International Journal of Software Engineering and Its Applications
Vol.7, No.5 (2013), pp.129-136
http://dx.doi.org/10.14257/ijseia.2013.7.5.13

On A New Hybrid Speech Coder using Variables LPF

Seong-geon Bae and Myung-jin Bae

*[1]Information and Telecommunication Department,
Soongsil University, 1-1 Sangdo 5 dong DongJak-Ku,
Seoul, 156-743, Republic of Korea
sgbae123@empal.com, mjbae@ssu.ac.kr*

Abstract

To encode the speech quality with reduce the redundancy within samples that resulted from domain processing method like PCM and LPC, Source coding or Waveform coding methods can be considered. However, it is well known that when conventional sampling methods are applied directly to speech signal, the required amount of data is comparable to or more than that of uniform sampling method. To overcome this problem, a new hybrid methods is proposed, in which hybrid domain coding is applied to two low-pass filters in lower bandwidth and the remain signals are compensated by the Gaussian white signal and harmonics, which is used to get high quality speech in higher bandwidth.

Keywords: *Hybrid coding, Sampling, Quantization, Peak and valley, Hybrid coding, Waveform coding*

1. Introduction

In many speech communication settings, the presence of background interference causes the quality or intelligibility of speech to degrade. When a speaker and listener communicate in a quiet environment, information exchange is easy and accurate. However, a noisy environment reduces the listener's ability to understand what is said. In addition to interpersonal communication, speech can also be transmitted across telephone channel, loudspeakers, or headphones. The quality of speech, therefore, can also be influenced in data conversion, transmission, re-production. The purpose of many enhancement algorithms is to reduce background noise, improve speech quality, or suppress channel or speaker interference. In this paper, we discuss the general problem of speech enhancement with particular focus on algorithms designed to remove additive background noise for improving speech quality. In our paper, background noise will refer to any additive broadband noise component as white Gaussian noise, aircraft cockpit noise, or machine noise in a factory environment. Other speech processing areas that are sometimes included in a speech enhancement include suppression of distortion from voice coding algorithms, suppression of a competing speaker in a multi-speaker setting, enhancing speech as a result of a deficient speech production system, or enhancing speech for hearing-impaired listeners. Since the range of possible applications is broad, we will generally limit our discussion to enhancement algorithms directed at improving speech quality in additive broadband noise for speakers and listeners with normal production and auditory systems.

The problem of enhancing speech degraded by additive background noise has received considerable attention in the past two decades. Many approaches have been taken, each attempting to capitalize on specific characteristics or constraints, all with varying degrees or success. The success of an enhancement algorithm depends on the goals and assumptions

ISSN: 1738-9984 IJSEIA

used in deriving the approach. Depending on the specific application, a system may be directed at one or more objectives, such as improving overall quality, increasing intelligibility, or reducing listener fatigue. The objective of achieving higher quality and/or intelligibility of noisy speech may also contribute to improved performance in other speech application, such as speech compression, speech recognition, or speaker verification.

A major objective in speech coding is to compress the signal, that is, to employ as few bits as possible in the digital representation of the speech signal. This efficient digital representation of the speech signal makes it possible to achieve bandwidth efficiency in the transmission of the signal over a variety of communication channels, or to store it efficiently on a variety of magnetic and optical media. Since the digitized speech is ultimately converted back to analog form for the user, an important consideration in speech coding is the level of signal distortion introduced by the digital conversion process.

The conversion of the speech encoding method for storing or transmitting in signal can largely be classified into the waveform coding, the source coding and the hybrid coding. And to maintain intelligibility and naturalness, the waveform coding method is mainly used. This coding process is the method of storing and synthesizing after eliminating repeated, unnecessary remaining components, and it is consisted of PCM, ADPCM, ADM, *etc*. However, it has the disadvantage that large amount of memory is required due to enormous quantity of data.

Main point of speech coding is to process it by considering, especially, transmission and compression rate of data, signal quality of playback, and processing velocity among the information transmitted by the encoded data. In addition, there is the source coding method which is the parameter method to have improvement by detecting characteristic section of signal, and this has the disadvantage that the complexity and the calculation is too much to be applied to various signals. It is the hybrid coding method which is made by taking advantages of coding methods, and the improvements have been made for this owing to the development of computing. In this paper, in order to use this hybrid coding method, the method of coding with non-uniform quantization and sampling of recognized signal obtained from the time domain and with analyzing frequencies in frequency domain will be used.

Especially, in the case of voiced signal, as it is concentrated in the speech with large amplitude, the proposed method which is searching and using peak and valley in voiced signal, is using the Gaussian random in unvoiced signal.

At Chapter 2, the conventional waveform method which is to get recognizable characteristics of signal, and at Chapter 3, the proposed method to use the hybrid coding will be described, and at Chapter 4, experimental and results, and at Chapter 5, conclusion and study direction in the future will be presented.

2. Existing Method

The coding method to use peak and valley of signal by analyzing waveform characteristic has been presented, and this has been used by considering the recognizable aspect. Therefore, this recognizable aspect of speech signal will be affected by the peak and valley in the sampling and quantization. In this characteristic of important information for recognition of speech signal, the roles of peak and valley point become remarkably important. By utilizing this characteristic, numerous applications for synthesis or coding section are possible, and especially for noisy environment, great support have been conducted for searching important factors of recognition signal by considering characteristics of peak and valley.

$$y_k(n) = [\frac{M(k-1) - M(k)}{2}\cos(\frac{\pi n}{I(k)}) + \frac{M(k-1) + M(k)}{2}], \qquad 1 \le n \le I(k) \qquad (2\text{-}1)$$

Figure 1 shows these two sampling methods for linear sampling and non-uniform one. where, *M(.)* are the magnitudes as peaks and valleys of non-uniformly sampled data and *I(.)* are the intervals of them. The interpolation method of non-uniform sampling in speech reconstruction is used as cosine interpolation. The waveform reconstruction is performed by using a cosine interpolation method based on such parameters as the magnitudes and the intervals of the peak and the valley. The reconstructed waveform, $y_k(n)$, obtained by cosine interpolation method is represented as Eq. 2-1.

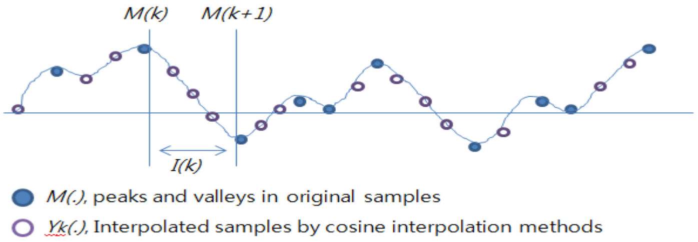
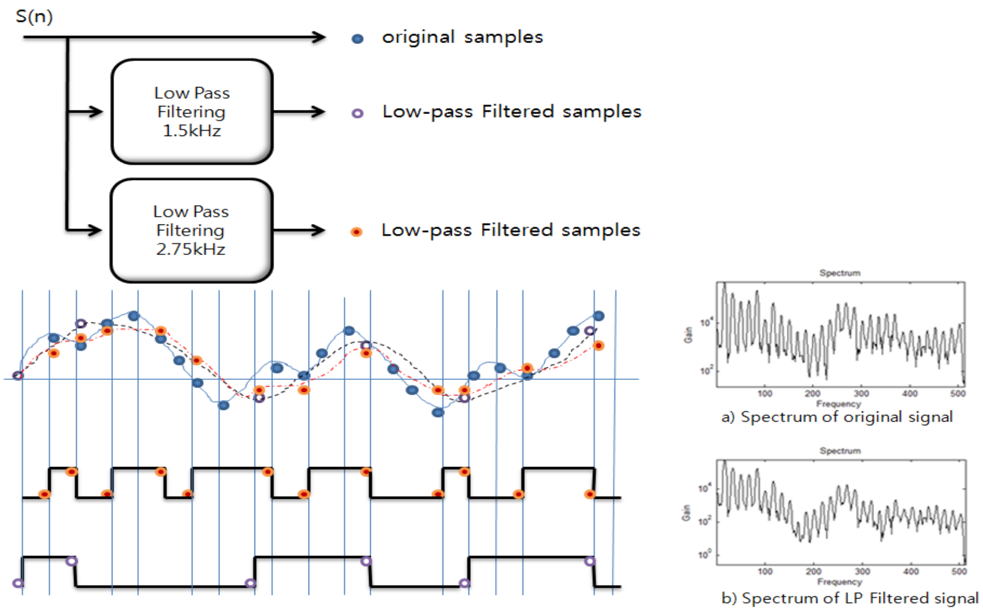


Figure 1. Examples of Cosine Interpolation Method

3. A New Proposed Method

According to the speech production mechanism, the 3rd and upper formants have broad bandwidths. Moreover, from the viewpoint of speech recognition, higher frequency band components are not significant, while the 1st and the 2nd formants are separated to reconstruct the high-intelligible speech. Therefore, the samples related to the frequency band higher than the 2nd formant are considered as redundant information in the speech perception.


S(n)
original samples
Low Pass Filtering 1.5kHz
Low-pass Filtered samples
Low Pass Filtering 2.75kHz
Low-pass Filtered samples
Spectrum
Gain
Frequency
Spectrum
Gain
Frequency

a) Spectrum of original signal

b) Spectrum of LP Filtered signal

Figure 2. Examples of Non-uniform Coding using LPF

The 1st and the 2nd formant frequencies in speech signal are less than 2.65 kHz. Also, the formants higher than this cut-off frequency have wide broad bandwidths. Therefore, non-uniform sampling can be only applied to the signal component of the original waveform less than 2.65 kHz to reduce loss of intelligibility. Since the low-pass filtered signal is smoother than the original one, fewer number of the peak and the valley sample are obtained when non-uniform sampling is performed on it. Also if the signal is voiced, the low-pass filtered signal is used by 2.65kHz, the unvoiced by 1.5kHz. So fewer number of the peak and the valley sample in voiced signal. But when unvoiced, the results is smaller it. So we consider it as a decision logic to detect the voiced or unvoiced signal. Then this makes it possible to achieve high quality and high compression ratio.

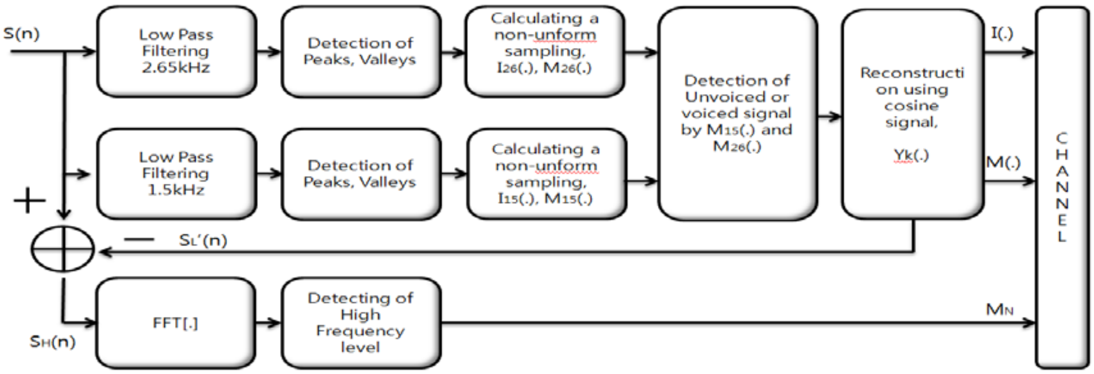

S(n)
Low Pass Filtering 2.65kHz
Detection of Peaks, Valleys
Calculating a non-unform sampling, I26(.), M26(.)
Detection of Unvoiced or voiced signal by M15(.) and M26(.)
Reconstruction using cosine signal, Yk(.)
I(.)
Low Pass Filtering 1.5kHz
Detection of Peaks, Valleys
Calculating a non-unform sampling, I15(.), M15(.)
M(.)
C H A N N E L
+
— SL'(n)
SH(n)
FFT[.]
Detecting of High Frequency level
MN

Figure 3. A New Speech Coding Scheme

Figure 3 shows the encode block diagram of the method proposed in this paper. In the encode block diagram, S(n) is speech signal digitized uniformly by A/D converter, and S_{L26}(n) is the law-pass filtered signal by 2.65 kHz as cutoff frequency, and S_{L15}(n), the law-pass filtered signal by 1.5 kHz. Then the conventional non-uniform sampling is applied to this two low-pass filtered signals and such parameters as the magnitudes, M(.) and the intervals, I(.) of the peak and the valley are quantized and sampled. At the same time, Re-constructed speech signal, S_L'(n), is reconstructed by the cosine interpolation as the non-uniform sampling.

$$S_L'(n) = [\frac{M(i-1)-M(i)}{2}\cos(\frac{\pi n}{I(i)}) + \frac{M(i-1)+M(i)}{2}], \qquad 1 \le n \le I(i) \qquad (3\text{-}1)$$

Then, the residual signal, S_H(n), between the original signal and the reconstructed signal is obtained in encoding. And then this is transmitted by the speech channel.

$$S_H(n) = S(n) - S_L'(n)\ , \qquad 1 \le n \le 512 samples \qquad (3\text{-}2)$$

where, *M(.)* is the magnitude of non-uniformly sampled data, *I(.)* is the interval of them and *i* are samples between *M(n)* and *M(n+1)* in the intervals. And, the Gaussian level, M_H is used by Eq. 3-3.

$$M_H = \frac{1}{N}\sum_{i=0}^{n-1} [S_f(i) - S_f'(i)]\ , \qquad 1 \le n \le 512 samples \qquad (3\text{-}3)$$

Where, $S_f(.)$ is the magnitude spectrum of *S(n)*, $S_f'(i)$ is $S_L'(i)$.

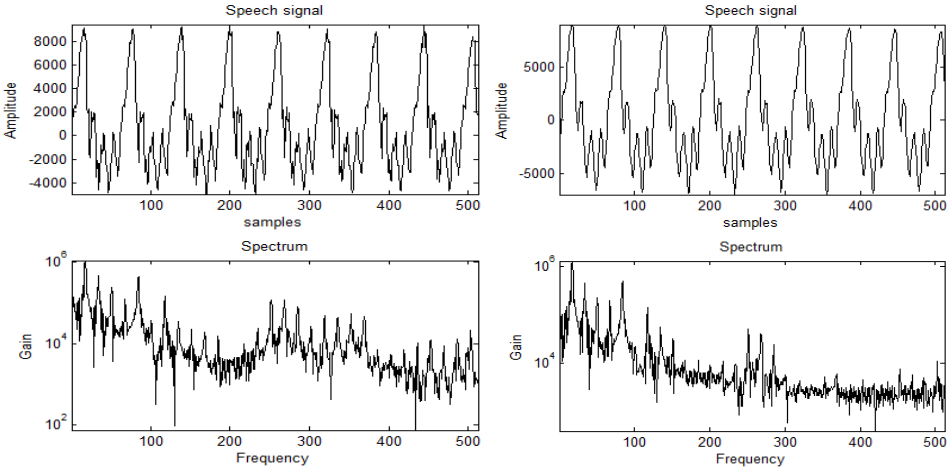


Figure 4. Results of Reconstructed Signals, S(N), $S_{L'}$(N) in Coding Part

Major components of the residual signal consist of the 3rd and upper formants. Generally, the magnitude spectrums of the higher formants have wide bandwidth than that of 1st or 2nd formant and these formants are assigned to a little important information to speech intelligibility. To preserve the naturalness of speech for the 3rd and upper formants, the Gaussian signal is added to the reconstructed roughly with Gaussian level from encoder. Since the characteristic of the residual signal between the original and the reconstructed signal is rather a pseudo colored than a white noise, we can roughly approximate the residual signal with Gaussian signal in decoder.

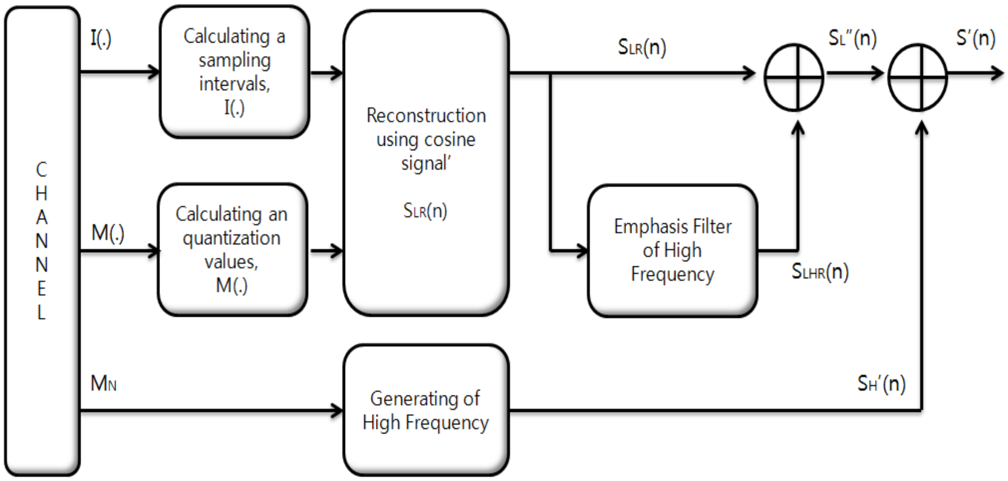

CHANNEL
I(.)
M(.)
MN
Calculating a sampling intervals, I(.)
Calculating an quantization values, M(.)
Reconstruction using cosine signal'
SLR(n)
SLR(n)
Emphasis Filter of High Frequency
SLHR(n)
Generating of High Frequency
SH'(n)
SL''(n)
S'(n)

Figure 5. A New Speech Decoding Scheme

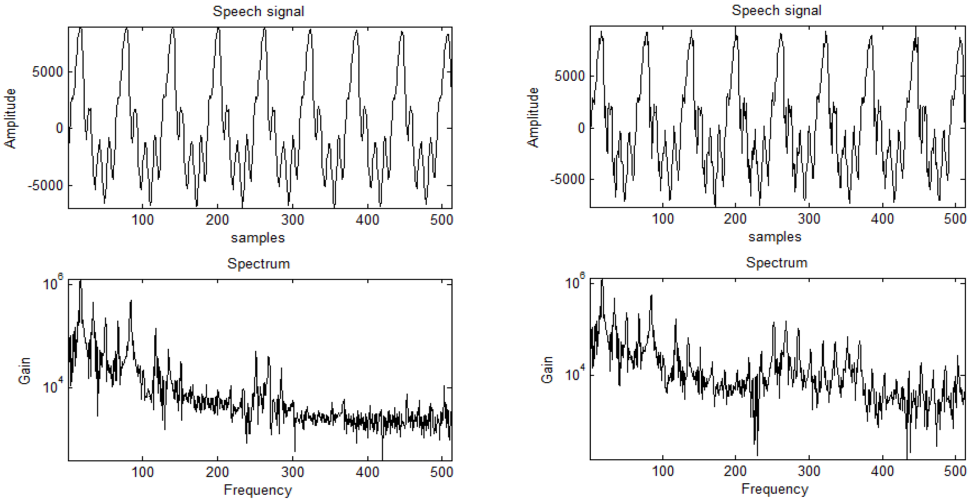


Figure 6. Results of Reconstructed Signals, S_{LR}(n), S'(n) in Decoding Part

$$S_{LR}(k)=[\frac{M(i-1)-M(i)}{2}\cos(\frac{\pi k}{I(i)})+\frac{M(i-1)+M(i)}{2}], \qquad 1\le k\le I(i) \qquad (3\text{-}4)$$

where, *M(.)* is the magnitude of non-uniformly sampled data, *I(.)* is the interval of them and *i* are samples in the interval.

$$S_{LHR}(n)=0.15*|S_{LR}(n)\,| \qquad , \qquad 1\le n\le 256 samples \qquad (3\text{-}5)$$

And, The Harmonics emphasis filter, $S_{LHR}(n)$ is used to compensate components of harmonics in frequency domain.

$$S_L''(n)=S_{LR}(n)+S_{LHR}(n) \qquad (3\text{-}6)$$

$$S_H'(n)=random*M_H \qquad (3\text{-}7)$$

$$S\ '(n)=S_L''(n)+S_H'(n) \qquad (3\text{-}8)$$

Then, the level, M_H, the Gaussian random signal is reconstructed by $S_H'(n)$. And finally we get *S'(n)*, reconstructed signals. This procedure can much reduce the data rate to achieve higher compression ratio than that of the conventional non-uniform sampling even in the noisy environment and also get a good quality of speech by reconstructed harmonics signal. The high frequencies component is considered with Gaussian random level from the M_H in encoder and decoder.

4. Experimental and Results

To compare the performances between the conventional method and the proposed method, three phoneme-balanced Korean sentences were used. Each sentence was pronounced by 8 female and 8 male speakers three times. For simulation test, speech signal was sampled at 8 kHz, filtered by anti-aliasing 10th order LPF with 3.6 kHz cutoff frequency and digitized with 16bit A/D converter.

Table 1. Results of Segmental SNR and Compression Rate

	Existing Method		Proposed Method	
	SNR(dB)	Compression Rate	SNR(dB)	Compression Rate
Sample 1	16.1	2.7	15.4	8.2
Sample 2	16.5	2.6	15.2	7.9
Sample 3	15.4	2.2	14.3	8.3
Sample 4	15.8	2.9	14.6	7.7
Sample 5	16.2	2.5	15.5	8.1
Avg.	16.0	2.5	15.0	7.8

Table 2. Results of 5-point MOS Test

	5-Level MOS	
	Existing Method	Proposed Method
	score	score
Sample 1	4.3	3.8
Sample 2	4.2	3.5
Sample 3	3.9	3.6
Sample 4	4.6	4.2
Sample 5	3.9	3.8

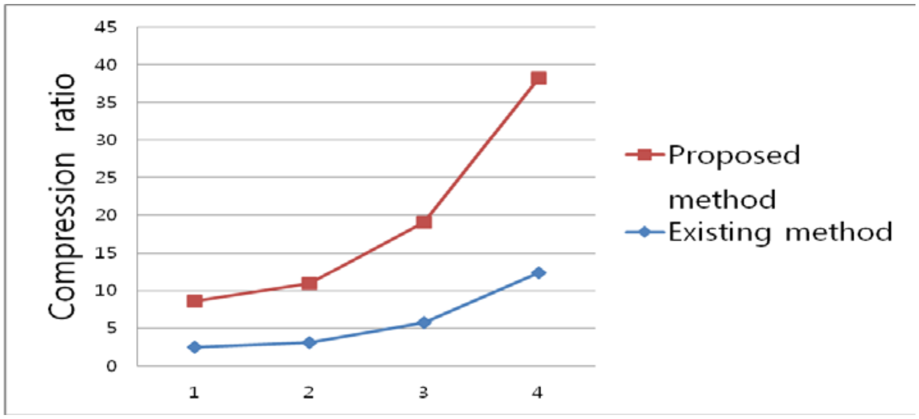


Figure 7. Results of Variables Data to Sampling

Table 4. Results of Variables Data to Sampling

	Existing Method	Proposed Method
8 kHz	2.5	6.1
10 kHz	3.1	7.8
22.01 kHz	5.7	13.4
44.1kHz	12.4	25.8

5. Conclusion

We proposed a new hybrid speech coding technique for high quality coding. It focuses on the naturalness and intelligibility of speech synthesis applications and the compression and signal-to-noise ratio of speech transmission applications. In this paper, to overcome this problem, we proposed new non-uniform sampling method using separated narrow and broad bandwidth by variable bandwidth LPF. The proposed technique is applied to two low-pass filters to speech signal to reduce the data rate without losing the 1st and the 2nd formants information.

In conclusion, Experimental results with phoneme balanced Korean sentences show that the proposed method can achieve higher compression ratio with good of segmental SNR compared with the conventional_method.

References

[1] S. G. Bae, H. W. Park and M. J. Bae, “On a New Enhancement of speech Signal using Nonuniform Sampling and Post Filter”, ICHIT2012, LNCS 7425, **(2012)** August, pp. 723-729.

[2] S. G. Bae, H. W. Park and M. J. Bae, “A Study on Enhancement of Speech using Non-uniform Sampling”, IJHIT, vol. 5, no. 2, **(2012)** April, pp. 237-242.

[3] J. chan-joong and B. myeong-jin, “Analysis regarding vocalizations of teuroteu singers”, The institute of electronics engineers of Korea, Summer scholarship contest of 2009, **(2009)**, pp. 1090-1091.

[4] Y. H. Song, J. H. Ahn and M. J. Bae, “On the noise detection from correlation of near pitch waveforms”, GESTS Society, GESTS Int'l Trans. Computer Science and Engineering, vol. 44, no. 1, **(2008)** January, pp. 45-54.

[5] S. Gazor and W. Zhang, “A soft voice activity detector based on a Laplacian-Gaussian model”, IEEE Trans. Speech Audio Processing, vol. 11, no. 5, **(2003)** September, pp. 498-505.

[6] T. Agarwal and P. Kabal, “Pre-processing of noisy speech for voice coders”, Proc. IEEE Workshop on Speech Coding, Tsukaba, Japan, **(2002)** October, pp. 169-171.

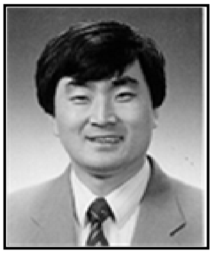
[7] J. Sohn, N. S. Kim and W. Sung, “A statistical model-based voice activity detector”, IEEE Signal Processing Lett., vol. 6, no. 1, **(1999)** January, pp. 1-3.

[8] A. J. Accardi and R. V. Cox, “A modular approach to speech enhancement with an application to speech coding,” J.Acout. Soc. Am, vol. 10, no. 3, **(2001)** September, pp. 1245.

[9] M. J. Bae and S. H. Lee, “Digital Voice Signal Analysis”, Books Publishing Dong Young, ch. 3, 6, **(1998)**.

Authors

Seong-Geon Bae received the M.S. degree in Electronics Engineering from Konkuk University in 1995. He is currently the under Ph.D. degree at Soongsil University. He is currently the Professor of the Dept. of Electronic Communications at Daelim University Collage. E-mail: sgbae123@empal.com



Myung-Jin Bae received the Ph.D. degree in Electronic Engineering from Seoul National University in 1987. He is currently the Professor of the Dept. of Information & Telecommunication at Soongsil University. E-mail: mjbae@ssu.ac.kr