THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.2E 2003. 6. pp. 62~68

Spectral Subtraction Using Spectral Harmonics for Robust Speech Recognition in Car Environments

Jounghoon Beh*, Hanseok Ko*
*Dept. of Electronics and Computer Engineering, Korea University
(Received February 14 2003; accepted April 16 2003)

Abstract

This paper addresses a novel noise-compensation scheme to solve the mismatch problem between training and testing condition for the automatic speech recognition (ASR) system, specifically in car environment. The conventional spectral subtraction schemes rely on the signal-to-noise ratio (SNR) such that attenuation is imposed on that part of the spectrum that appears to have low SNR, and accentuation is made on that part of high SNR. However, these schemes are based on the postulation that the power spectrum of noise is in general at the lower level in magnitude than that of speech. Therefore, while such postulation is adequate for high SNR environment, it is grossly inadequate for low SNR scenarios such as that of car environment. This paper proposes an efficient spectral subtraction scheme focused specifically to low SNR noisy environment by extracting harmonics distinctively in speech spectrum. Representative experiments confirm the superior performance of the proposed method over conventional methods. The experiments are conducted using car noise-corrupted utterances of Aurora2 corpus.

Keywords: *Robust speech recognition, Spectral subtraction*

I. Introduction

The mismatch between training and testing condition is a major problem in ASR systems. The techniques to solve this problem can be categorized into two principal approaches. First is the spectral subtractive-type of algorithm performing noise suppression using short-time spectral amplitude, such as spectral subtraction, nonlinear spectral subtraction and Weiner filter. The other is the feature compensation algorithm such as cepstral mean normalization or vector Tayler series. In general, it is well known that spectral subtractive-type algorithm is very simple and efficient especially in stationary noisy environments.

This paper is about a new spectral subtractive-type scheme based on the idea that even though speech is heavily corrupted by noise, the shape of spectral harmonics of speech is well preserved as when speech is not corrupted[1,2].

The weakness of conventional spectral subtractive-type algorithm is identified and presented in Section 2. The proposed remedial approach is described in Section 3. In Section 4, we show the proposed method's effectiveness over conventional methods with representative experiments using Aurora 2. Concluding remarks are provided in Section 5.

Corresponding author: Jounghoon Beh (jhbeh@ispl.korea.ac.kr)
Dept. of Electronics and Computer Engineering, Korea University, 5Ka-1, Anan-dong, Sungbuk-ko, Seoul, 136-701, Korea

II. Spectral Subtractive-type Algorithm

When speech signal $x(n)$ is corrupted by background

additive noise $b(n)$, the corrupted speech can be expressed as follows:

$$y(n) = x(n) + b(n). \tag{1}$$

In the frequency domain, the power spectrum of the noisy signal can be reperesented as:

$$|Y(k)|^2 = |X(k)|^2 + |B(k)|^2 + X(k) \cdot B(k)^* + X(k)^* \cdot B(k)^*, \tag{2}$$

where k is the index of frequency bin.

2.1. Power Spectral Subtraction

In the case of power spectral subtraction, the short-time power spectrum of enhanced speech signal can be obtained by subtracting its noise estimate off from the corrupted speech. In Eq. (1) the terms, $|B(k)|^2$, $X(k) \cdot B(k)^*$ and $X(k)^* \cdot B(k)$ cannot be obtained directly and are approximated as $E[|B(k)|^2]$, $E[X(k) \cdot B(k)^*]$ and $E[X(k)^* \cdot B(k)]$ respectably, where $E[\,]$ denotes the expectation operator. Generally, $E[|B(k)|^2]$ is estimated during the silence periods, and we denote it by $|\hat{B}(k)|^2$. If speech signal $x(n)$ and noise signal $b(n)$ are assumed to be uncorrelated, in frequency domain the terms $E[X(k) \cdot B(k)^*]$ and $E[X(k)^* \cdot B(k)]$ reduce to zero. Thus from the above based assumptions, the estimate of clean speech signal can be represented as follows:

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2 \left(1 - \alpha \dfrac{|\hat{B}(k)|^2}{|Y(k)|^2}\right), & \text{if } case1 \\ \beta |Y(k)|^2 & \text{if } case2 \end{cases} \tag{3}$$

with

$$case1: |Y(k)|^2 - \alpha |\hat{B}(k)|^2 > \beta |Y(k)|^2$$
$$case2: |Y(k)|^2 - \alpha |\hat{B}(k)|^2 \le \beta |Y(k)|^2$$

Note that every procedure is carried out in frame-by-frame basis. Instead of power spectrum, its magnitude spectrum can be made available. With over-subtraction factor α and floor factor β, this algorithm regularizes the trade-off between noise reduction and residual noise. Note that the enhanced short-time power spectral amplitude $|\hat{X}(k)|^2$ depends on *a posteriori* SNR[3]:

$$SNR_{post} = |Y(k)|^2 / |\hat{B}(k)|^2 \tag{4}$$

2.2. Nonlinear Spectral Subtraction

Nonlinear spectral subtraction algorithm is as follows [4]:

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2 \left(1 - \dfrac{\Phi(k)}{|Y(k)|^2}\right), & \text{if } case3 \\ \beta |Y(k)|^2 & \text{if } case4 \end{cases} \tag{5}$$

with

$$case3: |Y(k)|^2 - \Phi(k) > \beta |Y(k)|^2$$
$$case4: |Y(k)|^2 - \Phi(k) \le \beta |Y(k)|^2$$

Nonlinear function $\Phi(k)$ is calculated for each frame and can be chosen arbitrarily to implement the notion that relatively greater subtraction is applied to the low SNR region of spectrum and less subtraction to the high SNR region. In this experiments, we use the following function

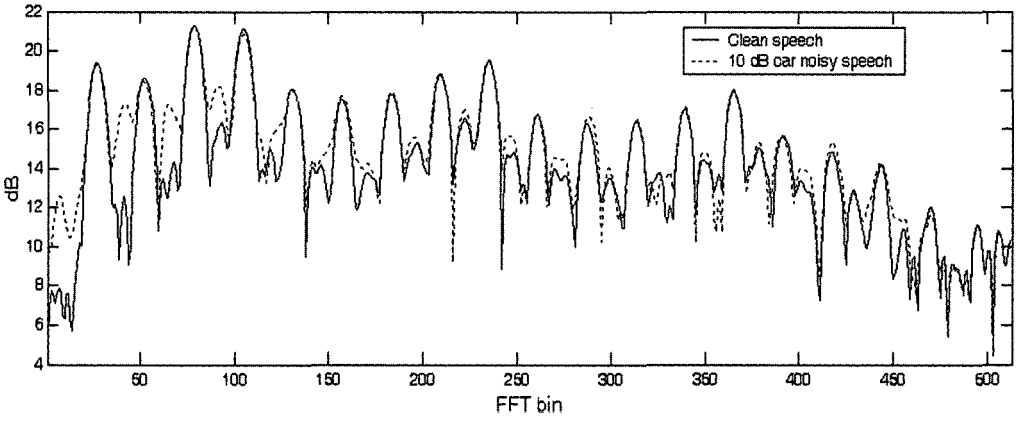

Figure 1. Example spectrum of a speech frame (pronunciation /oh/ by female).

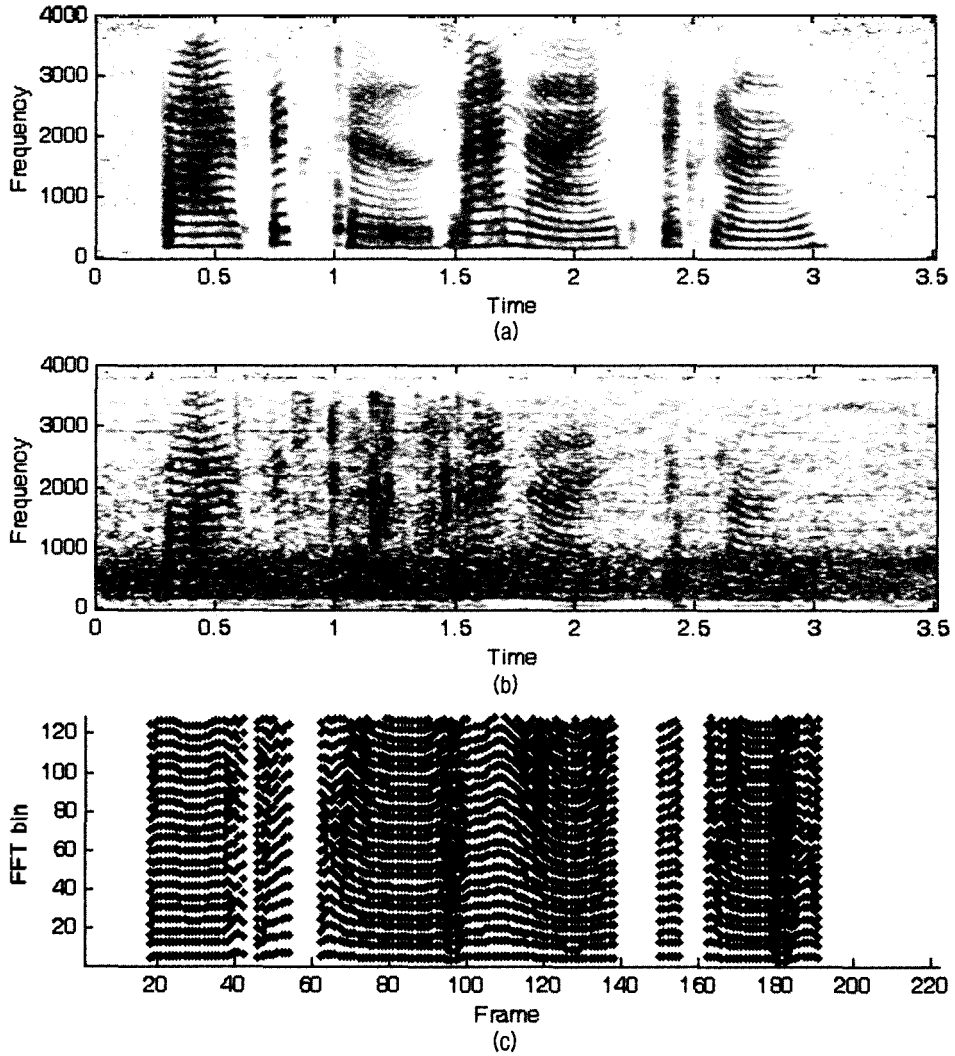

Figure 2. (a) Spectrum of the pronunciation 'five-six-two-nine-nine-six-nine' by female, (b) Corrupted version by 5 dB car noise, (c) Result of the harmonics detection.

and it is described more precisely in [4].

$$\Phi(k) = \alpha_i(k) F(k), \tag{6}$$

with

$$\alpha_i(k) = \max_{i-40 \le r \le i} |B_r(k)|^2 \tag{7}$$

$$F(k) = 1 - \left(1 - \frac{|\hat{B}(k)|^2}{\alpha_i(k)}\right) lin(\rho_{\min}, \rho_{MAX}) \tag{8}$$

where i is the frame index. $lin(\rho_{\min}, \rho_{MAX})$ refers to a linear weighting function of a posteriori SNR with the smoothed estimate of the corrupted speech $|\overline{Y_i(k)}|^2$ and the noise estimate in Eq. (14).

$$\rho(k) = \frac{|\overline{Y}(k)|^2}{|\hat{B}(k)|^2} \tag{10}$$

with

$$|\overline{Y_i}(k)|^2 = \lambda_Y |\overline{Y_{i-1}}(k)|^2 + (1-\lambda_Y)|Y_i(k)|^2, \; 0.1 \le \lambda_Y \le 0.5 \tag{11}$$

In general, the spectral amplitude of speech components (e.g. spectral harmonics) is higher than that of noise or the side lobes among harmonics so that those algorithms are rather suitable for reasonably high SNR (about 15~20 dB) noisy speech cases. However, it appears that in the case of low SNR environments (especially 0~10 dB), such that when the noise level is as close in amplitude as that of speech, there occurs unnecessarily subtracted speech

region or less subtracted noisy region in noisy speech spectrum. Consequently, instead of the mundane use of SNR as a measure for subtraction, we need a new and better measure for activating the subtraction procedure. In particular, the subtraction over spectrum requires a more accurate measure than mere SNR in order to apply the subtraction rule, which is selective to speech-dominant region vs. noise-dominant region of spectrum.

III. Proposed Algorithm

3.1. Speech Dominant Region vs. Noise Nominant Region

In speech, it is observed that the voiced speech segment has peaks positioned periodically in spectrum due to the vibration of vocal cords. These points are very critical in speech sound perception. Figure 1 illustrates this phenomenon with a sample spectrum of speech frame capturing the pronunciation /oh/ in one utterance contained in Aurora2 corpus. Note that at the peaks, their amplitudes are far greater than the amplitudes at the points between adjacent small peaks, or side lobes. Also, it is observed that the degree of corruption by the noise in the peaks is not as much as the degree of corruption at the points over the side lobes. From this observation, it can be deduced that in speech spectrum, the speech-dominant regions exist over or near the peaks and the noise-dominant regions exist over or near the side lobes.

The fundamental frequency can be obtained roughly by auto-correlation and then we can find the peak points from the roughly obtained fundamental frequency. For the frequency regions covering the peak points and their vicinity, we apply a small over-subtraction factor and large floor factor. In other regions, we apply a large over-subtraction factor and small floor factors. The detailed procedure is described in rest of Section 3.

Note that all procedures in the proposed algorithm are done on a frame-by-frame basis.

3.2. Peak Points Detection and Segmentation

It is known that the spectral harmonics are located at the points of multiples of fundamental frequency[5]. First, using autocorrelation function, the indices of local maxima in autocorrelation values are found. Among those indices, we can find the fundamental frequency f_0 by taking the reciprocal of the index that represents the maximum value. Secondly, by applying the nonlinear smoothing method [5], f_0 is modified to better reflect the true fundamental frequency. This algorithm works well even in high degree noisy environment. Autocorrelation function is expressed as follows[6]:

$$\phi(\tau)=\frac{1}{N}\sum_{n=0}^{N-1}x(n)x(n+\tau) \tag{12}$$

We then establish the index as k_0 for the frequency bin that corresponds to f_0. Using k_0, the peak points (harmonics) h_l are determined as $h_1=k_0$, $h_2=2k_0$, $h_3=3k_0$, ..., $h_L=Lk_0$ where L is the index of the last component of harmonics. Using this procedure, we accomplish determining all harmonics in the spectrum of input frame, which are assumed to be speech-dominant region. We illustrate the validity of the proposed method with spectrograms and the result of peak point detections in Figure 2.

For the purpose of implementing the proposed scheme (Section 3.4), frequency axis is divided into several non-overlapping bands in the following form.

$$\begin{aligned}&[1\ \ h_1],\left[h_1\ \ h_1+\frac{k_0}{2}\right],\left[h_1+\frac{k_0}{2}\ \ h_2\right],\left[h_2\ \ h_2+\frac{k_0}{2}\right],\ldots,\\&\left[h_{L-1}\ \ h_{L-1}+\frac{k_0}{2}\right],\left[h_{L-1}+\frac{k_0}{2}\ \ h_L\right],\left[h_L\ \ \frac{FFTorder}{2}+1\right]\end{aligned} \tag{13}$$

3.3. Voice Activity Detection (VAD) and Noise Estimation

Note that in each input frame, whether it is the speech frame or not, the fundamental frequency is calculated. As a result, the value of autocorrelation is made available at each frame. Since this value appears to be more effective measure to distinguish speech region from non-speech region than the mundane energy measure, so we employ the presence of fundamental frequancy instead of energy for VAD. Then, two separate procedures are employed for

robust detection of speech vs. non-speech region. First, we take a logarithm to this value, and then a smoothing procedure is carried out. If this value is greater than the mean value of autocorrelation obtained during the first 5 frames by 0.3, the relevant frame is decided as 'speech frame'. Secondly, for frames determined as 'speech frame', if the detected fundamental frequency is not in 50~400 Hz, then it is concluded as 'non-speech frame' once again. It is well known that 50~400 Hz is considered an appropriate fundamental frequency range for human voice. Input speech is coded into 32 ms frames, with a frame-shift of 16 ms. Then short-time FFT is applied. We performed 256 points FFT analysis on the relevant input frame.

For the noise estimation, we assume that any starting input speech is followed by a silence or background noise segment corresponding to 5 frames (96 ms). In that duration, mean of short-time spectral amplitude at each frequency bin are calculated. Then, if input frame is determined as 'non-speech frame', the estimated noise spectral magnitude is updated as follows:

$$|\overline{B_i}(k)|^2 = \lambda_B |\overline{B_{i-1}}(k)|^2 + (1-\lambda_B)|B_i(k)|^2,\ 0.65 \le \lambda_B \le 0.998 \tag{14}$$

where i is the frame index and k is the index of FFT bins and we use $\lambda = 0.75$.

This procedure is for the purpose of applying different subtraction rule to the speech frame and the non-speech frame. It is because of the 'non-speech frame', whose harmonics components cannot be determined. Consequently, the proposed scheme is inappropriate to 'non-speech frames'.

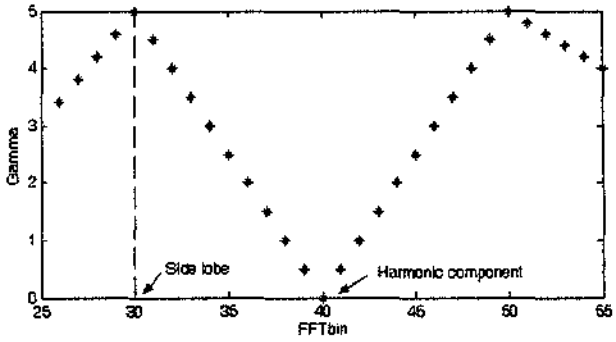

Figure 3. An illustration about implementing over-subtraction factor to FFT bins.

3.4. Spectral Subtraction

Based on the result of VAD procedure, we apply different subtraction rule.

3.4.1. Speech Frame

In order to implement the proposed scheme, we designed following simple linear function.

- If $k \in [1 \quad h_1]$ or $k \in \left[h_{l-1} + \frac{k_0}{2} \quad h_l\right]$, then

$$\gamma(k) = \frac{\alpha_{\min} - \alpha_{MAX}}{h_l - \left(h_{l-1} + \frac{k_0}{2}\right)}(h_l - k) + \alpha_{MAX} \tag{15}$$

$$\delta(k) = \frac{\beta_{MAX} - \beta_{\min}}{h_l - \left(h_{l-1} + \frac{k_0}{2}\right)}(h_l - k) + \beta_{\min}. \tag{16}$$

- If $k \in \left[h_l \quad h_l + \frac{k_0}{2}\right]$ or $k \in \left[h_L \quad \frac{FFTorder}{2} + 1\right]$ then

$$\gamma(k) = \frac{\alpha_{MAX} - \alpha_{\min}}{h_l - \left(h_{l-1} + \frac{k_0}{2}\right)}(h_l - k) + \alpha_{\min}. \tag{17}$$

Table 1. Word accuracy in mis-matched traing condition (%).

SNR	Baseline	SS	SS_D	NSS	NSS_D	Proposed
clean	99.02	98.78	98.78	98.75	98.75	98.81
20 dB	97.97	98.27	98.51	98.18	98.21	98.36
15 dB	94.24	97.08	97.55	97.26	97.20	97.52
10 dB	78.17	91.59	94.15	93.95	93.83	94.66
5 dB	42.59	79.39	81.78	84.61	84.49	85.00
0 dB	14.67	31.26	50.58	54.64	54.49	56.13
-5 dB	9.25	10.47	17.78	17.95	17.86	20.52
Avg.	69.79	79.52	84.51	85.73	85.64	86.33

$$\delta(k) = \frac{\beta_{\min} - \beta_{MAX}}{h_l - \left(h_{l-1} + \frac{k_0}{2}\right)}(h_l - k) + \beta_{MAX}. \quad (18)$$

$\gamma(k)$ applies the maximum over-subtraction factor to the middle point at each components. Then, over-subtraction factors of points exist in the interval between those points are interpolated linearly. Figure 3 illustrates a shape of $\gamma(k)$ which of fundamental frequency is about 156 Hz when sampling rate is 8 kHz and 1,024 points FFT analysis is applied.

On the contrary, $\delta(K)$ applies the maximum floor factor to each harmonic component and the minimum floor factor to the middle point at each component. Then, other floor factors between those points are generated by linear interpolation. The proposed subtraction rule can be represented as follows:

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2\left(1 - \gamma(k)\frac{|\hat{B}(k)|^2}{|Y(k)|^2}\right), & \text{if } case5 \\ \delta(k)|Y(k)|^2 & \text{if } case6 \end{cases} \quad (19)$$

with

$case5$: $|Y(k)|^2 - \gamma(k)|\hat{B}(k)|^2 > \delta(k)|Y(k)|^2$
$case6$: $|Y(k)|^2 - \gamma(k)|\hat{B}(k)|^2 \le \delta(k)|Y(k)|^2$

The value of parameters used are as follows:

$$\alpha_{MAX} = 5,\ \alpha_{\min} = 2,\ \beta_{MAX} = 0.2,\ \beta_{\min} = 0.05 \quad (20)$$

3.4.2. Non-speech Frame

We applied the subtraction rule as same as conventional method using Eq. (1) with α_{MAX}, $\beta_{\min}$ in Eq. (14).

IV. Experimental Results

4.1. Experimental Conditions

In these experiments, utterances corrupted by car noise among Aurora2 corpus are used. For comparison, spectral subtraction algorithm[7] and nonlinear spectral subtraction algorithm[4] are evaluated. Throughout all experiments, the VAD procedure and the noise estimation procedure are the same as the proposed method. Finally, for each algorithm, the enhanced speech signals are recovered by overlap-and-add manner followed by inverse FFT on frame-by-frame basis. Then, all performances are evaluated.

In Table 1, the meanings of the each abbreviation are as follows:

- 'SS' : general power spectral subtraction.
- 'SS_D' : general power spectral subtraction + different subtraction rule.
- 'NSS' : nonlinear spectral subtraction.
- 'NSS_D' : nonlinear spectral subtraction + different subtraction rule.

Experiments are evaluated with different subtractions rule as it s applied in the proposed method. That is, if the input frame is decided as 'non-speech frame', we applied the rule in section 3.4.2 which is the conventional subtraction rule Eq. (1) with the over-subtraction factor α_{MAX} and floor factor $\beta_{\min}$.

4.2. Experimental Results

'Avg' denotes the mean value of word accuracy over 0~20 dB. From the Table 1, the proposed method is seen effective at low SNR cases. The average word accuracy of the proposed method also shows that it is superior over other spectral subtraction approaches.

V. Conclusions

This paper proposes an efficient spectral subtraction scheme focused to specifically low SNR noisy environments by distinguishing the speech-dominant segment from the noise-dominant segment in speech spectrum. Consequently we let the shape of the spectral harmonics be preserved more clearly in noisy environments, which are very critical in speech sound perception. Representative experiments confirm the superior performance of the proposed method to conventional methods. In particular, the proposed method is seen effective at low SNR cases (< 10 dB). The average word accuracy of the proposed method also shows that it is superior to other recently introduced approaches. The experiments are conducted using car noise-corrupted utterances of Aurora2 corpus.

References

1. J. Jensen, and J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, 9 (7), 731-740, 2001.
2. D. Ealey, H. Kellher, and D. Pearce, "Harmonic tunneling: track-ing non-stationary noises during speech," *Eurospeech*, 437-440, 2001.
3. N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, 7 (2), 126-137, 1999.
4. P. Lockwood, and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, 11, 215-228, 1992.
5. W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.
6. L. Rabiner, and R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
7. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by additive noise," *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 208-211, 1979.
8. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transaction on Acoustics, Speech and Signal Processing*, 27 (2), 113-120, 1979.

[Profile]

• Junghoon Beh

2001, B. S. degree in elctronical, electronics and radio engineering from Korea university.
2003, M. S. degree in electronics and computer engineering from Korea university Now, Ph. D student in electronics and computer engineering from Korea university.

• Hanseok Ko

1982, B. S. degree in electrical engineering from Carnegie Mellon University.
1996, M. S. degree in systems engineering from University of Maryland, College Park.
1998, M. S. degree in electrical engineering from the Johns Hopkins University.
1992, Ph. D. degree in electrical engineering from the CUA.
1992~1995, Adjunct faculty member in the Dept. of Electrical Engineering at UMBC.
2001, Visiting Professor in the Dept. of ECE, Johns Hopkins University.
1995~present Faculty member in the Dept. of Electronics and Computer Engineering at Korea University.