Robust Automatic Speech Recognition Features using Complex Wavelet Packet Transform Coefficients

Tjong Wan Sen, Bambang Riyanto Trilaksono,
Arry Akhmad Arman & Rila Mandala

Bandung Institute of Technology, Jl. Ganesha 10, Bandung 40132, Indonesia
Email: mpsland@yahoo.com

Abstract. To improve the performance of phoneme based Automatic Speech Recognition (ASR) in noisy environment; we developed a new technique that could add robustness to clean phonemes features. These robust features are obtained from Complex Wavelet Packet Transform (CWPT) coefficients. Since the CWPT coefficients represent all different frequency bands of the input signal, decomposing the input signal into complete CWPT tree would also cover all frequencies involved in recognition process. For time overlapping signals with different frequency contents, e. g. phoneme signal with noises, its CWPT coefficients are the combination of CWPT coefficients of phoneme signal and CWPT coefficients of noises. The CWPT coefficients of phonemes signal would be changed according to frequency components contained in noises. Since the numbers of phonemes in every language are relatively small (limited) and already well known, one could easily derive principal component vectors from clean training dataset using Principal Component Analysis (PCA). These principal component vectors could be used then to add robustness and minimize noises effects in testing phase. Simulation results, using Alpha Numeric 4 (AN4) from Carnegie Mellon University and NOISEX-92 examples from Rice University, showed that this new technique could be used as features extractor that improves the robustness of phoneme based ASR systems in various adverse noisy conditions and still preserves the performance in clean environments.

Keywords: *complex wavelet packet coefficients; feature; noise; phoneme; principal component analysis; robust; speech recognition.*

1 Introduction

ASR is a human-computer interaction that is convenient, cost efficient and natural [1]. These systems will very assistive to us to exploit future intelligent machines that have intelligence and ability just like human being. Unlike any other existing interface, speech based interface provides hands-and-eyes-free operation mode. Efficacy in comprehending and exploiting these systems in the end will be very helpful to human being in improving their ability to solve problems.

Received July 28th, 2009.

Phoneme based ASR systems, at this moment, is expected to be operated in all kinds of different conditions and situations. Some of them are at environments where their circumstances of acoustic could not be controlled. For example roadway, train station, airport, commerce center, canteen, and other public places. In these situations, the system is negatively affected by stochastic and non-stationary acoustic sound sources, or simply noises. Therefore, to achieve a successful deployment, the system must be equipped with a robustness to overcome these problems.

The environmental differences, at the moment when the system was trained (training phase) in studio or laboratory with no noise and when the system is operated (testing phase) in noisy environment, make acoustic feature vectors value obtained by the feature extraction stage no longer the same as the one in its databases. The numbers of these differences are linearly correlated with system recognition accuracy. More acoustic feature vectors differences, generated by environment with noises, would produce higher word error rate (WER) level [2-4]. When the number of wrong recognized words is high enough, the system becomes inefficient and the benefits are lost. At that level of WER, the ASR system performance is no longer acceptable and useless.

The negative effects caused by environment with noises happened especially for ASR systems that use Mel Frequency Cepstral Coefficients (MFCCs, representing state of the art of general ASR systems in this time) [5-6] as its feature extraction method. Although cepstral coefficients perform well when tested on clean environment with data similar to training data, they are not robust to unexpected additive noises or spectrum distortions. With MFCCs feature extraction method, each of additive noise frequency magnitudes (and distortion) is directly added (superposition) to its corresponding true speech signal frequency magnitudes. Thus it gets wrong acoustic feature vectors (different feature values) as a result. Therefore, the frequency contained in speech signal and additive noise (and distortion) from environment play major role in recognition mistakes (more noises frequency involved would generally cause higher WER).

To achieve greater robustness, more robust acoustic features are needed. One way to do it is using CWPT coefficients [7-10]. In this paper, a new feature extraction technique using CWPT coefficients and PCA is developed. This technique is expected to minimize environment noises effect on clean phoneme signal and produces the right acoustic feature vectors. This is motivated by the ability of CWPT coefficients to capture important time and frequency features and PCA to find underlying principal component for every single phoneme in a certain language. CWPT decomposes the input signal into complete frequency bands involved in recognition process and the energy of its all transformation coefficients equal to that of the original input signal, hence preserve all information. These coefficients altogether represent all different frequencies contained in input signal. PCA finds a simplified structure that is hidden in a large complex training data set with big dimension. Those simplified structures are used later to reduce high dimension testing data to a lower dimension one without significantly loosing its information or feature.

In this paper we propose a new technique to obtain robust phoneme based ASR features using CWPT coefficients. In this technique, the features are produced by searching for the underlying structure from clean data set using PCA [11]. This paper is organized as follows. In Section 2 speech signal and phonemes are presented and in Section 3 complex wavelet packet transform is discussed. In Section 4 the new robust features extraction method is presented and in Section 5 the experiments and results are reported. Finally some conclusion and future work are presented in Section 6.

2 Speech Signal and Phonemes

Speech signal produced by human being is a sequence of air pressure changes yielded by human utterance instruments. These unique pressure changes have been trained by each individual in such a manner, for example in the process when babies learn to talk or when adults learn a second language, so that for the same certain meaning in language all human beings could produce the same (or almost the same) sequence of air pressure changes. Otherwise, these utterance instruments are trained over and over again until they produce that sequences of air pressure changes with certain level of similarity or native. In connection with written language, a unique sequence of air pressure changes produced by human being is then mapped individually into its related symbol that represents a related phoneme. Combination of these sequences together form words, which are known as oral language.

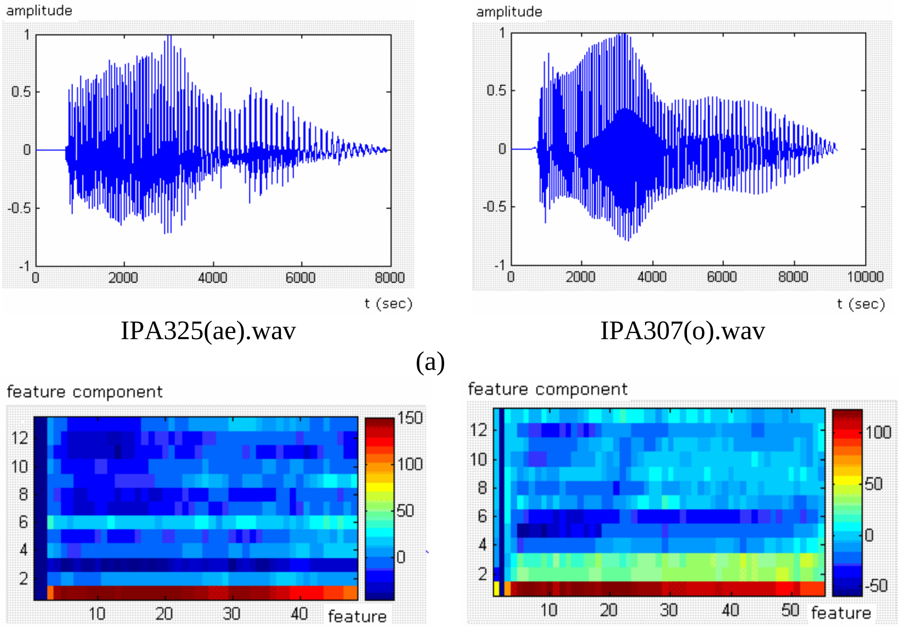
Each phoneme in phonemes set for every language has fundamentally significant differences. These sound differences heavily depend on the characteristic and capability of human speech apparatus. This is why languages in the same region are relatively similar and relatively more different for different regions. Without these significant differences phonemes sequence could not effectively represent different words. These sound differences have been developed in a evolutionary way that from time to time every insignificant phonemes or too similar phonemes causing misrecognition (when they are communicated) are gradually removed. As a result, every phoneme in the same phoneme set is significantly different to each other. This is the key factor for the main idea developed in this noise robust feature extraction technique.

AA
AE
AH
AO
AW
AY
B
CH
D
DH
EH
ER
EY
F
G
HH
IH
IY
JH
K
L
M
N
NG
OW
OY
P
R
S
SH
T
TH
UH
UW
V
W
Y
Z
ZH

THE GREAT YOUNG ACTOR
DH AH – G R EY T – Y AH NG – AE K T ER

Figure 1 Thirty nine English phonemes and its written language.

For example, English has 39 phonemes. With only 39 different phonemes, computation needed to exhaustively search one of them could be easily performed by most of today Personal Computer processing power. Figure 1 shows example of English phonemes and its written language.



IPA325(ae).wav IPA307(o).wav

(a)

IPA325(ae).mfcc IPA307(o).mfcc

(b)

Figure 2 (a) Two phonemes from International Phonetic Alphabet database. (b) Its related MFCCs.

In the form of sampling and quantization process result, each phoneme is represented as a sequence of specific value at the specific time (time-series). These values are determined by it's correspond signal amplitude in time domain. These values are inherently different for each phoneme because that is what the phonemes are about; each phoneme has to be different to effectively represent specific different meaning. Figure 2a shows two examples of different time series that represent different phoneme (vowel) from International Phonetic Association; its MFCCs are shown by Figure 2b.

3 Complex Wavelet Packet Transform

Fourier Transform is widely used and proven to be an important tool in signal analysis. By transforming into frequency (Fourier) domain one could easily separate different frequency contents from every time overlapping signal. The biggest drawback of the Fourier Transform is the lack of time information. Fourier transform knew what frequencies are contained in a signal, but did not know the time when they actually present. Time information is very important especially for time varying signals like speech. This drawback is handled by Short Time Fourier Transform (STFT). STFT provides not only frequency information but in the same time also time information. STFT uses small time frame (window) to relatively reserve time information. The smaller time frame gives higher time resolution, but lower frequency resolution. In opposite, wider time frame gives lower time resolution but higher frequency resolution, therefore there is a time-frequency resolution trade off to consider.

This time-frequency trade off is further exploited by Wavelet Transform which introduce non uniform time frame usage [12]. Different time frames are used for different frequency bands. Usually, for high frequencies band one could get lower frequency resolution and higher time resolution and for low frequency band one gets higher frequency resolution and lower time resolution. This way, one could take advantage by freely choosing the appropriate time frame that is most suitable for the signal under consideration.

The Wavelet Transform is a way to represent a time domain signal into time-frequency domain by projecting the signal into wavelet basis functions that consist of scaling and wavelet functions, φ and ψ. Different input signal would produce specific different set of coefficients and these specific set of coefficients, with perfect reconstruction mechanism, would produce the same specific input signal. There are two parameters for scaling and wavelet function called scaling coefficient (j) and shifting coefficient (k) as shown in equations (1) and (2).

$$\phi_{j,k}(t) = 2^{j/2}\phi\left(2^j t - k\right) \tag{1}$$

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right) \tag{2}$$

Bigger *j* value would produce shorter scaling function and increase the time resolution, but decrease frequency resolution; and the value of *k* from smallest to highest would traverse the input signal from beginning to end.

The Wavelet Transform coefficients are computed using decomposition shown in equation (3).

$$f(t) = \sum_k c_j(k)\,\phi_{j,k}(t) + \sum_j \sum_k d_j(k)\,\psi_{j,k}(t) \tag{3}$$

Where $c_j\ (k)$ are the scaling function coefficients at level *j*, and $d_j\ (k)$ are the wavelet function coefficients at that level from $k = 0$ to $j^2 - 1$. In this way, the same input signal must produce the same specific coefficients set. Any changes in input signal would definitely reflect in changes in coefficients set. In another word, the coefficients set perfectly represent the input signal. Thus, wavelet coefficients could be used for recognition features.

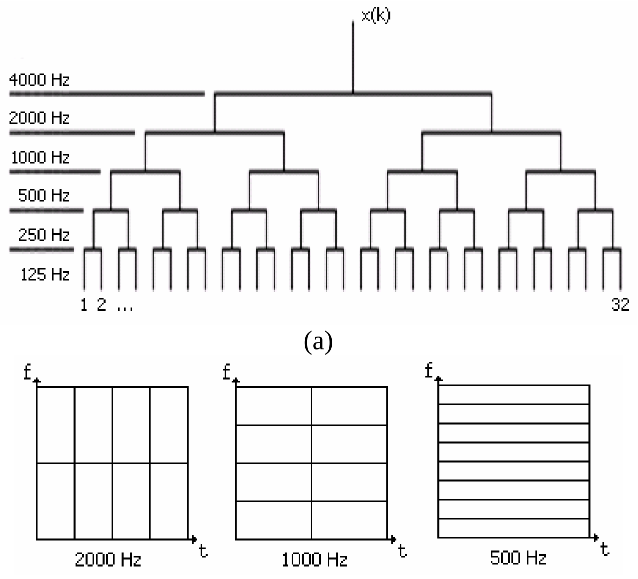
Wavelet coefficients computations need a lot of machine cycles and need to be simplified. One way to efficiently calculating the Wavelet Transform coefficients is using conjugate mirror filter banks [8-9]. In the form of discrete samples input, output coefficients from low pass filter ($c_j\ (k)$) corresponds to a projection of the input signal on to the scaling functions; and output coefficients of high pass filter ($d_j\ (k)$) corresponds to a projection of the input signal on to the wavelet functions as shown in equations (4) and (5). Coefficients computation using filtering operation and down-sampling by a factor 2 is more convenient and now is widely used.

$$c_j(k) = \sum_m g_0(m - 2k)\ c_{j+1}(m) \tag{4}$$

$$d_j(k) = \sum_m g_1(m - 2k)\ c_{j+1}(m) \tag{5}$$

The g_0 and g_1 are filters coefficients set that correspond to scaling and wavelet functions respectively.

Non uniform time frame in Wavelet Transform is extended further into Wavelet Packet Transform (WPT) in order to get more complete decomposition level. This decomposition also expands the detail tree nodes not just approximation. With WPT, it is possible to freely choose time-frequency resolution trade off (basis tree) to find the best suitable combination in respect to the processed signal. Different time-frequency plane combination produces different frequency bands partitions. The numbers of frequency bands depend on the decomposition depth level but the total energy for a specific input signal is the same for every level. Thus decomposing the input signal into complete WPT tree would always cover all frequencies involved in recognition process too.



(a)

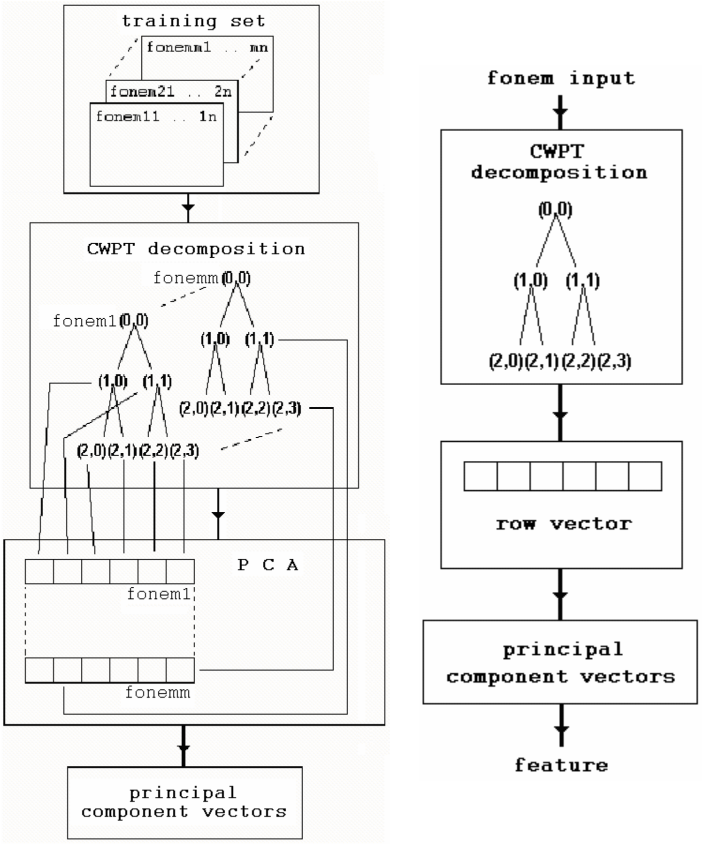
(b)

Figure 3 (a) The WPT splits a signal into frequency bands. (b) Different Time-Frequency Plan from the input signal.

In order to get richer frequency response characteristic, another extension to WPT is Complex Wavelet Packet Transform (CWPT) [8-9]. One implementation of this transform is seen in Dual Tree Complex Wavelet Packet Transform (DTCWPT). Usually, DTCWPT is used for images or other multidimensional signal. Using dual tree in DTCWPT, one could get two complete trees (sets of coefficients) with slightly different frequency responses. These two trees, one produces real part and the other produce imaginary part, computed in parallel to save time. Though the filters used in both trees are different, basically the main mechanism is the same. With DTCWPT one could decompose a time overlapping signal into two complete trees and obtain all frequencies contained in it in two different coefficients sets. Thus, decomposing any input signal into two complete DTCWPT trees would also cover all frequencies involved in the process under consideration, such as in speech recognition.

4 The Proposed Method

Figure 4 shows the mechanism of robust phoneme based ASR features extraction technique using CWPT coefficients and PCA. There are two mechanisms, one for training phase (4.a) and the other for testing phase (4.b).



(a) Training phase. (b) Testing phase.

Figure 4 The proposed method mechanisms.

The training phase begins with decomposition using CWPT of all phonemes from clean training dataset. The purpose is to obtain sets of coefficients that uniquely represent each phoneme used in training dataset. To provide better generalization, each phoneme is represented by a number of samples with variation, for example, the same phoneme but spoken by different person or spoken by the same person but at different times (or extracted from several different words). Each phoneme produces two complete tree sets (real and imaginary) with theirs coefficients. The number of coefficients produced by decomposition depends on the decomposition level. To cover more detail frequency, one could use more decomposition level.

After the coefficients are ready, the next step is forming the row vectors. Each row vector is formed by arranging all nodes from each tree couple into one row. Here, we begin to arrange nodes from real tree and follow up by nodes from imaginary tree. These row vectors all together form one big matrix from which we derive principal component vectors using PCA. The principal vector component represents the underlying structure of training dataset or features which have much lower dimensionality.

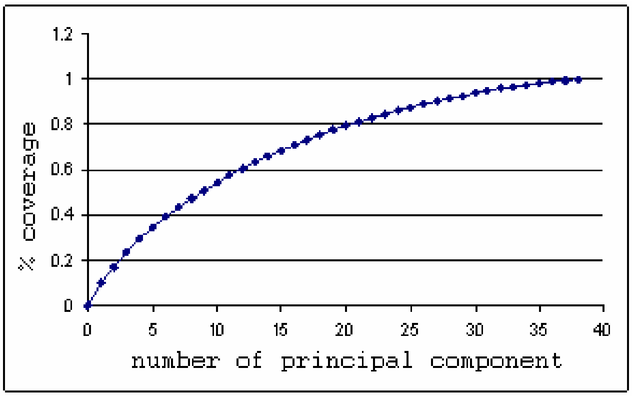
The testing phase also begins with decomposition of an input phoneme using CWPT. With the same mechanism, nodes from real and imaginary tree are also arranged to form row vector. This row vector is processed using the matrix multiplying process with the principal component vectors produced in training phase. As a result, the phoneme based ASR feature is obtained.

5 Experiments and Results

We implemented the proposed method in MATLAB 7 and simulated experiments on phonemes set extracted from clean speech and noises database. For clean speech database, we use AN4 database from CMU Sphinx Group. The AN4 database was recorded internally at Carnegie Mellon University circa 1991, and consisted of utterances from different persons, uttered by man and woman native speakers. The utterances are in PCM format, mono, 16 bit quantization and 16 kHz sampling rate. And for noises database, we use examples taken from NOISEX-92 database from Rice University Digital Signal Processing Group. The noises are in .wav format, mono, 16 bit quantization and 19.98 kHz sampling rate. Both speech and noises sampling rate are converted to 16 kHz in these experiments.

The phonemes set consists of 39 English phonemes or 39 classes. All phonemes are extracted manually from the particular utterance to match the desired number of samples for CWPT decomposition (dyadic). Each phoneme is represented by 1024 samples (64 ms). Phonemes larger than 1024 samples are truncated about the center of the phoneme. The noises set consist of 5 types. Babble (representing chat in public area), hfchannel (representing communication devices disturbances), pink (representing colored noises), Volvo (representing noises inside car cabin) and white noise. These noises are also truncated randomly to 1024 samples.

Using 39 English phonemes (representing 39 classes), we derive principal component vectors using mechanism explained in the training phase of the proposed method. Each class is represented by 10 pieces of truncated utterances from database with each truncated utterance is spoken by different person or by the same person at different time. The results are shown in Figure 5.



No.	Number of principal components	% coverage
1	14	65%
2	27	90%
3	38	100%

Figure 5 Principal component vectors coverage.

To measure separation between pairs of classes, we compute standardized Euclidian distance. Greater standardized Euclidian distance indicates greater separation. The measures are shown by Table 1.

Table 1 Standardized Euclidian distance measure.

Distance	From clean data set	From noisy data set
Minimum	2.9397	1.2253
Average	10.2700	10.7600
Maximum	19.1794	23.6831

Standardized Euclidian distances from clean dataset are shown by column 2. The minimum distance is 2.9397 and the maximum distance is 19.1794 with average distance 10.2700. With noisy dataset, shown by column 3, the minimum distance is decreasing 58.32% to 1.2253 and the maximum is increasing 23.48% to 23.6831 with average distance 10.7600. These noisy dataset distances are computed using combination of clean data set with 5 types of NOISEX-92 noises with SNR 0 dB. The 0 dB SNR could be considered as bad acoustic environment for ASR.

In the next experiment, we performed classification of 39 phonemes using one against one Support Vector Machine (SVM) [13]. First we train SVM using features derived from clean dataset to produce models. There are 3 types of feature produced by the method proposed. These features are produced by 38, 27 and 14 principal component vectors. We use combination of 5 types of noises with 0, -3 and -7 dB SNR each and calculate the number of misclassification. The result is shown by Figure 6.

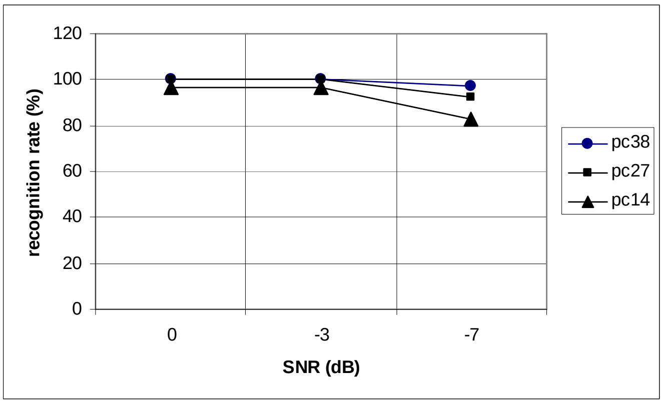


Figure 6 Recognition rate using SVM classification.

The recognition rate is defined by the number of misclassification divided by the total classification. With 38 and 27 principal component vectors, the recognition rate for 0 dan -3 dB SNR achieve 100% and for -7 dB SNR achieve 97.44% and 96.92% in average respectively. For 14 principal component vectors, the recognition rates never achieve 100%. Only 96.41% in average for 0 and -3 dB SNR and drop to 82.56% for -7 dB SNR.

6 Conclusions and Future Works

Through this works, robust automatic speech recognition features extraction method using complex wavelet packet transform coefficients and PCA has been developed. We tested this technique using phonemes from AN4 and NOISEX-92 database with several SNR combinations. The separation between classes experiment shows that the minimum distance is decreasing when noises are present. But the experiments using SVM classifier show that these features could achieve the recognition rate above 95% except for features produced by 14 principal component vectors at -7 dB SNR. This new technique adds robustness to phoneme based ASR feature when operated in noisy environment and still preserves the performance when operated in clean environments.

There are two further works that could be done. First is to test this technique using more wavelet basis function or develop a new one. This is important since different wavelet basis function produces different coefficients values. Second is to train the SVM models by adding various noises into training dataset to investigate the impact on recognition rate.

References

[1] Juang, B.H. & Furui, S., *Automatic Recognition and Understanding of Spoken Language – A First Step Toward Natural Human-Machine Communication*, Proceeding of the IEEE, **8**, pp. 1142-1165, 2000.

[2] Pallett, D., *DARPA HUB-4 rep.*, National Institute of Science and Technology, 1999.

[3] Hanai, N. & Stern, R.M., *Robust speech recognition in the automobile,* International Conference on Spoken Language Processing, **3**, pp. 1339-1342, 1994.

[4] Huerta, Juan M., *Speech Recognition in Mobile Environments*, Carnegie Mellon Universit, April 2000.

[5] Davis, Steven B. & Mermelstein, P., *Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Transaction on Acoustic, Speech and Signal Processing, **28**, pp. 357-366, 1980.

[6] Ganchev, T., Fakotakis, N. & Kokkinakis, G., Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, Proceedings of the SPECOM-2005, 1, 191-194, 2005.

[7] Hai, J., Meng, J.E. & Yang, G., *Feature Extraction Using Wavelet Packets Strategy*, IEEE Conference on Decision and Control, **5**, pp. 4517-4520, 2003

[8] Kingsbury, N.G., *Complex Wavelets for Shift Invariant Analysis and Filtering of Signals*, Journal of Applied and Computational Harmonic Analysis, **3**, pp. 234-253, 2001.

[9] Selesnick, I.W., Baraniuk, R.G. & Kingsbury, N., *The Dual-Tree Complex Wavelet Transform - A Coherent Framework for Multiscale Signal and Image Processing*, IEEE Signal Processing Magazine, **22**(6), pp. 123-151, 2005.

[10] Bayram, I. & Selesnick, I.W., *On the Dual-Tree Complex Wavelet Packet and M-Band Transforms*, IEEE Transactions on Signal Processing, **56**(6), pp. 2298-2310, 2008.

[11] Duda, R.O., Hart, P.E. & Stork, D.G., *Pattern Classification 2nd Edition*,Wiley-Interscience, 2000.

[12] Rioul, O. & Vetterli, M. *Wavelets and Signal Processing*, IEEE Signal Processing Magazine, **4**, pp. 14-38, 1991.

[13] Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery, **2**, pp. 121–167, 1998.