Several New DWT-Based Methods for Noise-Robust Speech Recognition

Jeih-Weih Hung, Hao-Teng Fan, and Syu-Siang Wang

***Abstract*—This paper proposes three novel noise robustness techniques for speech recognition based on discrete wavelet transform (DWT), which are wavelet filter cepstral coefficients (WFCCs), sub-band power normalization (SBPN), and lowpass filtering plus zero interpolation (LFZI). According to our experiments, the proposed WFCC is found to provide a more robust c0 (the zeroth ceptral coefficient) for speech recognition, and with the proper integration of WFCCs and the conventional MFCCs, the resulting compound features can enhance the recognition accuracy. Second, the SBPN procedure is found to reduce the power mismatch within each modulation spectral sub-band, and thus to improve the recognition accuracy significantly. Finally, the third technique, LFZI, can reduce the storage space for speech features, while it is still helpful in speech recognition under noisy conditions.**

***Index Terms*—Discrete wavelet tranSform, wavelet filter cepstral coefficients, sub-band power normalization, lowpass filtering and zero interpolation, speech recognition.**

I. Introduction

The conventional mel-frequency cepstral coefficients (MFCC) [1] have been one of the most widely used speech features for speech recognition over many years. In deriving the MFCC, the short-time Fourier transform (STFT) is applied. However, due to its time-frequency properties, STFT is actually not very suitable for analyzing a non-stationary signal like speech [2], which implies the resulting MFCC is not always optimal for representing the speech signal and possibly provides less recognition accuracy. One way to partially solve the problem is to apply the multi-resolution property in time and frequency domain, and the multi-resolution goal is achieved by replacing STFT with wavelet transform [3]. Unlike the Fourier transform, the finite-length basis functions [4] help wavelet transform analyze the non-stationary signal with better transformable ability. Although wavelet transform has a better performance in analyzing the nonperiodic signal, STFT performs better for presenting the periodic signal [5]. Thus in this paper we create the compound feature to integrate them both and find it is helpful for speech recognition.

The environmental mismatch caused by additive noise and/or channel distortions often degrades the performance of a speech recognition system. To overcome this problem, researchers have proposed many speech enhancement or robustness techniques to enhance the speech or alleviate the effect of noise. We find that the wavelet analysis can be also applied to constructing the noise-robust speech features. In the past research of our lab, the wavelet transform was used in the temporal speech feature stream and good recognition performance can be achieved [6]. Therefore, in this paper we follow this direction to provide more noise-robust features with wavelet transform, and come up with two novel robustness methods, sub-band power normalization (SBPN) and lowpass filtering plus zero interpolation (LFZI).

The remainder of this paper is organized as follows: Section II briefly introduces the discrete wavelet transform (DWT). Then we present three DWT-based noise robustness methods in Section III. Section V contains the experimental results together with the discussions. Finally, a brief concluding remark is given in Section VI.

II. Discrete Wavelet Transform

Here, we make a brief introduction of discrete wavelet transform. Consider a signal $f[n]$ that is decomposed by a discrete wavelet transform with the scaling ($\phi_{j,k}[n]$) and wavelet ($\psi_{j,k}[n]$) basis functions [7]:

$$f[n]=\sum_{j}\sum_{k}a_{j,k}\phi_{j,k}[n]+\sum_{j}\sum_{k}d_{j,k}\psi_{j,k}[n] \qquad (1)$$

Based on eq. (1), $f[n]$ is decomposed into $a_{j,k}$ and $d_{j,k}$, which are considered as the approximation (low-pass) part and the detail (high-pass) part, respectively.

In practice, the implementation of the discrete wavelet transform is sometimes accomplished with sequential filtering and down-sampling, as presented in Fig. 1. In this figure, $h[-n]$ and $g[-n]$ are the low-pass and high-pass filters respectively, followed by a down-sampling process. According to Fig. 1, the signal $a_{j+l,k}$ is first decomposed into the approximation and detail parts with filtering and down-sampling procedures, and then the approximation sequence is decomposed again with the same process.

In this paper, we viewed the discrete wavelet transform as a filtering process and develop several noise-robustness methods.

III. Several Novel Proposed Techniques

Based on DWT, we present three novel methods to improve the original MFCC in its recognition accuracy under noisy environments.

Manuscript received August 10, 2012; revised September 24,2012.

Jeih-Weih Hung, Hao-Teng Fan and Syu-Siang Wang are with the Department of electrical engineering, national Chi Nan University, Taiwan (e-mail: jwhung@ncnu.edu.tw, s99323904@ncnu.edu.tw).

DOI: 10.7763/IJIMT.2012.V3.295

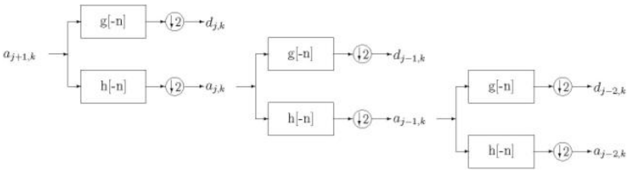

Fig. 1. Three-level discrete wavelet transform

A. Wavelet Filter Cepstral Coefficients (WFCCs)

In Section II, the DWT procedure recursively decomposes the approximation sequence. However, the wavelet packet transform (WPT) [8], a generalized version of DWT, is applied in the proposed WFCC construction here. Unlike DWT which only decomposes the approximation part, WPT also decomposes the detail part, and thus it can perform the frequency division with more flexibility. The flowchart of the proposed WFCC construction process is shown in Fig. 2. We use wavelet packet transform for frequency division to create a new filter bank. This new wavelet filter bank is used to replace the mel-filter bank in MFCC creating process, and we name the corresponding new features as wavelet-based filter cepstral coefficients (WFCC).

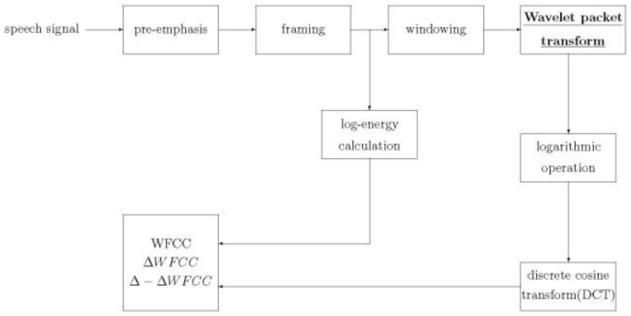

Fig. 2. The flowchart of WFCC feature extraction

The tree structure of the wavelet packet decomposition in creating WFCC for a given frame signal is shown in Fig. 3. The lower frequency range [0, 2000 Hz] is divided into equal-width sub-bands, while the bandwidths of the sub-bands get wider monotonically as the frequency increases in the higher frequency range [2000 Hz, 4000 Hz] (Here, the sampling rate is set to 8000 Hz).

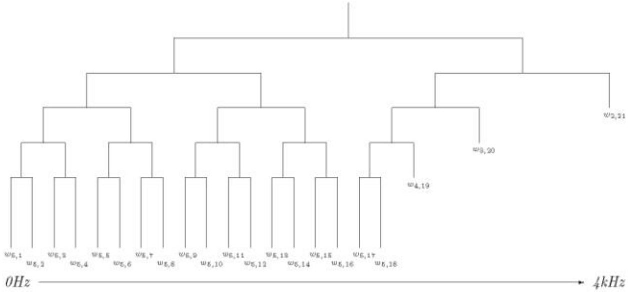

Fig. 3. The tree structure of wavelet packet transform in WFCC construction

There are 21 sub-bands finally created. This arrangement somewhat simulates the perceptual characteristics of human beings.

After processing the frame signal $x[n]$ with the above sub-band decomposition, as shown in Fig. 3, the output sequences $x[n]$, l=1,2,…,21, in the l-th sub-band are further applied to eq. (2) to produce 21 values, which present the sub-band signal power features for that frame.

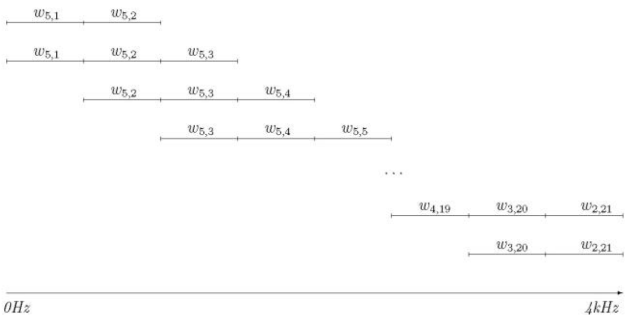

Fig. 4. The sub-band arrangement in the filter-bank used for WFCC

$$S_l = \begin{cases} \sum_n (x_l[n])^2 + \sum_n (x_{l+1}[n])^2, & l = 1 \\ \sum_n (x_{l-1}[n])^2 + \sum_n (x_l[n])^2 + \sum_n (x_{l+1}[n])^2, & l = 2,\cdots,20 \\ \sum_n (x_{l-1}[n])^2 + \sum_n (x_l[n])^2, & l = 21 \end{cases} \tag{2}$$

From eq. (2), we see that an overlapping filter structure as shown in Fig. 4 is used, which is similar to the filter-bank arrangement in deriving MFCC, and it produces the signal power values of the overlapped filter outputs. Except for the wavelet package transform for finally creating 21 power values, the remaining procedures of WFCC construction are the same as those of MFCC, as shown in Fig. 2.

B. Sub-Band Power Normalization

In this new method, we attempt to normalize the “power" of the sub-band features in the temporal domain of MFCC feature sequence to alleviate the noise effect. The discrete wavelet transform (DWT) is applied in the temporal domain to obtain the (modulation spectral) sub-band features. According to [9], different modulation frequency components possess different importance for speech recognition. General speaking, the most significant information for speech recognition is located between 1 Hz and 16 Hz in modulation frequency, and the most important part is around 4 Hz. Based on this observation, to split the modulation band into several ones with an unequal bandwidth will help process the important sub-band features individually.

The proposed sub-band power normalization (SBPN) is to normalize the power of each sub-band (in modulation frequency) temporal feature sequence for the utterances in the training and testing sets. The detailed procedure of SBPN process is depicted in Fig. 5.

We consider the mel frequency cepstral coefficients (MFCC) for speech recognition. An MFCC feature stream for an utterance is represented as:

$$\{c^m[n] \mid 0 \le n \le N-1, 0 \le m \le M-1\}, \tag{3}$$

where n and m are the frame index and feature index, respectively, and N and M are respectively the number of frames and the number of features in a frame.

In SBPN, each temporal feature sequence $\{\ c^m[n] \mid 0 \le n \le N-1\ \}$ is first divided into L sub-band sequences by using an L-level discrete wavelet transform (DWT). Therefore the l-th sub-band sequence, represented by { $c_l^m[n]$ }, is roughly within the following modulation frequency range:

$$\begin{cases} [0, \frac{1}{2^{L-1}}(\frac{F_s}{2})], & l=1 \\ [\frac{2^{l-2}}{2^{L-1}}(\frac{F_s}{2}), \frac{2^{l-1}}{2^{L-1}}(\frac{F_s}{2})], & 2 \le l \le L \end{cases}, \tag{4}$$

where F_s (Hz) is the frame sampling rate. Therefore, through DWT, we split the entire frequency band [0, $\frac{F_s}{2}$ Hz] into L sub-bands with unequal bandwidth.

Next, each sub-band sequence is updated for power normalization according to the following equation:

$$\tilde{c}_l^m[n] = c_l^m[n] \times \sqrt{\frac{P_{target\_l}^m}{P_{single\_l}^m}}, \tag{5}$$

where the $\tilde{c}_l^m[n]$ is the new speech feature, and $P_{target+_+l}^m$ and $P_{single+_+l}^m$ are the target power (obtained from the clean sub-band features in the training set) and the power of the currently processed $c_l^m[n]$, respectively.

Finally, all the updated sub-band sequences are used together to reconstruct the new full-band (temporal) sequence with an L-level inverse discrete wavelet transform (IDWT).

C. Lowpass Filtering and Zero Interpolation (LFZI)

According to the original DWT decomposition process, the low-pass and high-pass filters are first applied to the input data, and a down-sampling procedure is applied to the two filter outputs. As stated in [9], the main speech component (1 Hz□16 Hz) is just within the front half modulation spectrum. Therefore, if the original speech frame rate is 100 Hz, then after processing the feature frame sequence with a one-level DWT, the low-pass filter output (before down-sampling) is roughly within [0, 25 Hz], preserves the speech information, while the high-pass filter output seems less helpful in speech recognition. In addition, the down-sampling process doubles the bandwidth of filtered sequence.

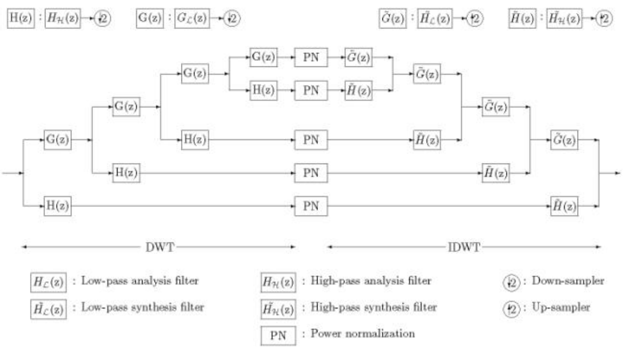
H(z) : $H_{\mathcal{H}}(z)$ → ↓2
G(z) : $G_{\mathcal{L}}(z)$ → ↓2
$\tilde{G}(z)$: $\tilde{H}_{\mathcal{L}}(z)$ → ↑2
$\tilde{H}(z)$: $\tilde{H}_{\mathcal{H}}(z)$ → ↑2
G(z)
H(z)
PN
$\tilde{G}(z)$
$\tilde{H}(z)$
DWT
IDWT
$H_{\mathcal{L}}(z)$: Low-pass analysis filter
$\tilde{H}_{\mathcal{L}}(z)$: Low-pass synthesis filter
$H_{\mathcal{H}}(z)$: High-pass analysis filter
$\tilde{H}_{\mathcal{H}}(z)$: High-pass synthesis filter
PN : Power normalization
↓2 : Down-sampler
↑2 : Up-sampler

Fig. 5. The SBPN process with four sub-bands

Accordingly, the approximation feature sequence via DWT (the lowpass-filtered and down-sampled sequence) may perform better than the original sequence since the irrelative components (high frequency parts) are removed. However, due to down-sampling, the approximation sequence is a half of the original feature sequence in length. The feature length reduction is found to be defective since there will be insufficient data for training accurate acoustic models.

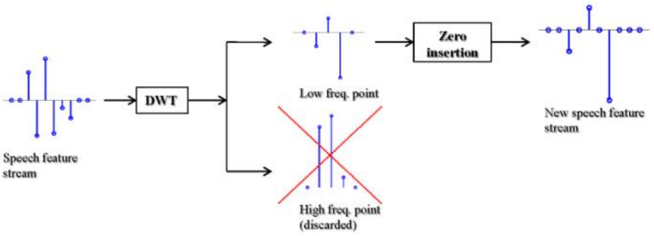

Fig. 6. The detailed procedure of the LFZI technique

In order to solve or alleviate the above problem of data insufficiency, we apply a zero interpolation (adding one zero between each sample) in the approximation sequence to make the resulting new sequence roughly equal the original sequence in length. We name the above process as “lowpass filtering and zero interpolation”, abbreviated as “LFZI”. The procedure of the LFZI method is depicted in Fig. 6.

IV. Experimental Setup

We use the AURORA-2 database [10], which is widely used for evaluating robustness algorithms under noisy conditions. For the recognition environment, three different subsets are defined: Test Sets A, B and C. Speech signals in Test Sets A and B are affected by additive noise (in Set A, the noise types are subway, babble, car and exhibition; and in Set B, they are restaurant, street, airport and train station), and speech signals in Test Set C is affected by additive noise and channel effects (subway or street noise together with an MIRS channel mismatch). Each noise instance is added to the clean speech at six SNR levels (ranging from 20 dB to -5 dB). Each utterance in the clean training set and three noise-corrupted testing sets is first converted into a sequence of 13-dimensional MFCCs ($c0$ □ $c12$) and the same dimensional WFCCs. The frame length and frame shift are set to 32 ms and 10 ms, respectively.

The Hidden Markov Model Tool kit (HTK) [11] is used for the training and recognition process. The resulting acoustic models include 11 digit models (zero, one, two, three, four, five, six, seven, eight, nine and oh) and a silence model. Each digit model contains 16 states and 20 Gaussian mixtures per state.

V. Experiment Results and Analyses

In this subsection, we will separately present the recognition performance achieved by our three novel methods and give the corresponding discussions.

Since WFCC alone do not perform very well (as will be shown as Feature Set iiiv in Table 2), we partition the original 13 cepstral features into two sets: { $c0$ } and { c1, $c2$, ..., $c12$ }, and then we have seven compound feature sets as listed in Table 1, and they are depicted in Fig. 7 for a clearer description, where g_1, g_2, g_3 and g_4 are the weights for $c0$ and the set { $c1$, $c2$, ..., $c12$ } of WFCC and MFCC, respectively. Besides, Sets VIII and IX in Table 2 presents using WFCC (c0, c1□c12) alone and using MFCC (c0, c1□c12) alone, respectively.

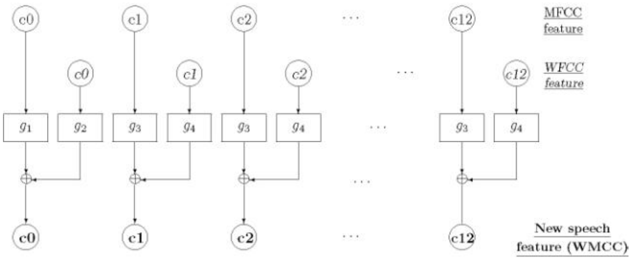

Fig. 7. The combination of WFCC and MFCC

A. Recognition Results for Combinations of WFCC and MFCC

TABLE I: SEVEN FORMS OF COMPOUND FEATURES

Set	c0	c1~c12	g1	g2	g3	g4
I	WFCC	MFCC	0	1	1	0
II	WFCC+MFCC	MFCC	1	1	1	0
III	WFCC+MFCC	WFCC+MFCC	1	1	1	1
IV	WFCC+MFCC	WFCC	1	1	0	1
V	MFCC	WFCC	1	0	0	1
VI	MFCC	WFCC+MFCC	1	0	1	1
VII	WFCC	WFCC+MFCC	0	1	1	1

Table II represents the recognition rates of these compound features, conventional MFCC and the WFCC. From this table, we have the following observations:

1) Based on Table 2, WFCC (Set VIII) performs worse than MFCC (Set IX). However, most of the compound features could improve the recognition accuracy, which implies WFCC provides some information for recognition that MFCC somewhat lacks.
2) The recognition accuracy always gets better when we use the WFCC-based c0 alone or combine it with MFCC-based c0 (feature sets I, II, III and VII). However, for MFCC-based c0 the situation is converse. This phenomenon implies for the component c0, WFCC is better than MFCC.
3) On the other hand, if the features c1 □ c12 are completely or partially from MFCC rather than WFCC alone, the recognition accuracy is significantly improved. This shows that MFCC deriving process is more capable of providing better nonzero quefrency ($c1$□$c12$, corresponding to the variation of the log-spectrum) components.

TABLE II: SUMMARY OF THE AVERAGED RECOGNITION ACCURACY (%) FOR ALL TYPES OF COMPOUND FEATURES

Set	Set A	Set B	Set C	Average
I	73.08	69.70	79.47	74.08
II	72.66	69.34	79.04	73.68
III	73.06	69.65	79.67	74.13
IV	67.42	64.02	72.75	68.06
V	66.63	63.06	71.97	67.22
VI	72.87	70.22	79.08	74.06
VII	74.35	71.04	80.13	75.17
VIII	68.15	64.44	72.35	68.31
IX	71.13	67.55	78.53	72.40

B. Recognition Results for Sub-band Power Normalization

Table III represents the recognition rates from of the SBPN process. From this table, we have several observations as follows:

1) Since in CMS, only one statistic (mean) is normalized, which is similar to our proposed SBPN that also processes only one statistic (power), here we additionally show the recognition accuracy from CMS. All the power normalization methods, including full-band power normalization (FBPN) and SBPN with different L, outperform the baseline and CMS. Therefore, we show that normalizing the power for feature streams is indeed helpful for improving the recognition accuracy.
2) The results of SBPN with L=2, 3, and 4 are very close to those of FBPN. However, when the number of sub-bands, L, in SBPN is greater than 4, a significant accuracy improvement can be achieved, and the best relative error reduction is 55.22% when L=6. Therefore, the results indicate that performing modulation frequency division in the proposed SBPN can effectively improve the recognition accuracy.
3) Finally, we find that increasing L from 6 to 7 results in lower accuracy rates. This is possibly due to the over-normalization effect. Besides, due to the down-sampling of DWT, higher level decomposition will give relatively short-length signals, which make the corresponding power estimate less accurate and thus lower the performance of SBPN.

TABLE III: THE AVERAGED RECOGNITION ACCURACY (%) FOR SBPN WITH DIFFERENT NUMBER OF SUB-BANDS.

	Set A	Set B	Set C	Average
FBPN	80.69	84.00	79.31	81.41
SBPN(2)	80.99	84.21	80.01	81.74
SBPN(3)	80.22	83.09	79.71	81.01
SBPN(4)	80.69	83.47	80.39	81.52
SBPN(5)	84.85	86.58	85.02	85.48
SBPN(6)	87.19	88.17	87.58	87.65
SBPN(7)	85.95	86.95	86.49	86.44
CMS	79.01	82.30	79.34	80.22
MFCC	71.13	67.55	78.53	72.40

TABLE IV: THE AVERAGE RECOGNITION ACCURACY (%) FOR THE VARIOUS METHODS AND THE BASELINE.

	Set A	Set B	Set C	Average
Approximate part of DWT-processed MFCC	45.78	46.71	46.10	46.20
DWT lowpass-filtered MFCC	73.02	70.03	79.90	74.32
LFZI	78.34	80.21	77.15	78.57
MFCC baseline	71.13	67.55	78.53	72.40

C. Recognition Results for Lowpass Filtering and Zero Interpolation

Table IV gives the performance of our third new DWT-based method: low-pass filtering and zero interpolation (LFZI). Here we also represent the results of two other methods for comparison, one of which uses the approximation part of the DWT-processed MFCCs (without zero interpolation) as the new speech features, and the other uses the low-pass filtered MFCCs in the DWT procedure (without down-sampling and zero interpolation). According to Table IV, we see that:

1) Our proposed LFZI provides 6.17% accuracy improvement over the baseline. In addition, LFZI performs the best among the three methods listed in this table.
2) The method that directly uses the approximate part of DWT-processed MFCCs has the worst performance, which is possibly due to the insufficient training data, causing the inaccurate acoustic models.
3) Although the low-pass filter-processed sequence has the same length as the original MFCC sequence and gets better accuracy rates than the MFCC baseline, it is worse than our proposed LFZI method. The results reveal that in addition to low-pass filtering, the down-sampling and zero-insertion processes in LFZI indeed help improve the noise robustness of MFCC features.

VI. Conclusion

In this paper, we propose three noise-robustness techniques. First, the new WFCC construction process gives a better $c0$ feature while MFCCs have superior $c1 \sqcup c12$ features, which makes the compound features behave better than WFCC alone and MFCC alone. Second, the sub-band power normalization (SBPN) attempts to normalize the power of each sub-band in the temporal domain. Finally, LFZI reserves the more important modulation spectral portions in the feature sequence and reduces the storage space for speech features simultaneously. Despite the simplicity in implementation, the proposed SBPN and LFZI significantly improve the recognition accuracy under noisy situations.

References

[1] X. Huang, A. Acero， and H. W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice Hall PTR, 2001.

[2] R. Modic, B. Lindberg, and B. Petek, “Comparative wavelet and MFCC speech recognition experiments on the slovenian and English speechDat2,” *ISCA Non-Linear Speech Processing (NOLISP)*, vol.16, 2003.

[3] R. Sarikaya and J. H. L. Hansen, “High resolution speech feature parametrization for monophone-based stressed speech recognition,” *IEEE Signal Processing Letters*, vol. 7, no. 7, pp. 182-185, 2000.

[4] F. B. Tuteur, “Wavelet transformations in signal detection,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1435-1438, 1988.

[5] O. Farooq and S. Datta, “Mel filter-like admissible wavelet packet structure for speech recognition,” *IEEE Signal Processing Letters*, vol. 8, no. 7, pp. 196-198, 2001.

[6] H. Fan and J. Hung, “Sub-band feature statistics normaliztion techniques based on discrete wavelet transform for robust speech recognition,” *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 806-809, 2009.

[7] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice-Hall PTR, 1995.

[8] W. Chong and J. Kim, “Speech and image compressions by DCT, wavelet, and wavelet packet,” *International Conference on Information, Communications and Signal Processing (ICICS)*, vol. 3, pp. 1353-1357, 1997.

[9] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the importance of various modulation frequencies for speech recognition,” *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1079-1082, 1997.

[10] D. Pearce and H. G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ICSA ITRW ASR2000*, 1999.

[11] HTK. [Online]. Available: http://htk.eng.cam.ac.uk/