THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.4E 2003. 12. pp. 161~166

# Adaptive Wavelet Based Speech Enhancement with Robust VAD in Non-stationary Noise Environment

Sungwook Chang*, Sungil Jung*, Younghun Kwon**, Sung-il Yang*
*School of Electrical and Computer Engineering, Hanyang University
**Department of Physics, Hanyang University
(Received May 6 2003; accepted August 28 2003)

**Abstract**

We present an adaptive wavelet packet based speech enhancement method with robust voice activity detection (VAD) in non-stationary noise environment. The proposed method can be divided into two main procedures. The first procedure is a VAD with adaptive wavelet packet transform. And the other is a speech enhancement procedure based on the proposed VAD method. The proposed VAD method shows remarkable performance even in low SNRs and non-stationary noise environment. And subjective evaluation shows that the performance of the proposed speech enhancement method with wavelet bases is better than that with Fourier basis.

***Keywords:*** *Speech enhancement, Wavelet, VAD, Speech dominant indicator, Spectral subtraction*

## I. Introduction

In many speech related applications, speech has to be processed in the background of undesirable noise. During the last two decades, various approaches to reduce the noise have been proposed. Among them, spectral subtraction is one of the widely used methods. This method requires the estimation of statistical information for a background noise. Hence, the accuracy of the estimated statistical noise information decides the performance of the enhancement system. However, conventional spectral subtraction method based on Fourier basis where assumes that the input signal is periodically sinusoidal cannot show good estimation of spectrum in noisy environment. Especially, it is almost impossible to perform an accurate estimation of noise information by Fourier basis in low SNRs or non-stationary noise environment. For that reason, various wavelet based denoising methods had been proposed[1,2]. Unfortunately, the conventional wavelet based denoising methods sometimes induce an artifact in speech enhancement system. Thus, it is one of the crucial issues of the wavelet based speech enhancement systems to alleviate the artifact in low SNRs or non-stationary noise environment[3,4].

To resolve the non-stationary noise problem (including artifact problem), we propose an adaptive wavelet packet based spectral subtraction method. The proposed method includes a new VAD method with adaptive wavelet packet transform and a modified spectral subtraction method based on a speech dominant indicator (SDI).

Corresponding author: Sungwook Chang (schang@ihanyang.ac.kr)
School of Electrical and Computer Engineering, Hanyang University, 17, Haengdang-dong, Seongdong-gu, Seoul, 133-791, Korea

## II. Voice Activity Detection

In many speech enhancement methods segments of pure noise are evaluated by detection of speech pauses. However, this is a difficult task in practical environments, especially if the background noise is not stationary or the SNR is low[5]. To solve the problem, we propose a new VAD

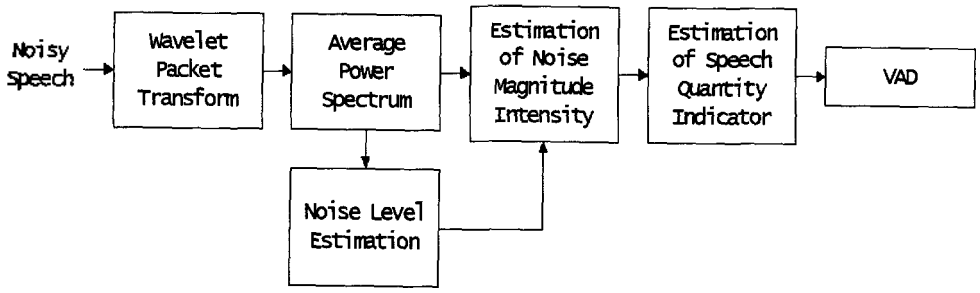



Figure 1. Proposed VAD process.

method. For the purpose, noise magnitude intensity (NMI) and speech dominant indicator (SDI) are suggested as new measures for the robust VAD. The proposed VAD is used to estimate noise information in noise only area. Fig. 1 depicts the proposed VAD.

## 2.1. Noise Magnitude Intensity

In this section, we present an estimation method of noise magnitude intensity (NMI).

At first, noise level is estimated as follows[5]

$$\hat{N}_j(n) = \begin{cases} \alpha \cdot \hat{N}_{j-1}(n) + (1-\alpha) \cdot P_j(n), & P_j(n) \le \beta \cdot \hat{N}_{j-1}(n) \\ \hat{N}_{j-1}(n) & , \ otherwise \end{cases} \quad (1)$$

*which* $\alpha = 0.98$, $1.5 \le \beta < 2.5$

$n$ : *wavelet packet filterbank index*

$j$ : *frame index*

$P_j(n)$: *average power spectrum of* $n^{th}$ *wavelet packet filterbank at* $j^{th}$ *frame*

$\hat{N}_j(n)$ : *estimated noise level*

Next, we estimate a NMI with the estimated noise level as a measure of noise intensity in spectrum. The below **Algorithm 1** shows the estimation method of NMI.

**Algorithm 1:** Estimation of NMI

*Step 1*: Initialize *NMI* and wavelet packet filterbank index (*i.e.* node number in wavelet packet tree), *NMI* = 0 and $n = 0$.

*Step 2*: If $n$ is larger than the highest wavelet packet filterbank index, then exit.

*Step 3*: If $\left(1 \le \sqrt{\frac{|\hat{N}(n)|}{|P(n)|}}\right)$, then *NMI* = *NMI* + 1.

$n = n + 1$

*which* $P(n)$: *average power spectrum of* $n^{th}$ *wavelet packet filterbank*

$\hat{N}(n)$: *estimated noise level*

go to *Step 2*.

The NMI provides an information about amounts of noise components concentrated in the current noisy speech. This is very useful and important information for robust VAD in noisy environment.

## 2.2. Speech Dominant Indicator

As we have mentioned, it is very difficult to perform accurate VAD in low SNRs or non-stationary noise environment. Thus, we propose a speech dominant indicator (SDI) $\kappa$ for robust VAD as shown in Eq. 2 and Eq. 3. The SDI is defined by multiplication of geometric mean and weight $\lambda^{\gamma}$ based on NMI.

$$\kappa := 10 \log_{10}\left(\left|\left(\prod_{i=1}^{N_{FB}} P(i)\right)^{\frac{1}{N_{FB}}}\right| \cdot \lambda^{\gamma}\right) \quad (2)$$

$$\lambda = \log_{10}\left(\frac{N_{FB}}{NMI}\right) \quad (3)$$

*which* $N_{FB}$ : *the number of filterbank*

*NMI* : *noise magnitude intensity extracted from* **Algorithm 1.**

$\lambda^{\gamma}$ : *weight*

$\gamma$ : *experimental exponent parameter A value between 1 and 3 is used, experi-mentally.*

Fig. 2 shows SDI for $\gamma$ = 0, 1, 2. In the figure, all regions (voice, unvoice, noise only) are hand-segmented. We can see that SDI using NMI gives rise to better discrimination than SDI only ( $\gamma$ = 0) in Fig. 2. Especially, SDI using NMI induces good discrimination between noise and unvoiced speech even in low SNRs and non-stationary noise environment.

## 2.3. VAD with SDI

For robust detection of a boundary between speech and noise, we propose some thresholds based on the SDI. At first, we assume that there is preliminary silence area at least for 200 msec ahead of utterance of speaker. Now, we introduce a procedure based on SDI for each frame to decide a threshold for detection of a boundary between speech and noise. Initially, a mean value of SDI $\kappa_{mean}$ is defined by mean value for an interval of preliminary silence area. And the last frame of the interval of preliminary silence area is marked by $F_{DF}$. Next, we find a frame which is continuously larger than $\kappa_{mean}$ during 80 msec from $F_{DF}$ to starting frame $F_{SF}$. The found frame is marked by $F_{DF\text{-}80ms}$ as shown in Fig. 3. And we find mean ( $\kappa'_{mean}$), maximum ( $\kappa'_{max}$) and minimum value ( $\kappa'_{min}$) of SDI for an interval between starting frame FSF and $F_{DF\text{-}80ms}$. At last, we define temporal threshold $TH_{TT}$ and final threshold $TH_{FT}$ as follows:

$$TH_{TT} = \kappa'_{mean} + (\kappa'_{max} - \kappa'_{min}) \tag{4}$$

$$TH_{FT} = \frac{2 \cdot \kappa'_{mean} + \kappa'_{max} + \kappa'_{min}}{4} \tag{5}$$

Most of local maximum introduced by background noise can be reduced by temporal threshold $TH_{TT}$.

Now, we suggest a robust VAD procedure using SDI and the proposed thresholds. The procedure is shown in Fig. 4. At first, we find tentative starting and ending points (set by TSP and TEP, respectively) of contact between temporal threshold $TH_{TT}$ and curve of SDI $\kappa$. The next step is to move backwards from TSP (forward from TEP) comparing SDI $\kappa$ to final threshold $TH_{FT}$. If SDI $\kappa$ falls below final threshold $TH_{FT}$, temporal starting point TSP is moved back (forward at TEP) to the corresponding point (set by SP or EP).

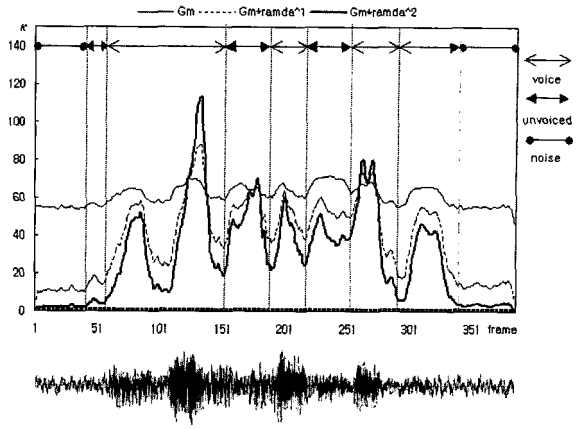



Figure 2. Speech dominant indicator $\kappa$ for $\gamma$ = 0, 1, 2: "three one zero six" (additive factory noise at a SNR = 5 dB) for decision of thresholds.

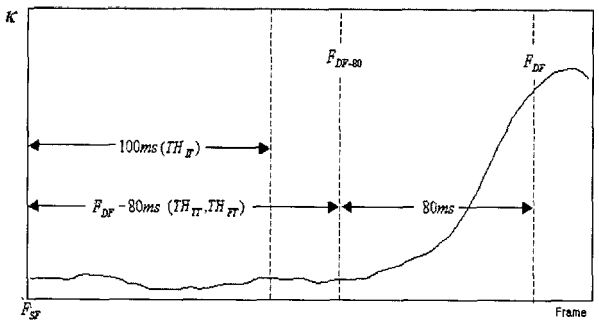



Figure 3. Decision of thresholds.

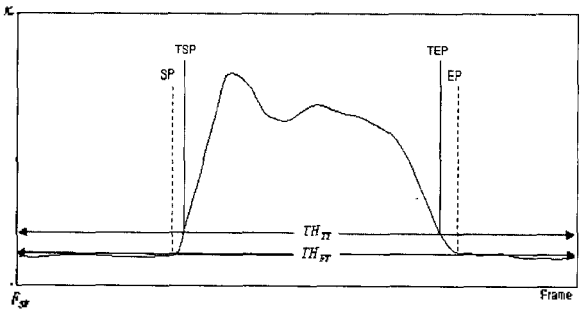



Figure 4. Detection of voice activity.

Fig. 5 shows examples of the proposed VAD algorithm. The first row of each figure is waveform of clean speech and second row shows noisy speech. Then the last row shows SDI for the corresponding noisy speech. Figs. 5(a) and (b) show examples of faulty VAD. When we fail in correct noise estimation, SDI derived from estimated noise introduces a faulty VAD. And unsuitable selection of VAD thresholds in fast varying noisy speech induces some error. Finally, low SNR environment is the factor leading

to faulty VAD. However, in most cases, we can take good VAD performance even in colored or non-stationary noise environments as shown in Figs. 5(c) and (d). In other words, the proposed VAD algorithm shows good detection rate with average $\pm 40 \sim \pm 50$ msec error range in SNR = 15 dB ~ 10 dB and average $\pm 60 \sim \pm 70$ msec error range in SNR=5 dB ~ 0 dB.

# III. Noise Estimation and Modified Spectral Subtraction

## 3.1. Noise Estimation

Now, we reestimate noise spectrum by modified noise estimation method taken from the proposed VAD as follows:

$$\hat{N}_j(i) = \begin{cases} \alpha \cdot \hat{N}_{j-1}(i) + (1-\alpha) \cdot P_j(i), & \begin{cases} \textit{noise area} \\ \text{or} \\ P_j(i) < \beta \cdot \hat{N}_{j-1}(i) \end{cases} \\ \hat{N}_{j-1}(i) & , \textit{otherwise} \end{cases} \quad (6)$$

*which* $\beta = 2$, $\alpha = 0.98$

$1 \le i \le$ *filterbank size*

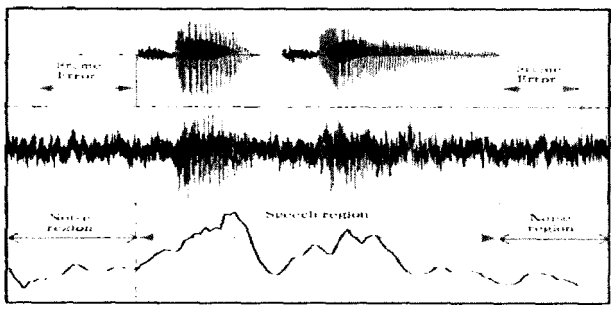
$j =$ *frame index*

This procedure is similar with Eq. 1 except that we estimate most of noise spectrum in noise only area.

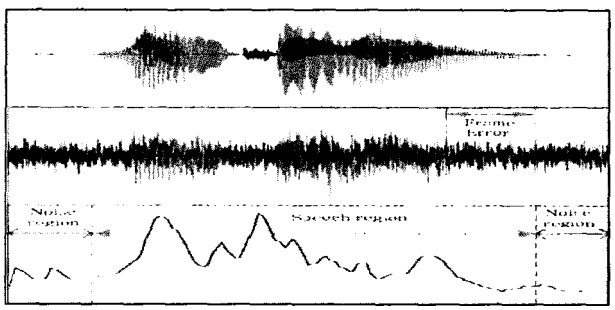
## 3.2. Spectral Subtraction Weight

Berouti *et al.* proposed a modified version of the power subtraction rule in which the amount of noise subtraction depends on the SNR of the particular frame[6]. A fixed subtraction weight is used over all frame in the Berouti *et al.*'s method. However, we need a new variable (not fixed) subtraction weight for each frame because a great deal of real environmental noise is non-stationary. That is, a real environmental noise has both time-varying statistical characteristic and time-varying SNR for each frame.

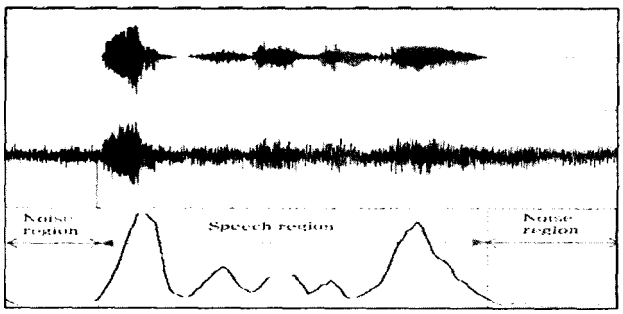
Thus, we propose a new estimation method of spectral subtraction weight at each frame for reliable speech enhancement method. The proposed method which depends on the SDI $\kappa$ is shown below.



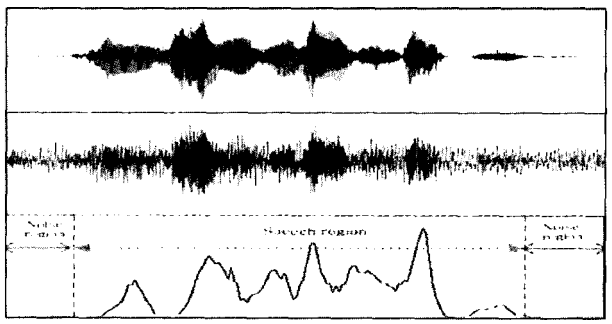

(a) An example of faulty VAD (SNR=0 dB with pink noise)




(b) An example of faulty VAD (SNR=0 dB with factory noise)




(c) An example of good VAD (SNR=0 dB with babble noise)




(d) An example of good VAD (SNR=0 dB with leopard noise)

Figure 5. Examples of the Proposed VAD with SDI.

$$\alpha_j = \left( \frac{\kappa_{\max} - \kappa_{\min}}{(\kappa_j - \kappa_{\max}) \cdot \lambda + \kappa_{\max} - \kappa_{\min}} \right)^2 \quad (7)$$

where $\alpha_j$, $j$ and $\lambda$ are spectral subtraction weight, frame index and normalization level, $0.4 \le \lambda \le 0.6$. The $\kappa_{\max}$, $\kappa_{\min}$, are maximum and minimum SDI over all frames. And $\kappa_j$ is the SDI at $j$th frame.

Finally, we define a filter for magnitude spectral subtraction with estimated noise level found from Eq. 6 as shown in Eq. 8 and perform the spectral subtraction by Eq. 9.

$$G_j(i) = \begin{cases} \sqrt{1 - \alpha_j \cdot \left( \frac{|\hat{N}_j(i)|}{|P_j(i)|} \right)^{0.5}} & , \text{if } \alpha_j \cdot |\hat{N}_j(i)| < |P_j(i)| \\ \beta & , \text{otherwise} \end{cases} \quad (8)$$

$$|\hat{S}_j(i)| = G_j(i) \cdot WPT_j(i) \quad (9)$$

*which* $0 \le \beta \le 0.02$, $1 \le i <$ *filterbank size*

$j$ : *frame index*

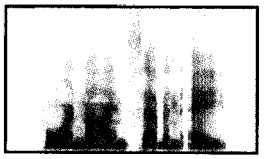
$|\hat{N}_j(i)|$: *estimated noise level*

$|P_j(i)|$: *average power spectrum of* $i^{th}$ *wavelet packet filterbank at* $j^{th}$ *frame*

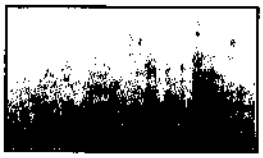
## IV. Evaluation

In noisy environment, the SNR cannot be used as faithful indication of speech quality. Thus, we employ subjective tests for evaluation of the proposed method. For subjective tests, we use an informal listening test and spectrum test. TIDIGIT 64 is used as speech database and various noise (babble, Leopard, pink, and Volvo noise), taken from Noisex-92 database, is added for our evaluation. Filterbank is composed of 64 uniform bands both in Fourier basis and wavelet bases.

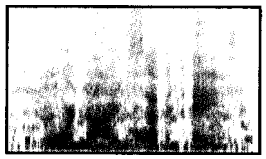
Fig. 6 shows speech spectra obtained by the proposed algorithm with Fourier basis and wavelet bases. The proposed algorithm with wavelet bases yields the better result than that with Fourier basis for the most noise conditions. Also, in informal listening test, the proposed algorithm with wavelet bases shows better performance than that with Fourier basis as shown in Table 1. And, as shown in Figs. 6(d) and (e), Coiflet and Daubechies' based results show similar spectrogram. However, as shown in Table 2, preference performance depends deeply on the characteristics of additive noise. That is, Daubechies' basis is suitable for speech babble noise and Leopard



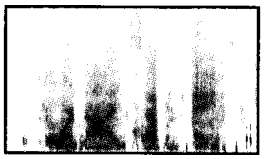
(a) Clean speech



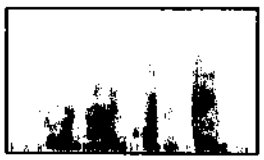
(b) Noisy speech (babble noise at SNR = 0 dB)



(c) Speech enhanced by proposed algorithm with Fourier basis



(d) Speech enhanced by proposed algorithm with Daubechies basis



(e) Speech enhanced by proposed algorithm with Coiflet basis

Figure 6. Speech spectrograms.

Table 1. Preference percentage between the outputs of proposed denoising algorithm using Fourier transform and wavelet packet transform (average of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB).

| | Fourier Transform | Wavelet Packet Transform | No Difference |
|---|---|---|---|
| Babble | 8% | 52% | 40% |
| Leopard | 28% | 52% | 20% |
| Pink | 28% | 72% | 0% |
| Volvo | 32% | 40% | 28% |

Table 2. Preference percentage between enhanced speech using Daubechies' basis and Coiflet basis (average of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB).

| | Daubechies' Basis | Coiflet Basis | No Difference |
|---|---|---|---|
| Babble | 50% | 14% | 36% |
| Leopard | 79% | 21% | 0% |
| Pink | 7% | 53% | 40% |
| Volvo | 7% | 53% | 40% |

Table 3. Preference percentage between the outputs of proposed denoising algorithm using wavelet packet transform based on Daubechies' basis and noisy speech signal (average of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB).

| | Proposed Algorithm | Noisy Signal | No Difference |
|---|---|---|---|
| Babble | 68% | 32% | 0% |
| Leopard | 68% | 32% | 0% |
| Pink | 72% | 28% | 0% |
| Volvo | 76% | 16% | 8% |

(military vehicle) noise, and Coiflet basis for the others. This result is matched with the characteristics of the wavelet bases. And Table 3 shows the preference percentage between the outputs of the proposed denoising algorithm using wavelet packet transform based on Daubechies' basis and noisy speech signal. Although the performance of informal listening test depends on the characteristic of background noise, most listeners preferred the output of the proposed algorithm to the non-processed noisy signal. But, there were cases that some listeners preferred the noisy signal due to the induced distortions and artifacts.

## V. Conclusions

We proposed the speech enhancement method with the VAD method based on adaptive wavelet packet for various noisy environments. The proposed speech enhancement algorithm shows good performance even though it is sensitive to choice of the wavelet basis. Furthermore, we can see that the proposed VAD with NMI based SDI is very useful to estimate the noise information even in low SNRs and non-stationary noise environment.

## References


1. D. L. Donoho, "Denoising by soft thresholding," *IEEE Trans. on Information Theory*, **41** (3), 613-627, 1995.
2. I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist. Soc. B*, 59, 319-351, 1997.
3. Sungwook Chang, Sung-il Jung, Younghun Kwon, and Sung-il Yang, "Speech enhancement using wavelet packet transform," *ICSLP 2002*, 1809-1812, 2002.
4. Sungwook Chang, Younghun Kwon, and Sung-il Yang, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," *ICASSP 2002*, 1, 561-564, 2002.
5. H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *ICASSP 95*, 153-156, 1995.
6. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *ICASSP-79*, 208-211, 1979.


## [Profile]


• Sungwook Chang

Sungwook Chang was born in Anyang, Kyunggi, Korea in 1971. He received the B.S. and M.S. degrees in Control and Instrumentation Engineering from Hanyang University, Korea in 1997 and 1999, respectively. Currently, he is graduate student for Ph.D. degree in Electronic, Electrical, Control and Instrumentation Engineering at Hanyang University. His current research interests include speech recognition, speech enhancement, wavelets, and bioinformatics. He is also a member of the Acoustical Society of Korea.

• Sungil Jung

Sungil Jung was born in Busan, Korea in 1972. He received the B.S. and M.S. degrees in Computer Engineering from Korea Maritime University, Korea in 2000 and 2002, respectively. Currently, he is graduate student for Ph.D. degree in Electronic, Electrical, Control and Instrumentation Engineering at Hanyang University. His current research interests include speech enhancement, speech recognition, and wavelets. He is also a member of the Acoustical Society of Korea.

• Younghun Kwon

Younghun Kwon was born in Seoul, Korea in 1961. He received his B.S. degree in Mathematics and Physics with the greatest honors from Hanyang University, Seoul, Korea, 1984, and his M.S. and Ph.D. degrees in Physics from the University of Rochester, Rochester, New York, 1986 and 1987, respectively. Since 1995, he has been with Hanyang University and he is now an Associate Professor at Department of Physics. His current research interests include Mathematical Physics, Theoretical Physics, Artificial Intelligence, Signal Processing and Quantum computing. He is also a fellow of the International Society for Complexity, Information, and Design, and member of American Mathematical Society, Korean Mathematical Society and Korean Physical Society.

• Sung-il Yang

Sung-il Yang was born in Geosan, Chungbuk, Korea in 1956. He received his B.S. degree in Electronics Engineering with the greatest honors from Hanyang University, Seoul, Korea, 1984, and his M.S. and Ph.D. degrees in Electrical & Computer Engineering from the University of Texas, Austin, Texas, 1986 and 1989, respectively. Since 1990, he has been with Hanyang University and he is now a Professor at the School of Electrical & Computer Engineering. His current research interests include speech recognition, digital signal processing, and responsible technology. He is also a member IEEE, Korea Institute of Telematics and Electronics, and the Acoustical Society of Korea.