification. While in sections 7, 8 and 9 we present the results, discussions and conclusions respectively.

2. Background

Sarikaya*et al*. [5] used WPT to replace the Discrete Cosine Transform (DCT) to obtain a set of features called Wavelet Packet Parameters (WPP). Wu and Lin proposed an Irregular Wavelet Packet decomposition feature based on the energy of an uttered word in order to improve the performance speaker identification systems [11]. Their proposed method applies the WP decomposition to frequency regions that are observed to have high energy values as a final result 96.6% recognition rate were obtained.

DWT were also used by Gowdy and Tufekci to obtain a new feature vector called Mel-Frequency Discrete Wavelet Coefficients (MFDWC) [9]. The MFDWC were obtained by applying DWT to the Mel-scaled log filterbank energies of a speech frame. Results showed that the MFDWC performed better in terms of recognition over other features that were used for the test.

The use of Admissible Wavelet Packet (AWP) by Farooq and Datta to recognize phonemes showed good results when compared to MFCCs [12]. Their proposed FE also yielded better results than MFCCsunder different types of noise. The use of (AWP) were also proposed by Deshpande and Holambe [13]. Here, the authors proposed a filter structure that best represents the signal without taking any human auditory scale into consideration for use in speaker identification application.Thus, it can be concluded from these studies that the wavelet can be utilized in feature extraction phase to increase speech recognition accuracy. The experiments also suggest that wavelets can be used as an alternative to MFCCs for feature extraction.

2.1. Wavelets and Cepstrum Calculation

Several works have been done in using wavelets for computing the cepstrum as in [1, 14, 15]. The paper by Kinney [14] proposed decomposing the speech signal using wavelet packet transform (WPT) and then calculating the real cepstrum for each coefficients or atoms obtained from that decomposition. Promising results were obtained for text-dependent speaker recognition considering the methods few feature coefficients. It was shown that 90% of speakers were recognized when using 9 training vectors.

In [15] a wavelet based cepstrum calculation was proposed. The proposed wavelet based calculation was used for pitch extraction in speech signals. Different types of wavelet family were used to find the optimal accuracy for pitch extraction.

Recently, Pavez and Silva [1] proposes the Wavelet Packet Cepstral Coefficients (WPCC) as an alternative to filter-bank energy based feature extraction. In their work, detailed filter design were presented to obtain the WPCC as an alternative to the widely used MFCCs. Results show that the WPCC are better than MFCCs and has the ability to retain more phone discriminating information in the speech signal at lower frequency ranges.

3. Theoretical Background

3.1. Cepstral Analysis

Cepstral analysis is an important concept in many task related to speech processing. For example, cepstrum could be used for pitch detection and formant estimation. Also, the MFCC as previously stated is also considered to be a cepstral analysis method in which the analysis is done in the Mel frequency scale.Computing the cepstral coefficients for each speech frame involves three steps (refer to Figure 1):

1. DFT of the speech frame.
2. Log energy spectrum calculation.
3. DCT of the log energy spectrum.

A frame of speech is first subjected to the DFT with the following equation.

$$S_w(K) = \sum_{n=0}^{N-1} s_w(n) e^{\frac{-2\pi kjn}{N}} \quad (1)$$

Next, the spectrum undergoes log power calculation to obtain the log power spectrum.

$$m_k = \log\left|S_w(K)\right|^2 \quad (2)$$

Finally, taking the DCT of the log power spectrum yields (3)

$$c_i = \sqrt{\frac{2}{N}} \sum_{k=1}^{N} m_k \cos\left(\frac{\pi i}{N}(k-0.5)\right) \quad (3)$$

Where c_i is the cepstrum or cepstral coefficient.

Figure 1. Block diagram of cepstrum calculation

3.2. Mel-Frequency Cepstral Coefficients

As mentioned in Section 1 the MFCC is the most widely used feature for speech recognition purpose. Computing the MFCC from a given speech signal follows several steps.

First, the speech is passed to a pre-emphasis filter with the form of $H_{pre}(z) = 1 + a_{pre} z^{-1}$ where a typical value of a_{pre} is usually near -1 [16]. Next the speech signal is framed and windowed.

The duration of the frame are usually set to 256 samples with 128 samples of overlapping between the frames while the hamming window with the function $w(n) = 0.54 + 0.46 \cos\left(\frac{2\pi t}{T}\right)$ is used. Then, for each windowed speech frame the discrete Fourier Transform (DFT) is computed and the power spectrum of the DFT is binned with a set of Mel-scaled triangular filterbank. Next, the logarithm of the mel-spectral coefficients is taken. The final step is applying the discrete cosine transform (DCT) to the logarithm scaled mel-spectral coefficients to obtain the MFCCs. A detailed computation of the MFCC is presented in [3].

3.3. Discrete Wavelet Transform

A wavelet is a short oscillating signal or function that has a finite duration with an average value of zero. Given a discrete signal $s(n)$ with period N, the Discrete Wavelet Transform (DWT) of the signal is:

$$DWT\left[n, 2^j\right] = \sum_{m=0}^{N-1} s[n] \psi^*_{2^j}[m-n] \quad (4)$$

Where j is the level of decomposition and

$$\psi_{2^j}[n] = \frac{1}{\sqrt{2}} \psi\left(\frac{n}{2^j}\right) \quad (5)$$

The DWT decomposes an input signal into a set of approximation and detail coefficient. The approximation is recursively decomposed into a binary tree like structure leaving the details without further decomposition.

3.4. Neural Network Classifier

A multilayer perceptron (MLP) back propagation with adaptive learning rate Neural Network (NN) was employed to train and classify all the 26 English alphabets. Depending on the features used, the input node to the NN classifier would have 800 for MFCC, 90 for WCC at level 8, 60 at level 5 and 40 at level 3.

The learning rate and the momentum coefficient of the NN were varied to obtain the best recognition rate. For the classification, the input features were normalized between -1 and 1 in which 1 indicates true classification while -1 indicates false classification. The activation function for the NN is the hyperbolic tangent activation function as it guarantees the output to fall between this range.The hyperbolic tangent activation function also speeds up the learning process of the NN [17]. This activation function was used for both hidden and output nodes.

4. Method

We propose that the DFT block in Figure 1 be replaced with the DWT because of the advantages of the wavelet transform. The coefficients from the DWT of the speech signal are then subjected tolog power spectrum and DCT. The final output is what we call Wavelet Cepstral Coefficient (WCC).For each WCC computed from the DWT coefficients, only the first ten coefficients from each WCC are concatenated (Figure 2) to form the overall feature vector of the WCC.

4.1. Database

The speech samples used for training and testing were from the standard TI46 isolated alphabets. This Dataset of isolated speech utterance was developed by Texas Instruments (TI). Although the dataset contains both isolated alphabet and digit speech recordings the experiments conducted only used alphabets. Overall, the TI46 database contains 16 speakers, eight female speakers (F1 to F8) and eight male speakers (M1 to M8). Each of the alphabets A to Z was uttered 16 times from each speaker.

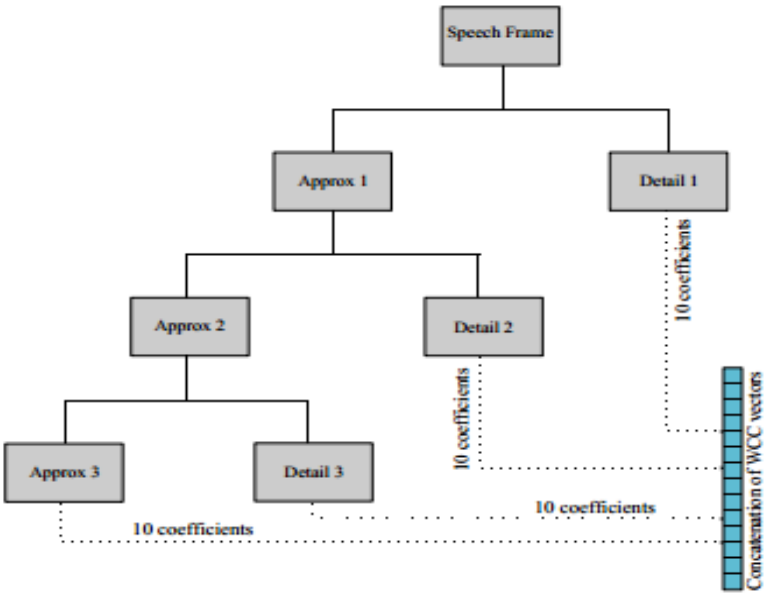

Figure 2. Process of obtaining WCC from DWT coefficients

To train the NN, five female speakers F1 to F5 and three male speakers M1, M2 and M3 were used. Thus, the tests were conducted in either speaker dependent (SD) or speaker independent (SI) mode. SD involves testing the trained NN classifier with the same speakers that were used for training while SI tests were conducted with the ones not used for training.

5. Proposed Feature Extraction

The first step in computing the proposed WCC is speech end point detection or silence removal. Here, the silence at the beginning and end of the speech is omitted. Then the speech undergoes pre-emphasis filtering, framing and windowing as explained in section 3.2 for the MFCC calculation. Next, instead of undergoing the DFT (As in Section 3.1) the speech frames are decomposed with the DWT (Refer Section 3.3).The coefficients produced from the DWT are then subjected to log power spectrum and DCT calculations. Each wavelet coefficient produces the WCC however, to reduce the dimension of the feature extracted only ten coefficients are retained from each wavelet decomposition (Figure 2).Finally, the WCC are fed into the NN classifier for either training or testing.

6. Experimental Verification

The experiment was conducted by varying the level of DWT. WCC were extracted from level 8, level 5, and level 3. To benchmark our proposed WCC we compared the results with the MFCC. The task was to recognize all 26 English alphabets which is quite a difficult task because of several acoustic similarities between the letters. Thetest was conducted to observe the effects of decomposition level of the DWT with regard to the recognition rate. We also wanted to observe the effectiveness of the WCC in classifying acoustically confusable letters when compared to the MFCCs.Table 1 and 2 are several setup for the experiment.

For the MFCC features, an 800 feature vector was obtained for each speech. This value was fixed by means of zero-padded normalization as mentioned in [18]. For the WCC 90, 60 and 40 coefficient feature vector was used for each speech signal. The number of feature used also denotes the number of input nodes needed for the NN classifier.

Table 1. MFCC parameter settings

Parameter	Value
Frame size	256 samples
Frame overlap	128 samples
Pre-emphasis coefficient (α_{pre})	-0.95
Number of triangular band-pass filters	20
Number of MFCC coefficients	13 with energy

Table 2. Neural network classifier setup

Parameter	Value
Input layer	800 nodes for MFCC, 90, 60 and 40 for WCC
Hidden layer	1 hidden layer with 250 nodes
Output layer	26 nodes
Hidden layer activation function	Hyperbolic tangent
Output layer activation function	Hyperbolic tangent

The number of coefficients for the WCC is obtained by retaining only ten coefficients from each DWT decomposition level (Figure 2). In this study we have restricted for only ten coefficients from each level for evaluation. Ten coefficients from approximation coefficient and ten coefficients from each level of detail coefficients are used.

Let their be n levels of wavelet decomposition in which each level contains both detail and approximation coefficients. Except for the n^{th} level, ten coefficients from the detail

coefficients are taken while ten coefficients are taken for both detail and approximate coefficients on the level n^{th} wavelet decomposition. The relationship is presented in (6) as

$$N_c = 10n + 10 \tag{6}$$

Where N_c is the number of WCC obtained from a n level wavelet decomposition. Thus, for example an 8 level wavelet decomposition yields $10 \times 8 + 10 = 90$ WCC coefficients.

7. Results

Table 3 and 4 shows the results for the proposed WCC on speaker dependent and speaker independent tasks. Comparisons with MFCCs were done to evaluate the effectiveness of the WCC. Figure 3 and Figure 4 shows the histogram plot for the the average recognition of both MFCC and WCCs. Average recognition were obtained from different values of the learning rate (LR) and momentum constant (MC) of theNN classifier. Different values of LR and MC were needed to obtain the best results for each of the WCCs and MFCC.

Table 3. Comparative results between MFCC and WCC (Speaker dependent)

Speaker	Recognition MFCC (%)	Recognition WCC lvl 8 (%)	Recognition WCC lvl 5 (%)	Recognition WCC lvl 3 (%)
F1	81.31	76.74	77.71	72.96
F2	78.79	66.65	73.02	69.95
F3	84.50	75.54	78.55	82.81
F4	87.56	79.81	81.13	78.25
F5	81.01	81.37	81.97	75.24
M1	74.94	66.23	72.42	61.42
M2	93.03	88.16	91.29	85.10
M3	93.03	72.36	76.38	70.07
Average	84.27	75.86	79.06	74.48

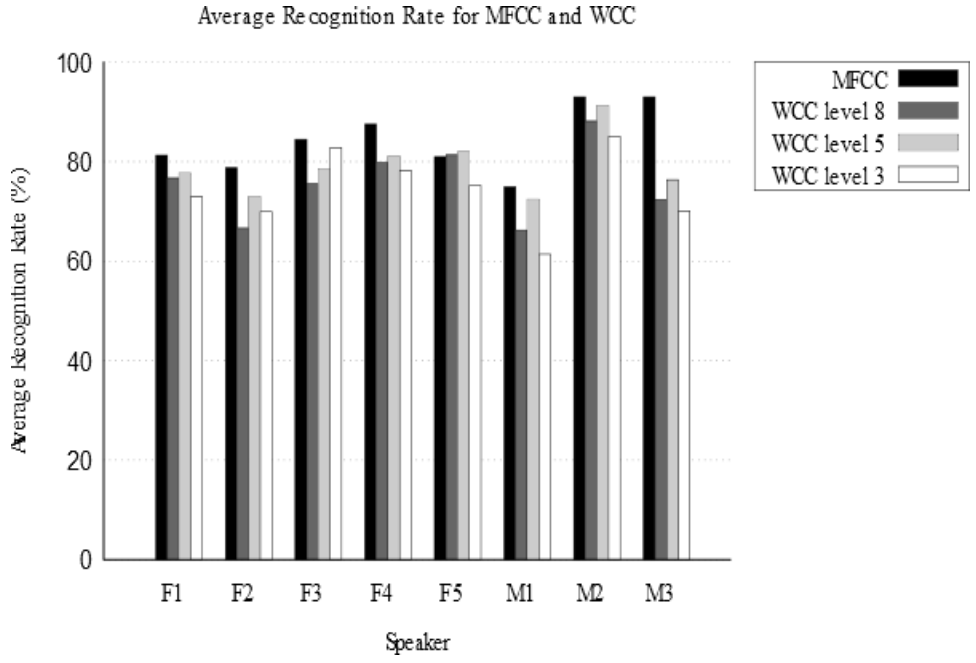

Figure 3. Results for speaker dependent tasks

Table 4. Comparative results between MFCC and WCC (Speaker independent)

Speaker	Recognition MFCC (%)	Recognition WCC lvl 8 (%)	Recognition WCC lvl 5 (%)	Recognition WCC lvl 3 (%)
F6	75.60	61.84	68.75	65.87
F7	74.52	72.96	79.45	75.72
F8	68.13	59.94	64.88	63.86
M4	85.60	80.15	82.60	78.19
M5	69.55	71.32	70.10	65.20
M6	73.29	62.35	60.82	54.04
M7	71.48	67.72	70.15	64.57
M8	56.97	64.90	68.81	62.82
Average	71.89	67.65	70.70	66.29

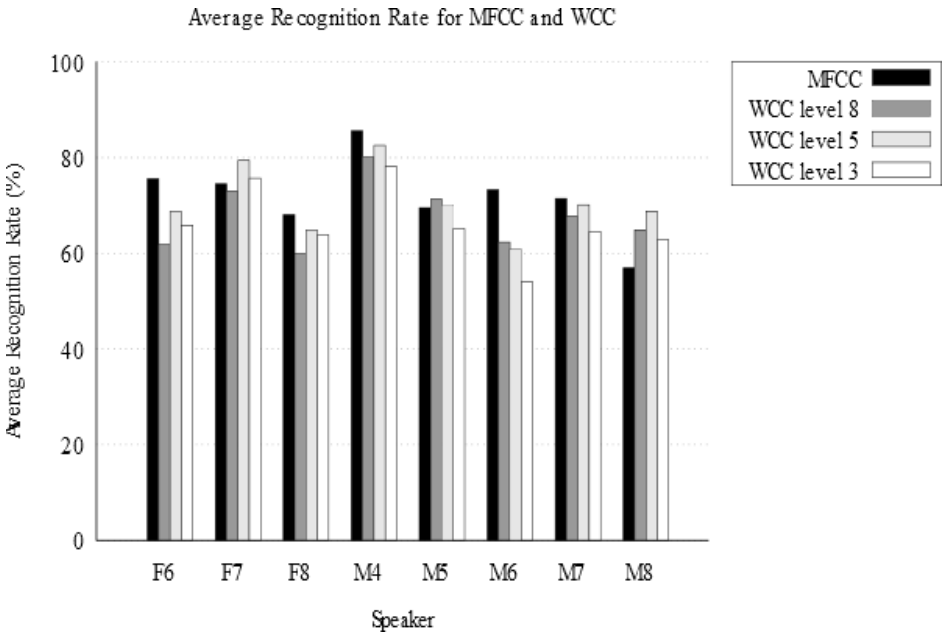

Figure 4. Results for speaker independent tasks

8. Discussion

From the experiments conducted, there are several interesting discussion that can be presented. First, the WCCs had a considerable small feature vector when compared to the MFCC. This is true for the WCCs obtained from level 8, level 5, and level 3. Each WCC feature vector contained only 90, 60, and 40 coefficients respectively. These small values show a considerable amount of feature reduction when compared to the MFCC which uses 800 coefficients.

It is found from the results (Table 3 and Table 4) that level 5 of wavelet decomposition yielded the best results among the WCC. This is true for either speaker dependent or speaker independent taks. However, the MFCC results were still higher in most of the cases.

Results for speaker independent task showed interesting observations. It is found that for speaker independent tests the recognition rate (RR) were quite comparable with each other for the WCCs and MFCC. The recognition for MFCC is 71.89% while WCC level 5 is 70.70%. The percentage difference between the two is only 1.19% which is quite small. The significance of this result stems from the fact that the feature vector for the WCC used only 60 coefficients while the MFCC used 800 coefficients. For speaker dependent task, the percentage difference of WCC at level 5 and MFCC is 3.21% still maintaining a small percentage difference.

9. Conclusion

The proposed WCC feature extraction produced a considerably small feature vector when compared to MFCCs. Although the feature vector for the WCC were small, results showed that with further studies and improvement the WCC can be improved to outperform the MFCCs. To improve the WCCs, the structure of the DWT must be experimented with. In this study, it is shown that the best accuracy for WCC was at level 5 of the DWT decomposition. Results from the speaker independent task shows that the WCC could be improved and well suited for text dependent speech recognition tasks.

Future works include experimenting with different numbers of coefficients, wavelet families and wavelet structure. The WCCs should also be tested under different noise conditions in future experiments to observe its robustness towards noisy speech. We expect by experimenting with these parameters the WCCs may surpass MFCCs in terms of RR.

References

[1] E Pavez and JF Silva. "Analysis and Design of Wavelet-Packet Cepstral Coefficients for Automatic Speech Recognition". *Speech Communication.* 2012; 54: 814-835.
[2] LD Vignolo, DH Milone and HL Rufiner. "Genetic Wavelet Packets for Speech Recognition". *Expert Systems with Applications*. 2012; 40: 2350-2359.
[3] S Davis and P Mermelstein. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*. 1980; 28: 357-366.
[4] M Anusuya and S Katti. "Front End Analysis of Speech Recognition: A Review". *International Journal of Speech Technology*. 2011; 14: 99-145.
[5] R Sarikaya, BL Pellom, and JHL Hansen, "Wavelet Packet Transform Features With Application To Speaker Identification". in *Third IEEE Nordic Signal Processing Symposium*. 1998; 81-84.
[6] P Kumar and M Chandra. "Hybrid of Wavelet and MFCC Features for Speaker Verification". in *World Congress on Information and Communication Technologies (WICT)*. 2011: 1150-1154.
[7] SY Lung. "Improved Wavelet Feature Extraction Using Kernel Analysis for Text Independent Speaker Recognition". *Digital Signal Processing*. 2010; 20: 1400-1407.
[8] K Daqrouq and KY Al Azzawi. "Average Framing Linear Prediction Coding with Wavelet Transform for Text-Independent Speaker Identification System". *Computers and Electrical Engineering*. 2012; 38.
[9] JN Gowdy and Z Tufekci. "Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition". in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2000: 1351-1354.
[10] E Didiot, I Illina, D Fohr, and O Mella. "A Wavelet-Based Parameterization for Speech/Music Discrimination". *Computer Speech and Language*. 2010; 24: 341-357.
[11] JD Wu and BF Lin. "Speaker Identification using Discrete Wavelet Packet Transform Technique with Irregular Decomposition". *Expert Systems with Applications*. 2009; 36: 3136-3143.
[12] O Farooq and S Datta. "Wavelet Based Robust Sub-Band Features for Phoneme Recognition". *IEEE Proceedings Vision Image and Signal Processing*. 2004; 151: 187-193.
[13] MS Deshpande and RS Holambe. "Speaker Identification Using Admissible Wavelet Packet Based Decomposition". *Journal of Signal Processing World Academy of Science Engineering and Technology*. 2010; 6: 20-23.
[14] A Kinney and J Stevens. "Wavelet Packet Cepstral Analysis for Speaker Recognition," in *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*. 2002; 1: 206-209.
[15] FL Sanchez, S Barbon Júnior, LS Vieira, RC Guido, ES Fonseca, PR Scalassara, CD Maciel, JC Pereira, and SH Chen. "Wavelet-Based Cepstrum Calculation". *Journal of Computational and Applied Mathematics*. 2009; 227: 288-293.
[16] JW Picone. "Signal Modeling Techniques in Speech Recognition". *Proceedings of the IEEE*. 1993; 81: 1215-1247.
[17] M. Negnevitsky, *Artificial Intelligence: A Guide to Intelligent Systems*, 2 ed. Harlow UK: Addison Wesley, 2005.
[18] MSH Salam, D Mohamad and SHS Salleh. "Temporal Speech Normalization Methods Comparison in Speech Recognition Using Neural Network". in *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Melacca, Malaysia. 2009.