On Wavelet Transform Based Feature Extraction for Speech Recognition Application

*Jae-Gil Kim, *Sung-Joo Ahn, and *Han-Seok Ko

※This research is performed by the University Research Grant.

Abstract

This paper proposes a feature extraction method using wavelet transform for speech recognition. Speech recognition system generally carries out the recognition task based on speech features which are usually obtained via time-frequency representations such as Short-Time Fourier Transform (STFT) and Linear Predictive Coding (LPC). In some respects these methods may not be suitable for representing highly complex speech characteristics. They map the speech features with same time and frequency resolutions at all frequencies. Wavelet transform overcomes some of these limitations. Wavelet transform captures signal with fine time resolutions at high frequencies and fine frequency resolutions at low frequencies, which may present a significant advantage when analyzing highly localized speech events. Based on this motivation, this paper investigates the effectiveness of wavelet transform for feature extraction focused on enhancing speech recognition.

The proposed method is implemented using Sampled Continuous Wavelet Transform (SCWT) and its performance is tested on a speaker-independent isolated word recognizer that discerns 50 Korean words. In particular, the effect of mother wavelet employed and number of voices per octave on the performance of proposed method is investigated. Also the influence on the size of mother wavelet on the performance of proposed method is discussed. Throughout the experiments, the performance of proposed method is compared with the most prevalent conventional method, MFCC (Mel-Frequency Cepstral Coefficient). The experiments show that the recognition performance of the proposed method is better than that of MFCC. But the improvement is marginal while, due to the dimensionality increase, the computational loads of proposed method is substantially greater than that of MFCC.

I. Introduction

The error-free speech recognition system has remained unattainable over the years. Commercial systems have been developed to handle small to medium vocabulary with moderate performance. Large vocabulary systems able to handle 6,000 to 10,000 words have been developed and demonstrated under laboratory conditions. However, all these systems suffer severe degradations in recognition accuracy when there are any deviations between the conditions where the systems are trained and the conditions where the systems are tested. Factors causing these differences are as follows: (1) speaking style, (2) linguistic content of the task, and (3) environment [1].

Among these factors, robustness to the environment is an essential requirement for the widespread use of speech recognizer. Environmental factors are additive background noise, convolutional channel distortion etc. To make a recognizer more robust to environment, the following four types of approaches are known to exist:

(1) Speech enhancement [2],

(2) Robust feature extraction [3],

(3) Robust distance measure [4], and

(4) Model-based compensation [5].

This paper focuses on the robust feature extraction employing wavelet transform.

All speech recognizers include an initial signal processing front-end required to extract important features from the speech waveform that are relatively insensitive to talker and channel variability unrelated to speech message content. This first stage also reduces the data rate for later stages of the recognizer and attempts to decrease redundancy inherent in speech waveform. Vast majority of front ends is based on the standard signal processing techniques such as filter banks, linear predictive coding (LPC), homomorphic analysis (cepstral)[6]. There has also been interest in front-ends based on known properties of the human auditory system. Some of these front-ends are linear but with parameters that correspond to auditory properties. Front-ends based on the auditory system have been shown to outperform more conventional signal processing schemes for speech recognition tasks [7]. Recent work has also explored the use of data reduction techniques such as linear discriminant analysis (LDA) to generate reduced

* School of Electrical Engineering Korea University

Manuscript Received : April 23, 1998.

feature set [8]. Such techniques have shown to be successful in speech recognition tasks, especially when noise or spectral tilt degrades speech. Among these front ends, most popular methods are LPCC (Linear Predictive Cepstral Coefficient)[9], Perceptual Linear Prediction (PLP)[10], and Mel-Frequency Cepstral Coefficient (MFCC) etc. [11].

Speech recognition systems generally carry out some kind of recognition tasks based on speech features which are usually obtained via time-frequency representations such as Short Time Fourier Transform (STFT) or Linear Predictive Coding (LPC) techniques. In some respects, these methods may not be suitable for representing speech. They assume the signal stationarity within a given time frame and may therefore lack the ability to analyze localized events accurately. In other words, they analyze all frequency components of a signal with same time and frequencies resolutions. Wavelet transform overcomes some of these limitations; it can provide fine time resolutions at high frequencies and better frequency resolutions at lower frequencies. This takes the best features of narrow band and wide band analysis and places it in one transform. Better time resolutions at high frequencies will allow more precise location of the beginning of an utterance and short time features will not be smeared by averaging that is inherent in the conventional method[12]-[19]. Based on these motivations the paper proposes a feature extraction method using wavelet transform and investigates its effectiveness for speech recognition by varying wavelet parameters.

This paper is organized in four sections. After a brief introduction in Section 1, Section 2 describes the proposed feature extraction method using wavelet transform. Section 3 presents the experimental results of speech recognition based on the proposed method. A concluding remark commenting on the findings is provided in Section 4.

II. Proposed Feature Extraction Method: Wavelet Transform based Cepstral Coefficient

The proposed method, Wavelet Transform based Cepstral Coefficient (WTCC), extracts information in cepstrum domain and its calculation procedure is very similar to MFCC. WTCC uses wavelet transform in order to obtain power spectrum instead of STFT in MFCC. The power spectrum obtained by wavelet transform has fine time resolutions at high frequencies and fine frequencies at low frequencies while the power spectrum via STFT has same time and frequency resolutions at all frequencies.

2.1 Employed wavelet transform: Wavelet coefficients are obtained by computing the correlation between each wavelet and the signal. The DWT computes wavelet coefficients on a dyadic grid. This makes it difficult to use DWT as a part of feature extraction for HMM-based speech recognizer because speech recognizer of this type requires frame synchronous input data. In other words, input data rate must be same at all scales. The Sampled Continuous Wavelet Transform (SCWT), a variation of the DWT meets this condition. Therefore, SCWT is employed for our feature extraction. The SCWT is given by

$$SCWT(a^j,\ n) = a^{-\frac{j}{2}} \sum_k \varphi\left(\frac{k-n}{a^j}\right) x(k) \qquad (1)$$

By restricting a shift factor, n, to a fixed value, this transform will generate frame-synchronous coefficients in a redundant fashion but still retains the features that are offered by the wavelet transform. It is common to discretize the scale parameter by choosing $a = a_0 2^{m/v}$ where v is the number of voices per octave [20].

2.2 Mother wavelet: Windowed modulated wavelets have been widely used in speech analysis. Examples include Morlet wavelet (Gaussian window), Hanning window, and Hamming window. Morlet wavelet has no scaling function but is explicit. Hamming window is widely using in speech analysis with its peak side lobe of 41dB below the main lobe. Hanning widow has been used in wavelet analysis with its peak lobe being 31 dB below the main lobe. It is important to choose a wavelet function that suits the applications. Modulation of window functions allows the control of the center frequency and bandwidth of the wavelet. We explore how differently modulated window functions may affect the recognition performance due to the different frequency and time responses.

2.3 Construction of SCWT: In our implementation, the frequency range to be covered is from 300 to 3400 Hz because the speech is sampled at 8 kHz and is band-limited to the frequency band by telephone line. Mother wavelet is modulated to 3400 Hz and dilated in order to analyze lower frequencies. This requires 3 octaves of wavelets and several voices per octave are used to resolve the frequency components of speech between octaves.

2.4 Feature extraction procedure of WTCC: Feature extraction procedure using WTCC is very similar to MFCC. Instead of STFT, Mel-scale filter banks in MFCC, the WTCC uses SCWT to extract energy of each frequency band. The feature extraction procedure (Figure 1) using WTCC is as follows:

(1) The input speech signal is sampled at a frequency of 8 kHz.

(2) A pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ is applied to the sampled speech.
The pre-emphasis filter is used to reduce the effects of glottal pulses and radiation impedance and to focus on the spectral properties of the vocal tract.

(3) For each time shift, wavelet-based spectral coefficients are derived using SCWT (Sampled Continuous Wavelet Transform).

(4) The magnitudes and logarithms of these coefficients are taken to yield log-energy outputs, which represent the log-energy of each frequency band.

(5) Discrete Cosine Transform (DCT) (Equation 2) is applied to these log-energy outputs, yielding cepstral coefficients.

$$x_t[k] = \sum_{i=0}^{N-1} X_t[i] \cos\left(k(i+\frac{1}{2})\frac{\pi}{N}\right) \qquad (2)$$

where N = the number of wavelets
$X_t[i]$ represents the log power output of the i^{th} wavelet at time shift, t, and
$x_t[k]$ represents the k^{th} cepstral coefficient at time shift, t.

In our implementation, the number of cepstral coefficients is 13.

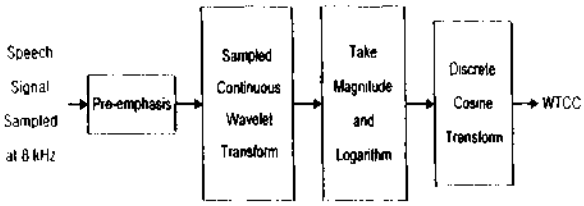

Figure 1. WTCC Block Diagram.

III. Experimental Results

This section describes the speech recognition system used for experiments and presents the results of the recognition performance using WTCC.

An isolated word recognition system that discerns 50 Korean words is implemented for experiments. The baseline system (Figure 2) is based on Discrete Hidden Markov Model (DHMM). Each word is modeled by a 30-state left-right HMM. The feature vector is composed of 4 sets:

(1) 12 cepstral coefficients,

(2) 12 1st-order derivatives of cepstral coefficients,

(3) 12 2nd-order derivatives of cepstral coefficients,

(4) 1st and 2nd-order derivatives of power.

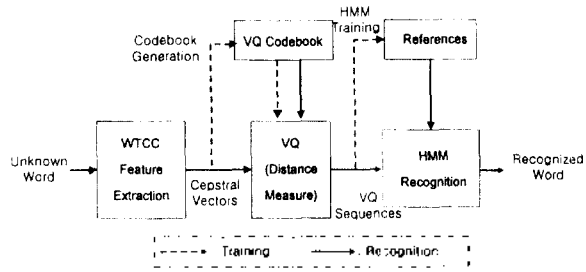

Figure 2. Baseline Speech Recognition System.

The vector quantization (VQ) codebook size of each feature set is:

(1) cepstral coefficients, 1st derivatives and 2nd derivatives: 256,

(2) 1st and 2nd-order derivatives of power: 64.

Our training and recognition tests are performed using the database for VDS (Voice Dialing System). Being collected via telephone line, the database are sampled at 8 *kHz* and band-limited to the frequency range from 300 to 3400 *Hz*. They are contaminated by additive non-stationary background noise and convolutional channel distortion, which is caused by communication channel and microphone. The database consists of 68 speakers from 10 to 50 year-old females and males. Among the database 1169 words from 63 speakers are used for training and 111 words from 5 speakers are used for recognition tests

3.1 The effect of the number of voices per octave in SCWT on the performance of WTCC: Collected over telephone line, the database used for the experiments are sampled at 8 *kHz* and band-limited to the frequency band from 300 to 3400 *Hz*. It requires 3 octaves of wavelets to cover this frequency range and several voices per octave to resolve the frequency components of speech. The number of voices may affect the performance of WTCC due to different frequency resolutions. So the effect of the performance of the WTCC with respect to the number of voices is investigated. The number of voices used for the experiments is 6, 7 or 8 and the corresponding number of wavelets of SCWT is 18, 21 or 24. Figure 3 visualizes the effect of the number of voices on the scalogram via SCWT. Table 1 shows the recognition rates of WTCC with respect to the number of voices. The mother wavelet employed for these experiments is 8ms Morlet wavelet. The recognition results

show that the increased number of voices per octave improves the recognition performance due to finer frequency resolutions.

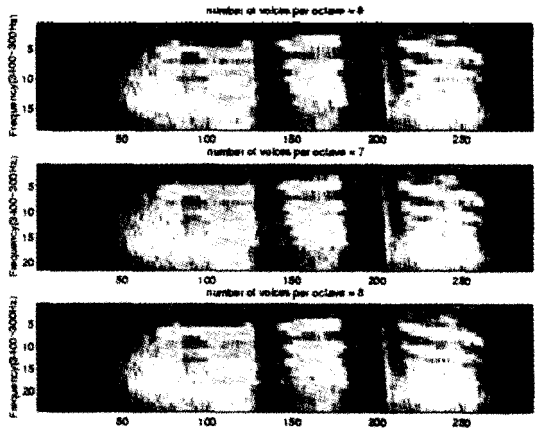

Figure 3. Scalogram of a Korean word "Harabujy" with respect to the number of voices; (a) 6, (b) 7, and (c) 8.

Table 1. Recognition rate of WTCC with respect to the number of voices per octave.

Number of voices	6	7	8
Recognition Rate (%)			
Top1	81.98	84.68	85.58
(Top2)	(90.09)	(96.39)	(96.39)

3.2 The effect of mother wavelet on the performance of WTCC: In implementing SCWT, it is important to choose an appropriate mother wavelet, which suits the application. Different mother wavelets may affect recognition performance due to different frequency and time responses. So, the effect of 3 different mother wavelets on the performance of WTCC is investigated. Mother wavelets used for experiments are Morlet window (Gaussian window), Hamming window and Hanning window. Figure 4 shows the differences among the scalograms resulting from different mother wavelets used. The number of voices per octave used for these experiments is 8 and the size of mother wavelet is 8ms. Table 2 shows the recognition tests of WTCC with respect to 3 different mother wavelets. The results indicate that Morlet wavelet gives WTCC the best performance.

3.3 The effect of the size of mother wavelet on the performance of WTCC: The size of mother wavelet determines the time and frequency resolution of WTCC. The smaller size of mother wavelet gives WTCC finer

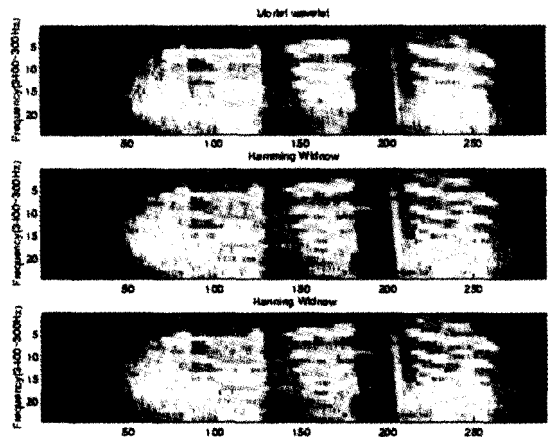

Figure 4. Scalogram of a Korean word "Harabujy" when mother wavelets are; (a) Morlet wavelets, (b) Hamming window, and (c) Hanning window.

Table 2. Recognition rates of WTCC with respect to different mother wavelets.

Mother wavelet	Morlet wavelet	Hamming window	Hanning window
Recognition Rate (%)			
Top1	85.59	83.78	83.78
(Top2)	(96.39)	(92.79)	(93.69)

time resolutions but increases computational loads. Finer time resolutions may improve the performance of WTCC. So it is necessary to investigate the effect of the size of mother wavelet on the performance of WTCC. In this experiment, we change the size of mother wavelet from 8ms to 6ms and check the performance of WTCC. The corresponding sizes of time shifts are half of the sizes of mother wavelets. Mother wavelet used for this experiment is Morlet wavelet and the number of voices is 8. Figure 5 visualizes differences between scalograms of a Korean Word "Harabujy" when the size of mother wavelet is 8, 6 or 4ms. The Figure 5 shows that the size of mother wavelet plays a role on improving the time resolutions. Table 3 presents the recognition rates of WTCC with respect to the size of mother wavelet. The results show that the smaller size of mother wavelet improves the performance of WTCC.

3.4 Comparison of WTCC with MFCC: Experiments conducted a comparative test for recognition performance of WTCC with the most popular conventional method, MFCC. Table 4 shows the results of these experiments. MFCC parameters used for these experiments are as follows:

(1) Number of filter banks: 24,

(2) Window size: 25 ms, and

(3) Frame rate: 10 ms.

The WTCC employed uses Morlet wavelet and 8 voices per octave.

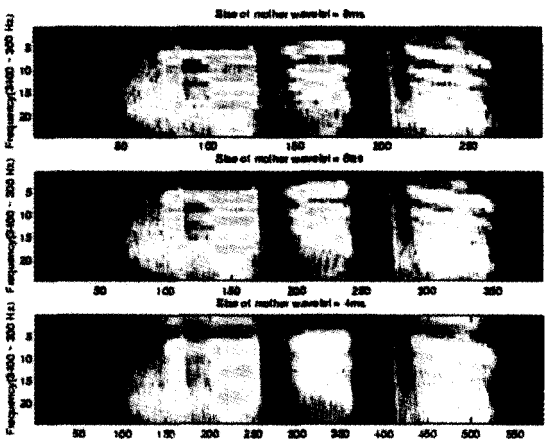

Figure 5. Scalograms of a Korean word "Harabujy" with different sizes of mother wavelet; (a) 8 ms, (b) 6 ms, and (c) 4 ms.

Table 3. Recognition rates of WTCC with respect to the size of mother wavelet.

Size of mother wavelet (ms)	8	6
Recognition Rate (%)		
Top1	85.59	88.29
(Top2)	(96.39)	(96.39)

Table 4. Recognition rates of WTCC and MFCC.

Feature extraction	WTCC	MFCC
Recognition rate (%)		
Top1	88.29	87.39
(Top2)	(96.39)	(94.59)

The result shows that top1 and top2 recognition rates of WTCC are higher than that of MFCC, but the improvement is marginal. The reasons for WTCC not being able to give significant improvement over MFCC despite of finer time resolution are interpreted by the following reasoning:

(1) The main difference between WTCC and MFCC is different time resolution. WTCC has fine time resolutions at high frequencies, while MFCC has same time resolutions at all frequencies. Figure 5 represents the differences between the mel-scale spectrogram via MFCC and the scalogram via WTCC. In the implementation of WTCC, the size of mother wavelet is changed from 8ms to 6ms. It is considered that this size of mother wavelet is not short enough to give enough fine time resolutions at high frequencies and may not result in significant improvement of the performance of WTCC over MFCC. Since choosing an even shorter size will make WTCC impractical due to the increased processing time, the shorter sizes are not considered in these experiments.

(2) Speech recognition can be thought of as being performed by locating the formants of speech. There are about the first three formants of significance. These three formants are located at low frequencies. So, the analysis of WTCC with fine time resolutions at high frequencies does not give significant advantages over MFCC.

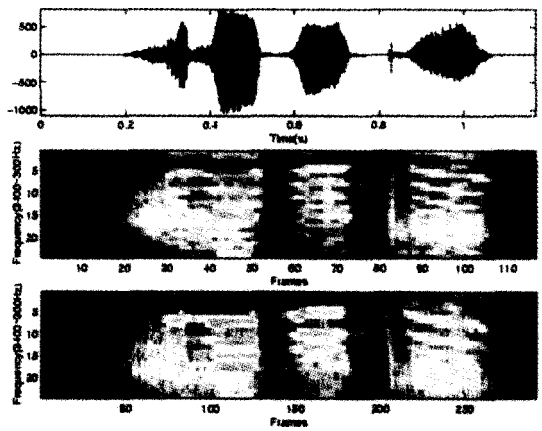

Figure 6. (a) Waveform, (b) Mel-scale Spectrogram and (c) Scalogram of a Korean word "Harabujy".

3.5 Comparison of WTCC with MFCC compensated with CMS(Cepstral Mean Subtraction): The use of the convolutional channel distortion compensation method, CMS, results in a higher recognition rate in MFCC. To verify whether this is the case also in WTCC, an experiment of applying CMS to WTCC is performed. Table 5 shows the results of the experiments. The results suggest that applying CMS to WTCC indeed improve the recognition performance. In this case, the top1 recognition rate of WTCC is about equal to that of MFCC while the top2 recognition rate indicate an improvement over MFCC.

Table 5. Recognition rates(%) of WTCC and MFCC compensated by CMS.

Feature extraction	WTCC+CMS	MFCC+CMS
Recognition rate (%)		
Top1	89.19	89.19
(Top2)	(96.39)	(95.50)

3.6 Comparison of WTCC and MFCC on compu-

tational load: A comparison of the computational load of WTCC and MFCC can be made directly by counting the number of data points involved in the computation of coefficients in each frame. The frame rate involved in each method represents the amount data consumed to generate 13 coefficients/frame. The frame rate essentially indicates the period of producing a frame worth of data. Both WTCC and MFCC produced 13 coefficients/frame. The frame rates of WTCC and MFCC are shown in Table 6. From Table 6, it is readily seen that WTCC (8ms) produces 10/4 times larger data amount than that of MFCC while WTCC(6ms) produces 10/3 times larger than that of MFCC. Accordingly, the result reflects that WTCC requires stiffer computational load than MFCC.

Table 6. Frame rate of WTCC and MFCC.

Feature extraction	MFCC	WTCC (size of mother wavelet = 8ms)	WTCC (size of mother wavelet = 8ms)
Frame rate	10 ms	4 ms	3 ms

IV. Conclusions

This paper proposed a wavelet transform based feature extraction method for speech recognition and presented a comparative analysis on the results of recognition rates using the proposed method. The proposed method, WTCC, extracts spectral information using wavelet transform instead of the conventional methods such as STFT and LPC. The investigation is motivated to exploit the strengths of Wavelet transform having several advantages over the conventional methods. Wavelet transform captures signal with fine time resolutions at high frequencies and better frequency resolutions at low frequencies, while the conventional methods analyze the signal with same time and frequency resolutions at all frequencies. Based on this anticipation, this paper focused on the effectiveness of wavelet transform for feature extraction for speech recognition.

The proposed method is implemented using Sampled Continuous Wavelet Transform (SCWT) and generates coefficients in cepstrum domain. The performance of the proposed method was measured with a speaker-independent isolated word recognizer that discerns 50 Korean words using speech database for voice dialing service. In particular, the effect of mother wavelet and number of voices per octave in SCWT on the performance of the proposed method was investigated. Also the influence of the size of mother wavelet on the performance of the proposed method was explored. The performance of the proposed method was compared with the most prevalent method, MFCC. Through the experiments, it has been determined that the Morlet wavelet as a mother wavelet gave the best performance among candidate wavelets. Also, greater the number of voices per octave improves the performance of the proposed method due to its ability for finer frequency range analysis. In addition, shortening the size of mother wavelet gives the proposed method finer time resolutions and improves its performance. A comparison of the proposed method with MFCC shows that the performance of the proposed method is better than that of MFCC. But the improvement is marginal while, due to the dimensionality increase, the computational load of the proposed method is substantially greater.

References

1. A. Acero, "*Acoustical and Environmental Robustness in Automatic Speech Recognition*," Ph.D. Thesis, Dept. of Electrical and Computer Engineering, Carnegie-Mellon University, 1990.
2. J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth compression of Noisy speech," *Proceeding of IEEE*, Vol. 67, No. 12, pp. 1586-1604, December 1979.
3. J. C. Junqua and J.P. Hanton, "*Robustness in Automatic Speech Recognition*", Kluwer Academic Publishers, pp. 233-272, 1996.
4. D. Mansour and B.H. Juang, "A Family of Distortion Measures Based upon Projection Operation for Robust Speech Recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, No. 11, pp. 1659-1671, November 1989.
5. A. Varga and R. Moore, "Hidden Markov Model decomposition of speech and noise," *ICASSP'90*, pp. 845-848, 1990.
6. J. W. Picone, "Signal modeling Techniques in Speech Recognition," *Proceeding of IEEE*, vol. 81, pp. 1215-1247, Sep 1993.
7. C. R.. Jankowski, V. Hoang-Doan and R. P. Lippman, "A comparison of Signal Processing Front Ends for Automatic Word Recognition," *IEEE Trans. on Speech and Audio Processing*, vol.3, pp. 286-293, Jul. 1995.
8. O. Siohan, "On the Robustness of Linear Discriminant Analysis as a Preprocessing Step for Noisy Speech Recognition," *ICASSP'95*, pp. 125-128, 1995.
9. S. Young, J. Odell, D. Ollason,V. Valtchev, and P. Woodland, "*The HTK BOOK*," Cambridge University, 1997.
10. H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "RASTA-PLP Speech Analysis Technique," *ICASSP'92*, vol.1, pp. 121-124, 1992.
11. S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol 28, pp. 357-366 , 1980.
12. R. Olivier and V. Martin, "Wavelets and Signal Processing," *IEEE Signal processing Magazine*, pp.14-38, Oct. 1991.

13. M. J Shensa, "The Discrete Wavelet Transform : Wedding the a Trous and Mallat Algorithms," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 40, pp. 2464-2482, Oct. 1992.
14. M. Vetterli and C. Herley, "Wavelets and Filter banks: Theory and Design," *IEEE Trans. on Signal Processing*, vol. 40, pp. 2207-2232, Sept. 1992.
15. S. Mallat, "A Theory for Multiresolution Signal Decomposition: the Wavelet Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp.674-693, July 1989.
16. I. Daubechies, "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE Trans. on Information Theory*, vol. 36, pp. 961-1005, Sept. 1990.
17. O. Rioul and P. Duhamel, "Fast Algorithms for Discrete and Continuous Wavelet Transforms," *IEEE trans. on Information Theory*, vol. 38, pp. 569-586, Mar 1992
18. L. Janer, "Modulated Gaussian Wavelet Transform based Speech Analyzer(MGWTSA) Pitch Detection Algorithm," *EUROSPEECH'95*, vol. 1, pp. 401-404, 1995.
19. S. Kadambe and G. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE trans. on Information Theory*, vol. 38, pp. 917-924, 1993.
20. M. Vetterli and J. Kovacevic, "*Wavelets and Subband Coding*," Prentice Hall, 1995.

▲Jaegil Kim DOB : 12 March 1972

Jaegil Kim received the B.S. degree from Korea University in 1996, M.S. degree from the Korea University in 1998, all in electrical engineering. From 1998 to present, he is Research Engineer (H/W), WLL Development Division, LG Infomation & Communications, Ltd

E-mail : jgkim@ispl.korea.ac.kr

▲Sungjoo Ahn DOB : 9 July 1975

Sungjoo Ahn received the B.S. degree from Korea University in 1998, in electrical engineering. From 1998 to present, he is M.S course from the Korea University in electrical engineering.

E-mail : sjahn@ispl.korea.ac.kr

▲Hanseok Ko DOB : 10 August 1960

Hanseok Ko received the B.S. degree from Carnegie-Mellon University in 1982, M.S. degree from the Johns Hopkins University in 1988, and Ph.D. degree from the Catholic University of America in 1992, all in electrical engineering. He also received the M.S. degree in Systems Engineering in 1986 from the University of Maryland, College Park. From 1982 to 1987, he was with the Search and Track Division of the Naval Surface Warfare Center in White Oak, Maryland, where his work involved radar signal and infrared image processing for the detection of low observable targets. From 1988 to 1994, he directed research and development projects for the 21st Century ship automation system with the Sensors and Electronics Division, NSWC. In addition, Dr.Ko was a part-time faculty member in the Dept of Electrical Engineering at UMBC from 1992 to 1995. In March of 1995, Dr.Ko joined the faculty of the Department of Electronics Engineeringat Korea University. His professional interests include signal processing, data fusion, navigation and tracking, and visualization.

Research Interests

(1) Speech Signal Processing - preprocessing, recognition, coding, synthesis
(2) Image data fusion - GIS fusion, adaptive wavelet classification, medical imaging
(3) Detection, Estimation, and Tracking - bias correction, sensor fusion, observability enhancement, navigation

E-mail : hsko@ispl.korea.ac.kr