Pitch Detection by Synchronizing the Phase of Noise-Corrupted Speech Signals

위상 동기화에 의한 잡음 음성의 피치 검출

(Byung-Gook Lee*, Myungjin Bae**, Souguil Ann*)

이 병 국*, 배 명 진**, 안 수 길*

※본 연구는 한국 전자통신연구소의 장기기초 연구비 지원으로 이루어졌음.

ABSTRACT

A new pitch detection algorithm is proposed. It takes advantage of the fact that if the phases of the fundamental and its harmonics are synchronized, the superimposed waveform shows peaks at the same peak position of the fundamental. We set the phase of the Fourier-transformed speech signal to zero, effectively synchronizing the phases of all harmonics. The algorithm gives robust performance even in 0 dB SNR environment, with a gross error rate of 3.63%. The gross error for clean speech is only 0.18%. It also exhibits good pitch resolution since the decision logic works in the time domain. Overall experimental results indicate that the proposed algorithm is quite effective for pitch detection.

요 약

시간 영역에서 음성의 피치 정보를 추출하는 새로운 알고리즘을 제안한다. 이 알고리즘은, 위상이 일치하는 고조파 성분의 합은 위상이 일치하지 않는 고조파 성분의 합의 경우보다 주기 정보를 분명히 나타낸다는 사실을 이용한 것이다. 즉, 음성 신호의 위상 성분을 0으로 되도록 하여 실질적으로 기본파와 모든 고조파 성분의 위상을 일치시킨다. 이 알고리즘은 잡음이 없는 음성의 경우 0.18%의 조오류(gross error)를 보이며, 0 dB SNR의 경우에도 3.63%의 조오류를 보임으로써 잡음에 강건한 성질이 있음을 알 수 있다. 또한 시간 영역에서의 결정 논리를 사용하므로 피치 해상도가 우수하다. 전반적인 실험 결과는 제안된 알고리즘이 피치 검출에 상당히 효율적임을 나타낸다.

I. Introduction

Pitch determination is one of the most important problems in speech signal processing. Numerous methods for detecting the fundamental frequency of voiced speech signals have been developed and proposed during the past 20 years, only with problems of their own which are yet to be solved. Prior to 1980s when integrated circuits were still relatively underdeveloped, researchers concentrated on reducing the amount of computation and thereby shortening the length of processing time, while after many digital signal processing chips became commercially available, the chief aim of algorithm development shifted to detecting the fundamental frequencies (or pitch periods) as accurately as possible.

Measuring the pitch periods of speech signals accurately and reliably is very difficult because of the reasons listed below[1].

The first reason is that the glottal excitation

*Dept.of Electronics Engineering, Seoul National University.
**Dept.of Electronics Engineering, Hoseo University.

waveform is not a pulse train that is perfectly periodic. Speech is in principle a nonstationary process : the instantaneous position of the vocal tract may change abruptly and lead to drastic variations in the temporal structure of the signal, even between adjacent pitch periods. The second difficulty we encounter is the interaction between the vocal tract and the glottis. The formant of the vocal tract may, in certain instances, completely alter the structure of the glottis waveform, making it difficult to detect actual pitch periods. Generally this interaction occurs when the speech-generating anatomy changes suddenly, and is harmful in detecting pitch periods.

The third reason is that it is hard to define the exact beginning and ending locations of each pitch period for voiced speech signals. This is because selecting the exact beginning and ending locations of pitch periods can be arbitrary. For example, the zero crossing intervals inside a period can be a candidate for defining the beginning and ending points of a pitch period. The fourth is that the distinction between unvoiced sound and voiced low-level voiced sound is difficult to be precise. It is not an easy task to find the exact boundary since in many cases the change from an unvoiced interval to a low-level voiced interval, or vice versa, is not very abrupt.

The fifth problem is that of additional complexity encountered in pitch extraction from speech signals transmitted over telephone lines. Most of the systems that require pitch extraction have to process toll-quality speech. The telephone system adds noise to the speech signals, while including linear and nonlinear processing steps. If we consider this telephone system as a linear filter, it can also be regarded as a bandpass filter (low cutoff frequency of about 200 Hz, high cutoff frequency of about 3200 Hz) that can weaken the fundamental frequency and its harmonic components. This makes it more difficult to detect periodicity. Nonlinear characteristics of the telephone system with respect to speech signals depend on the channel of the transmission system.

To summarize, the basic considerations in detecting the fundamental frequencies of speech signals are [2] :

1) The algorithm must be immune to individually varying characteristics, such as sex or age.

2) It must emcompass a wide range of frequency to represent the emotions and feelings of a speaker.

3) It must produce usable results even in noise or under the influence of the transmission channel.

4) It must be able to detect the pitch period regardless of consonant combinations.

Detecting the fundamental frequency, or pitch, is important in many speech processing applications. It can be used in eliminating the individual influences for speaker-independent speech recognition. It can also be utilized in speech synthesis to easily change or maintain naturalness and individual characteristics. Finally it may be applied to speech analysis, aiding in pitch-synchronous analysis, or in obtaining accurate vocal tract parameters which are removed the influence from glottis characteristics[3,4].

I. Methods for detecting fundamental frequencies

The pitch determination algorithms (PDAs) that have been developed so far can be categorized as pertaining to the time domain, the frequency domain, and the time-frequency hybrid domain methods. We briefly review these respective methods before we go on to Section II.

1) Time domain methods[3,4,5,6,7] :

Typical pitch period range in human speech is 2.5−25 msec, and this translates to 20−200 samples when the speech signal is sampled at 8 kHz. Thus this method has nice time resolutions. There is a setback however. That is, speech is usually processed on a frame basis, and changes in phonemic contents can occur within a single frame. This introduces error in estimating the envelope contour of the speech waveform, and in turn affects the extraction of pitch periods.

Moreover, pitch detection performance is also hampered when the transmission channel is noisy or other background noise is added.

Most algorithms extracting the pitch period in the time domain emphasize the periodicity of the voiced speech waveform before deciding on the period using a decision logic. Emphasizing periodicity eliminates the influence of formant frequencies, which is due to the resonance in the vocal tract, and puts emphasis only on the fundamental frequency of the excitation source.

Successful pitch detecting algorithms in the time domain are parallel processing technique, autocorrelation method, average magnitude difference function (AMDF) method, simple inverse filter tracking (SIFT) algorithm, and harmonics matching technique. We will not delve into specifics about each of these algorithms : readers are referred to other excellent review papers or books.

In summary, these time domain methods do not need to perform any transform into frequency domain and reduces the time required for finding pitch periods. They also exhibit good resolution capabilities. A problem for these methods is that they are rather weak and prone to errors when there are phonemic transitions within an analysis frame.

2)Frequency domain methods[9,11] :

Frequency domain methods usually measure the spectral distance between the harmonics of speech spectrum to obtain the fundamental frequency. Spectral pitch detectors give more accurate estimates than time domain approaches but require about an order of magnitude more calculations due to the transformation to the frequency domain. The transformation focuses information about speech periodicity in a way that time domain methods cannot. The spectrum in general is obtained on a frame basis (typically 20−40 msec long), hence the influence of phonemic transitions or shift, or background noise is averaged over this transform interval, lessening the degrading effect. However it takes longer to process when one wants higher frequency resolution and decides to increase the number of FFT points. This computational disadvantage is not much of a problem any more thanks to recent development of fast digital signal processing chips.

3)Time-frequency hybrid techniques[8,10] :

This technique combines both the time and the frequency domain methods to take advantage of all the good things in both. In short, this technique tries to ahieve good time resolution and good accuracy. Cepstrum method and spectra comparison method fall into this category. This hybrid technique also requires a lot of computations due to time-frequency domain transforms. But the advent of DSP processors has alleviated the problem.

We propose a novel algorithm for detecting the fundamental frequency in speech signals[12]. Our algorithm sets the phase component of the transformed speech signals identically to zero, and then does inverse-transform to easily find the pitch period in the time domain. In Section II we look into the period and frequency characteristics of speech signals, and justify our algorithm and explain the decision logic. In Section III we show the experimental results. Finally in Section IV we summarize and draw conclusions.

Ⅱ. Phase Synchronization in the Frquency Domain

1. Characteristics of speech signals

Voiced speech is generated by the vibration of the vocal cords. And the vocal cords determine the pitch period. Moreover we observe from two to four resonance peaks of the vocal tract, which is called formants. The energy tends to group around these formant frequencies, showing distinct difference in energy level compared to those of non-peaks. Also, in the frequency domain, the fine structure corresponding to the integer multiples of the fundamental period in the time

domain appears as harmonics, and the spectral distance between these harmonics is termed the fundamental frequency F_0, or the reciprocal of the pitch period. That is, we can determine the pitch period (fundamental frequency) by measuring the spectral distance between the harmonic frequencies. But this method falls short on the precision aspect since the frequency resolution is rather poor.

2. Synchronizing the phase of the speech spectrum[13]

Assuming that the frequency of the speech signal s(t) is bandlimited to $f_s/2$ in the time domain, where f_s is the sampling frequency, the Fourier series representation of s(t) is

$$s(t) = \sum_{k=-N}^{N} S(k)\, e^{j2\pi kt/T} \tag{1}$$

The spectral speech signal S(k) is written as a combination of the real part and the imaginary part, as in eq. (2),

$$S(k) = \mathrm{Re}[S(k)] + j\mathrm{Im}[S(k)] \tag{2}$$

where the magnitude expression equals

$$P(k) = |S(k)| \tag{3}$$

and the phase expression is

$$\psi(k) = \tan^{-1}\left\{\frac{\mathrm{Im}\,[S(k)]}{\mathrm{Re}\,[S(k)]}\right\} \tag{4}$$

We cite an intuitive example to explain our idea of extracting pitch information by controlling the phase of speech spectra. Three curves, one fundamental and others its two harmonics, are shown in Fig.1. If the respective phase of these sinusoids is different, the superimposed waveform will show little uniformity within a period of the fundamental. This complex shape, as appears in Fig.1(b), gives harmful effects in extracting the pitch period in the time domain and thus makes the pitch decision logic more complex in turn. Moreover the logic has to become even more complex when the speech signal is subjected to noise, since it affects the phase of the observed signal.

On the other hand, when we synchronize these three sinusoids, that is, make them to possess the same phase and add them, the final waveform would peak at the position where the three sinusoids are identically at their peaks. The period of the peak position coincides with that of the fundamental sinusoid, as can be seen in Fig. 1(c). We utilize this fact in extracting the pitch information from the speech signal.

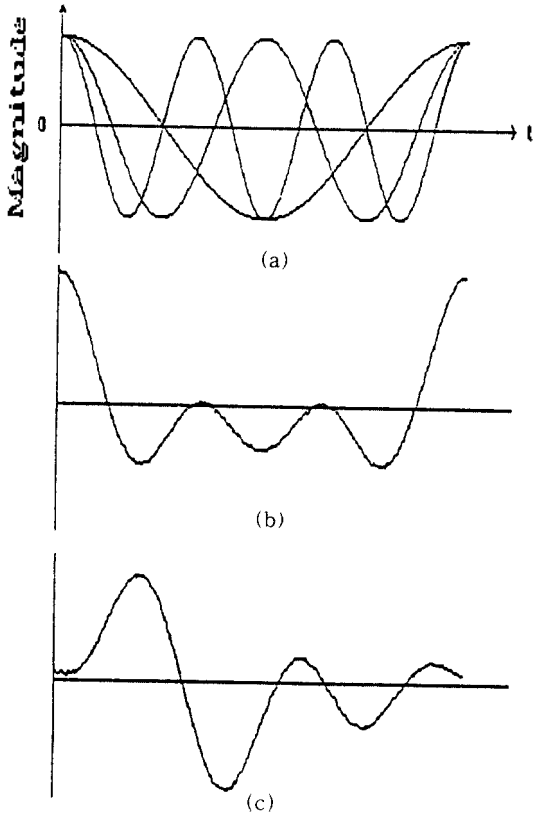
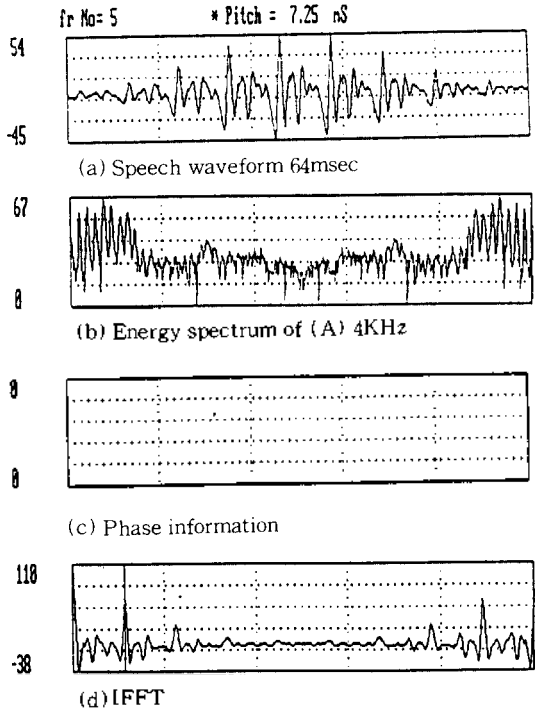


Fig. 1. The fundamental and its two harmonics showing the effect of phase synchronization,
(a)the three sinusoids,
(b)waveform when unsynchronized sinusoids are added,
(c)waveform when synchronized sinusoids are added.

The spectrum of voiced speech signals consists of the sum of the spectra of the harmonics. We

would get the same result with this speech spectrum as with the example of the three harmonic sinusoids when we set the phase of all frequency components identically to zero. And it would allow us to easily extract the pitch period information in the time domain, making the pitch stand out more clearly.

In other words, the speech signal shown in Fig. 2(a) first undergoes Fourier transformation and is represented in its real part and imaginary part form, where the energy spectrum is shown in Fig. 2(b). Using the fact that the spectral phase $\psi(k)$ equals $\tan^{-1}$ (Im [S(k)]/Re [S(k)]), the imaginary part Im [S(k)] of the speech spectrum S (k) is set to zero for all k, as in Fig.2(c). Then all the harmonics would be synchronized in the time domain. The speech spectrum, with its phase component set to zero, is then inverse Fourier transformed to come back to the time domain.



(a) Speech waveform 64msec

(b) Energy spectrum of (A) 4KHz

(c) Phase information

(d) IFFT

Fig. 2. The effect of phase synchronization for real speech,
(a)speech waveform, (b)energy contour,
(c)setting phase to zero,
(d)resulting time-domain waveform after IFFT.

The signal in the time domain in Fig. 2(d) now clearly shows the pitch period information that was rather obscure before setting the phase to zero.

Next, we know that typical range of the pitch period is from 2.5 msec to 25 msec. If the speech signal is sampled at 8 kHz, the pitch period would then be from about 20 to 200 samples. The position of the highest peak within this time interval equals simply the pitch period.

Ⅲ. Experimental Results

The experimental procedure is shown in the block diagram of Fig. 3.

The speaker was a 28-year old male. The sample sentences used were "Insoo's young boy likes a genius kid", "Speech signal processing team at the department of electronics engineering, Hoseo University", "Jesus spoke of the lessons of the creation of the earth", and "Thank you", all spoken in Korean. The speech signal was stored in an IBM 386 computer after being low-pass filtered at 8 kHz and quantized at 12 bits. The stored speech was then windowed, the frame length being 512 samples and overlapping by 384 samples, sliding 128 samples at a time. The rest of the procedure is as been explained in Section II.

Fig. 4(a),(b),(c) and (d) show the result of processing noiseless speech. The peak clearly appears in Fig. 4(d), allowing us to detect the pitch period quite easily. Next we added zero-mean white Gaussian noise to the uncorrupted speech to make the signal-to-noise ratio (SNR) to 0 dB. The result from this noise-corrupted speech is shown in Fig. 4(e). We can still see the peak corresponding to the pitch period, which stands out among smaller peaks due to added noise. This figure shows that our algorithm is able to extract pitch periods easily and effectively even when the SNR is 0 dB.

We show in Table 1 the gross error rates for each speech sample. The gross error rate is de-

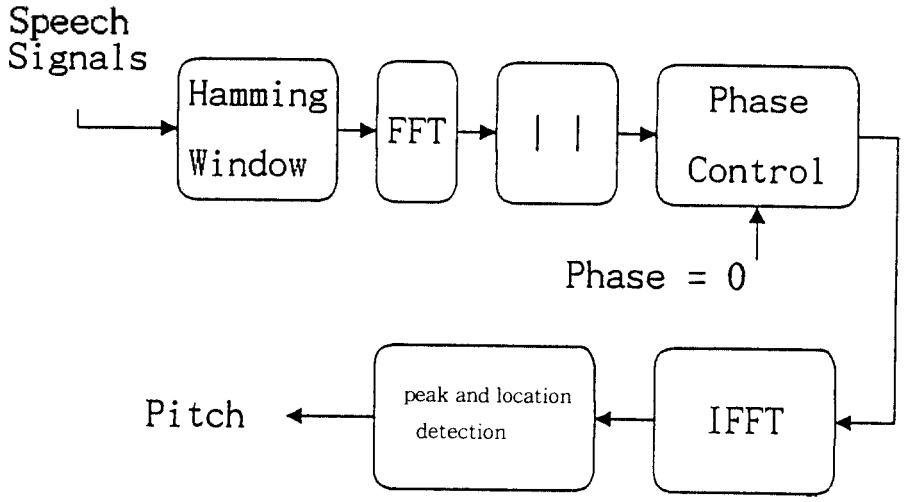


Fig. 3. Block diagram of the experimental procedure.

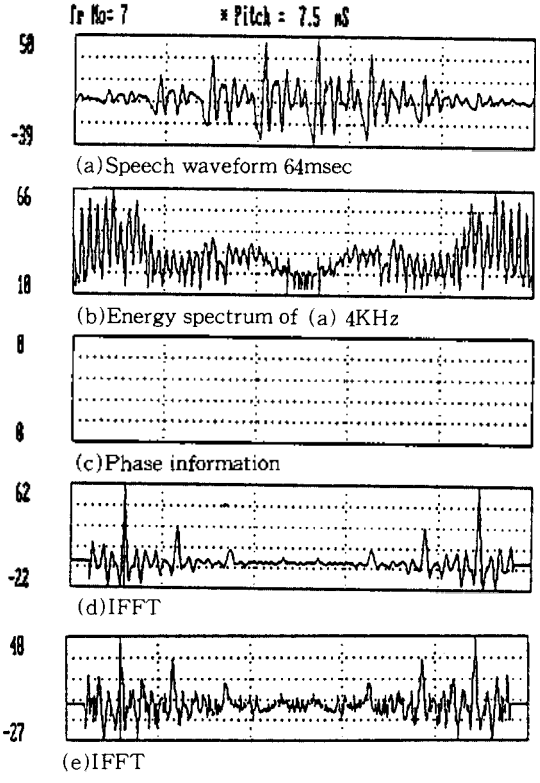


Fig. 4. The result of the proposed algorithm ;
(a)original waveform, (b)energy spectrum,
(c)phase control,
(d)time-domain waveform after IFFT,
(e)same as (d)but with 0 dB SNR.

fined as follows [1] : we compare the result of our algorithm with that of eye detection. When the result of our algorithm differs with the eye-detected result by more than 1 msec for a frame, we increase the error count by 1. This 1 msec corresponds to 8 samples. There are 62 frames in one second of speech. If there are 5 frames that contain errors, the gross error in that case would be

$$(5/62) \times 100 = 8.06(\%).$$

Examples of the gross errors are the case of doubling or tripling, or the effect of formant on pitch detection.

We did not consider the fine error, which is the error of less than 1 msec time difference, since

Table 1. The gross error rates for each speech sample.

speech sample	no. of analyzed frames	gross error rates(%)			
		clean speech	SNR 6dB	SNR 3dB	SNR 0dB
1	192	0.00	1.04	1.04	3.64
2	192	0.00	0.52	1.04	3.12
3	192	0.52	1.04	1.04	3.12
4	64	0.00	0.00	0.00	1.56
average	550	0.18	0.91	1.09	3.63

there were virtually none. This is because we employed a time-domain decision logic that looks for only the largest peak. Fine errors occur when the pitch detector let the resolution become poor to reduce computation, or when the resolution in the transform domain is low.

Ⅳ. Conclusions

We proposed a new pitch detection algorithm. It utilizes the fact that if the phases of the fundamental and its harmonics are synchronized, the superposed waveform shows peaks at the same position as the fundamental. The experimental results support our argument, and our algorithm is quite capable of detecting pitch periods in adverse conditions. Table 1 shows that our algorithm performs well regardless of the speaker's age or sex. Moreover our algorithm extracts pitch information in the time domain and hence has good resolution and is almost identical to eye-detected results. Its decision logic is very simple since it looks only for the largest peak in the time domain. Our method may be used in such diverse applications as speech enhancement, speech recognition, or speech analysis.

References

1. M. BAE, "A Study on the Pitch Detection and Alteration of Speech Signals," Ph. D. Dissertation, Seoul National University, 1991.
2. L. R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C. A. McGonegel, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoust., Speech, and Signal Processing, Vol.24, no.5, Oct. 1976.
3. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
4. P. E. Papamichalis, *Practical Speech Processing*, Prentice-Hall, 1987.
5. M. BAE, D. SHIN and S. ANN, "The Pitch Extraction of Voiced Speech by the Comparison Between the Original and the Repeated Partial Waveform," J. Acoust. Soc. Korea, vol.7, no.5, Nov. 1988.
6. M. BAE and S. ANN, "The High Speed Pitch Extraction of Speech Signals Using the Area Comparison Method," KITE, vol.22, no.2, pp.101-105, Feb. 1985.
7. L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans. Acoust., Speech, and Signal Processing, vol.ASSP-25, pp.24-33, Feb. 1977.
8. M. BAE and S. ANN, "Fundamental Frequency Estimation of Noise Corrupted Speech Signals Using the Spectrum Comparison," J. Acoust. Soc. Korea, vol.8, no.3, pp.57-64, 1989.
9. S. Seneff, "Real Time Harmonic Pitch Detector," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-26, pp.358-365, Aug. 1978.
10. M. BAE and S. ANN, "On the Time-Frequency Hybrid Technique for Detecting the Pitch of Noise Corrupted Signals (Time Domain Processing)," J. Acoust. Soc. Korea, vol.9, no.1, 1990.
11. M. Lahat, R. J. Niederjohn, and D. A. Krubsack, "A Spectral Autocorrelation Method for Measurement of the Fundamental Frequency of Noise Corrupted Speech," IEEE Trans. Acoust., Speech, and Signal Processing, vol.ASSP-35, no.6, pp. 741-750, June, 1987.
12. Chansou PARK, Kinam HAN, Myungjin BAE, Souguil ANN, "On the Pitch Detection by Using Phase Control of Speech Signals," Proc. KITE Summer Conf., vol. 15, no.1, pp.677-680, June, 1992.
13. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, 1989.

▲Byung-Gook Lee

was born in Seoul, Korea in 1965. He received his B. Sc. and M. S. degrees in electronics engineering from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively. He is currently pursuing his Ph.D. degree at the same institution. His research interests include speech recognition, speech enhancement and noise cancelling. He is a student member of IEEE.

▲Souguil, Ann

has been professor at the Department of Electronics Engineering, Seoul National University, Seoul, Korea since 1969. He received his B. Sc., M. S. and Ph. D. degrees in electronics engineering from Seoul National University in 1957, 1959 and 1974, respectively. He was lecturer at the Department of Electronics Engineering, Korea Military Academy during 1957-1959. He worked at CEN Saclay Research Institute, France, as a research member during 1959-1960. From 1960 to 1963 he lectured at the Department of Electronics Engineering, Seoul National University and from 1964 to 1968 he worked on a ground tracking station project at Centre Nationale d'Etudes Spatielles, Paris, France. He was at the Aerospace Department, Schlumberger France, as a research member. He is currently serving as Director of Region 10, IEEE, and is consultant to Korea Telecom and the Ministry of Science and Technology, Government of Korea. Dr. Ann is member of Board of Directors, IEEE, and senior member of IEEE.

▲Myungjin Bae

was born in Kyungbook-do, on May 20, 1956. He received the B.S. degrees in electronics engineering from Soongsil University, Seoul, in 1981. He also received the M.S. and Ph.D. degree in electronics engineering from Seoul National University, Seoul, in 1983 and 1991, respectively.

Since 1986, he has been with the department of Electronics Engineering, Hoseo University, Chunan-si, where he is currently an Assistant Professor. His research interests include speech signal processing, adaptive signal processing, and communication system.