트리코딩과 시영역 하모닉 스케일링을 결합한 음성 부호화

Tree Coding Combined with TDHS for Speech Coding

이 인 성*, 구 본 응**
(In Sung Lee*, Bon Eung Koo**)

※이 논문은 경기대학교 해외파견 연구비에 의하여 연구되었음.

요 약

트리코딩과 시영역 하모닉 스케일링을 결합하여 6.4 및 4.8 kbits/s급 음성부호화기를 제안하였다. 부호화기는 완전 후방 적응적이고 또 하모닉 스케일링 때문에 저지연은 아니다. 부호화기의 에러 성능을 향상시키기 위하여 트리코더에 새로운 적응 피치 예측기, 적응 이득 함수, 단구간 적응 예측 알고리듬 등을 제안하였다. 새로운 코드 트리와 적응 이득 함수, 새로운 후방 적응 피치 예측기, 잡음에 강인한 단구간 적응 예측 알고리듬 등을 이상적인 채널과 잡음의 영향을 받는 채널에 대하여 각각 그 성능을 평가하였다. 두 문장씩 쌍으로 비교한 청취실험 결과, 6.4 kbits/s coder (2-to-1 TDHS/2 bits/sample tree coding)의 음질은 6400 samples/s로 표본화된 6-bit logPCM의 음질과 대등하였다.

ABSTRACT

Tree coding is combined with time-domain harmonic scaling(TDHS) for speech coding at 6.4 and 4.8 kbits/s. The coders are fully backward adaptive but not low delay because of the TDHS. To improve the error performance of the speech coder, adaptive pitch predictor, gain adaptation rules, and short-term adaptation algorithms are proposed. New code trees with appropriate gain adaptation rules, a new backward adaptive pitch predictor, and robust short-term predictor adaptation algorithms are evaluated for both ideal and noisy channels. Paired comparison listening tests show that the 6.4 kbits/s coder (2-to-1 TDHS/2 bits/sample tree coding) has speech quality equivalent to 6-bit logPCM at a sampling rate of 6400 samples/sec.

I. Introduction

Speech coding at 4 to 8 kbits/s has applications in telephony[1,2], digital cellular mobile radio[3], European and North American half-rate mobile radio standards, and more recently, to Internet speech transmission. Tree coding methods have produced some excellent results at 16 kbits/s[4, 5] and have achieved good performance at rates of 9.6[6] and 8 kbits/s[7]. In the present paper, we combine tree coders with time-domain harmonic scaling (TDHS)[8] to obtain transmitted bit rates of 6.4 and 4.8 kbits/s. There have been other efforts at combining TDHS with speech coders at higher bit rates[9-12], but the "look-ahead" search in tree coding seems to provide a critical performance improvement. Objective and subjective performance, ideal and noisy channels, coder delay, and complexity are considered in evaluating the TDHS/tree coder combination.

A block diagram of the system is shown in Fig. 1. The sampling rate of the input speech is 6400 samples/sec, and we limit the time compression of the TDHS operation to a ratio of 2-to-1, so the sampling rate of the tree coder input is 3200 samples/sec. The coder described here are fully backward adaptive in the sense that no side information is transmitted and hence, the pitch extraction for TDHS expansion at the receiver is performed on the tree decoder output. Although backward adaptive, the coder is not low delay because of the TDHS operation.

The tree coder has five basic components, namely, the code generator, the code tree, the distortion measure, the

*충북대학교 전기전자공학부
**경기대학교 전자공학과
접수일자 : 1997년 9월 29일

tree search algorithm, and the path map symbol release rule. In this paper, different code generators including adaptive pitch predictor, gain adaptation rules, and short-term adaptation algorithms are proposed for tree coder in order to improve the error performance of speech coder.

In Section II, we briefly describe the TDHS operation and the tree coder components, highlighting the extensions implemented to achieve the performance goals. Objective and subjective performance results are presented in Section III for both ideal and independent bit error rate channels. The final section summarizes the advantages and disadvantages of this coder and notes possible applications.

Ⅱ. Time Domain Harmonic Scaling and Tree Coding

We briefly outline the choices made concerning the several coder components and the experiments performed in making these selections. Specifically, we mention the choice of TDHS window function, the pitch extraction algorithm, the code generator, the gain adaptation, and the long and short term predictor adaptation algorithms.

TDHS and Window Functions

The time compression in the TDHS step is 2-to-1 and no side information is transmitted, so the two choices needed for TDHS are the window function and the pitch extraction method. Triangular, trapezoidal, cosine, and Hanning window functions were compared by performing a back-to-back 2-to-1 compression followed by a 1-to-2 expansion. Signal to reconstruction error calculations and informal listening tests revealed that the Hanning and cosine windows generally outperform the triangular window and these three seem better than the trapezoidal window. In these comparisons, th AMDF pitch extraction algorithm was employed[13].

Experiments comparing the autocorrelation, the AMDF (on every sample), and the autocorrelation with 3-level center clipping pitch detectors were performed, with the AMDF producing slightly better results than the other two.

Code Tree Design and Gain Adaptation

A block diagram of the code generator for the tree coder is given in Fig. 2. In our work, th excitation sequences are sequences of symbols taken from a tree structure. The tree can be a deterministically populated tree where the branch labels are (say) MMSE Gaussian quantizer outputs or it can be a stochastically populated tree with branch labels taken from a random variate generation routine [14]. The gain adaptation rules must be

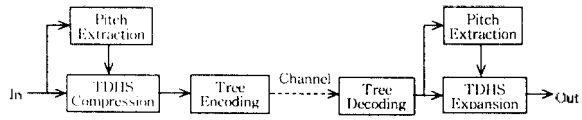

Figure 1. System Configuration for Combining TDHS and Tree Coding

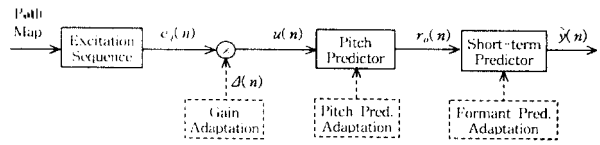

Figure 2. Code Generator in a Tree Coder

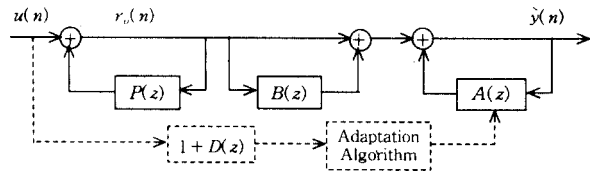

Figure 3. Filtered Residual Driven Method for All-Pole Predictor Adaptation in Pole-Zero Short-Term Predictor

matched with the trees used, and in our experiments, the deterministically populated trees produced better objective performance than the stochastic trees, primarily because the gain adaptation rule for the stochastic tree did not perform well either for ideal or noisy channels. We did not train the code tree on typical sequences but simply used random variates, and that could be an additional source of degradation.

Short-Term Predictor Adaptation

For the short-term predictor adaptation, we used the class of adaptation algorithms developed and reported in [7, 15]. Shaping or smoothing the residual driving terms into these algorithms provides substantial robustness to channel errors while maintaining good ideal channel performance. Experiments indicate that the shaping shown in Fig. 3 gives the best performance, where the all-zero shaping filter is chosen to satisfy

$$\frac{1+B(z)}{1-A(z)} \cong 1+D(z)$$

and the shaping filter coefficients, d_k, are obtained as

$$d_k = \begin{cases} \sum_{j=1}^{N} a_j d_{k-j} + b_k, & 1 \le k \le M \\ \sum_{j=1}^{N} a_j d_{k-j}, & M < k \le P \end{cases}$$

where M is the order of the all-zero predictor. The filtered residual signal is given by

$$\tilde{e}_d(n) = e_q(n) + \sum_{k=1}^{P} d_k e_q(n-p).$$

Long-term Predictor Adaptation

The long-term predictor in our coder has three taps and uses blockwise backward adaptation [4, 5], with the stability correction of Ramachandran and Kabal [16], combined with recursive sample-by-sample updates in-between [17]. We leave details of these algorithms to the references.

The pitch predictor has a long memory and is an often-cited source of problems when there are bit errors. This is particularly true in the backward adaptive algorithms and has led, in some applications, to the long-term predictor being discarded [18]. To address the error sensitivity, we modify the pitch predictor input as shown in Fig. 4, where $S(z)$ is a 3-tap smoother or interpolator $S(z) = s_1 z^{-1} + s_0 z^0 + s_{-1} z^1$. The pitch predictor becomes more robust if the coefficients of the smoother

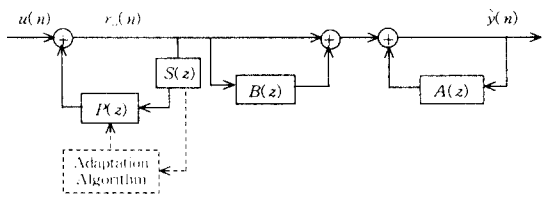

Figure 4. Pitch Predictor with Smoother

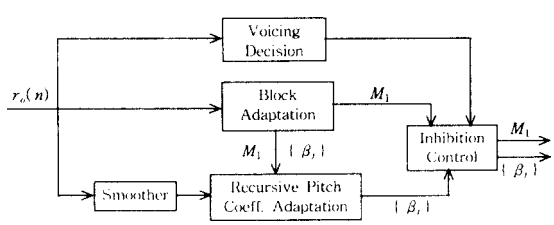

Figure 5. Block Diagram for Pitch Predictor Adaptation with a Smoother

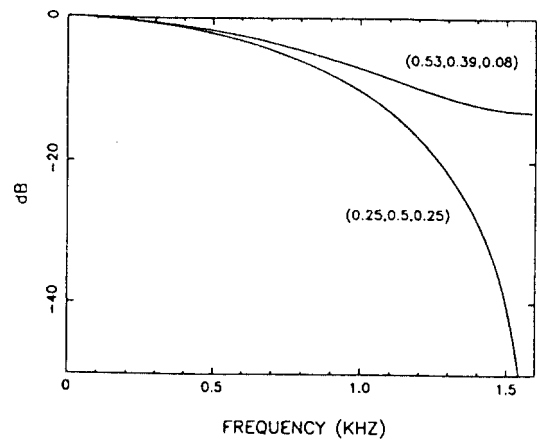

Figure 6. Frequency Response of Smoother

are chosen to implement a low-pass filter. Moreover, in order to track the pitch period change in a block, the coefficients of the smoother can be made variable according to the autocorrelation function of the output of the pitch synthesizer. The coefficients of the smoother are decided by autocorrelation values of the present pitch synthesizer output and three samples that are pitch period lagged. Fig. 5 shows the additional logic associated with the pitch predictor adaptation, including a voicing decision to inhibit th pitch (long-term) predictor in non-periodic segments. The coefficients of the fixed smoother are given by (0.25, 0.5, 0.25), and coefficients used in the variable smoother are decided as either (0.53, 0.39, 0.08) or (0.08, 0.39, 0.53) depending on autocorrelation values of the pitch synthesizer output. The magnitude frequency response of the smoother is shown in Fig. 6. It is clear that the smoother performs low-pass filtering.

The coefficients of the smoother are assigned as $s_1 > s_0 > s_{-1}$ if the following conditions are satisfied: $\hat{\rho}_{M_1+1}(n) > \hat{\rho}_{M_1}(n)$, $\hat{\rho}_{M_1+1} > \hat{\rho}_{M_1-1}(n)$, $\hat{\rho}_{M_1+1}(n) > \hat{\rho}_k$ where the autocorrelation function $\hat{\rho}_k(n)$ is estimated by

$$\hat{\rho}_k(n) = \lambda\, \hat{\rho}_k(n-1) + \frac{r_0(n) r_0(n-k)}{\sigma^2_{r_0}(n)},$$

where $\lambda = 0.95$, and the variance of the pitch synthesizer output, $\sigma^2_{r_0}$, is updated by

$$\hat{\sigma}^2_x(n) = \lambda\, \hat{\sigma}^2_x(n-1) + (1-\lambda) x^2(n).$$

The coefficient s_1 is weighted more than s_0 and s_{-1} because the pitch period of the input is considered to increase by one sample in this case. The coefficients of the smoother are assigned as $s_{-1} > s_0 > s_1$ if the following conditions are satisfied:

$\hat{\rho}_{M_1-1}(n) > \hat{\rho}_{M_1}(n)$, $\hat{\rho}_{M_1-1} > \hat{\rho}_{M_1+1}(n)$, $\hat{\rho}_{M_1-1}(n) > \hat{\rho}_{\min}$.

The coefficient s_{-1} is weighted more than s_0 and s_1 because the pitch period of the input is considered to decrease by one sample. Otherwise, s_{-1}, s_0 and s_1 are the same as that of the fixed smoother.

The smoothed pitch synthesizer output is used as the input in the recursive pitch coefficient adaptation and the calculation of the pitch prediction value. The gradient recursive algorithm is

$$\beta_k(n) = \lambda\beta_k(n-1) + \frac{\mu_s}{\hat{\sigma}_{e_q}(n)\,\hat{\sigma}_{r_s}(n)}\, e_q(n) r_s(n-M_1+k),$$

$k=-1, 0, 1$ $k=-1, 0, 1$

where

$$r_s(n-M_1+k) = s_1 r_0(n-M_1-1) + s_0 r_0(n-M_1+k) + s_{-1} r_0(n-M_1+k+1)$$

and the pitch synthesizer output is

$$r_0(n) = e_q(n) + \sum_{k=-1}^{1} \beta_k r_s(n-M_1+k) = e_q(n) + \sum_{k=-2}^{2} \beta'_k r_s(n-M_1+k)$$

where $\beta'_{-2} = \beta_{-1} s_1$, $\beta'_{-1} = \beta_{-1} s_0 + \beta_0 s_1$, $\beta'_0 = \beta_{-1} s_{-1} + \beta_0 s_0 + \beta_1 s_1$, $\beta'_1 = \beta_0 s_{-1} + \beta_1 s_0$, and $\beta'_2 = \beta_1 s_{-1}$.

The performance comparisons of four pitch adaptation methods including Cuperman's hybrid adaptation[17], Cuperman's hybrid adaptation without the pitch tracker, a hybrid adaptation with a fixed smoother, and a hybrid adaptation with a variable smoother were conducted. The performance comparisons are shown in Table 1. The variable smoother is only nominally better than the fixed smoother, but the importance of including a smoother is clear.

Table 1. Performance comparisons of pitch adaptation methods

Speech	BER	SNR/SNRSEG [dB]					
		Cuperman		Fixed Smoother		Variable Smoother	
Female	0	21.70	19.71	21.03	19.88	21.61	20.21
	10^{-4}	15.70	16.26	19.91	19.13	19.36	19.01
	10^{-3}	8.17	9.18	13.64	14.51	13.35	13.86
	10^{-2}	3.18	3.01	4.11	4.77	4.43	4.70
Male	0	13.20	16.96	13.23	16.83	15.31	17.56
	10^{-4}	8.46	12.29	12.68	16.09	14.63	16.78
	10^{-3}	4.63	7.16	10.39	12.25	11.81	13.20
	10^{-2}	−0.26	1.79	2.56	3.95	2.79	3.86

Ⅲ. TDHS/Tree Coder Performance

To establish the TDHS/tree coder subjective performance, we conducted paired comparison listening tests of the TDHS/tree coder at 6.4 kbits/s versus 6400 samples/s log-PCM at 4, 5, 6, and 7 bits/sample. Two female sentences (sentence 1, sentence 2) and two male sentences (sentence 4, sentence 5) were used in these comparisons, so for each bit rate, eight pairs of speech signals, which consisted of A-B and B-A comparisons for four sentences, were presented through a headphone to a listener. Each listener compared 32 pairs of test signals presented in random order. Twenty persons participated in the subjective listening test. From the test results, the preference percentage of the TDHS/tree coder with respect to logPCM was calculated for each rate. The results of these comparisons are presented in Fig. 7. The 50% preference level is located at about 6 bits in Fig. 7. Narrowband spectrograms of the TDHS/tree coder at 6.4 kbits/s for bit error probabilities of 0, 10^{-3}, and 10^{-2} are shown in Figs. 8, 9 and 10, respectively. The subjective performance at all error rates is surprisingly good.

The 4.8 kbits/s TDHS/tree coder has noticeable, granular-type distortions, and hence, formal subjective listening tests were not undertaken. A spectrogram of the TDHS/tree coder at 4.8 kibts/s with Sentence 1 as input is shown in Fig. 11.

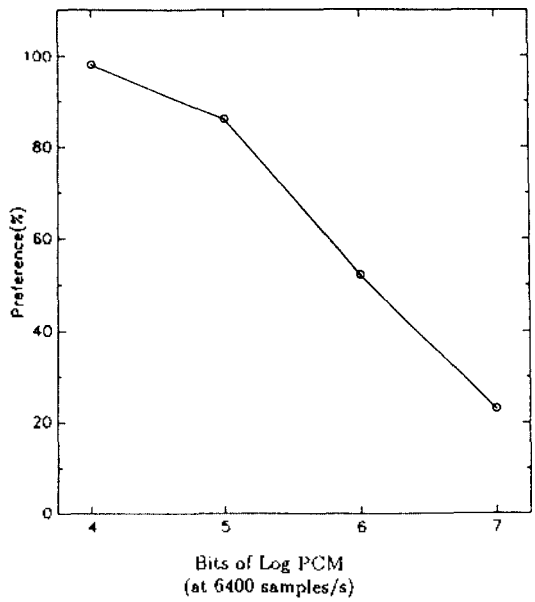

Figure 7. Subjective Evaluation of TDHS-Tree Coder at 6.4 kbits/s

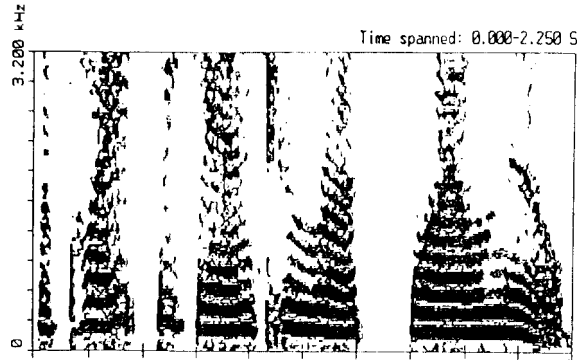

Figure 8. Spectrogram of Reconstructed Speech in Noiseless Channel for 6.4 kbits/s TDHS-Tree Coder : Sentence 1

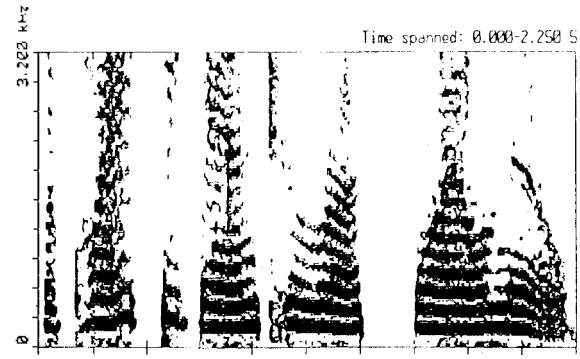

Figure 9. Spectrogram of Reconstructed Speech in Noisy Channel for 6.4 kbits/s TDHS-Tree Coder : BER = 10^{-3}, Sentence 1

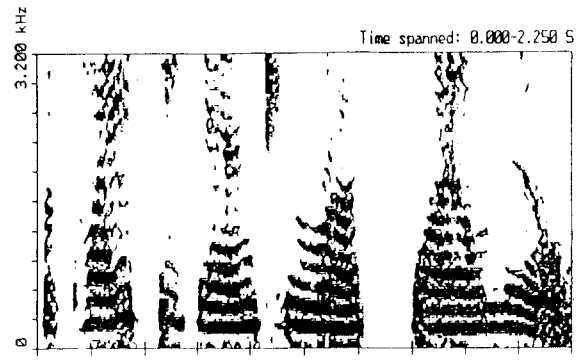

Figure 10. Spectrogram of Reconstructed Speech in Noisy Channel for 6.4 kbits/s TDHS-Tree Coder : BER = 10^{-2}, Sentence 1

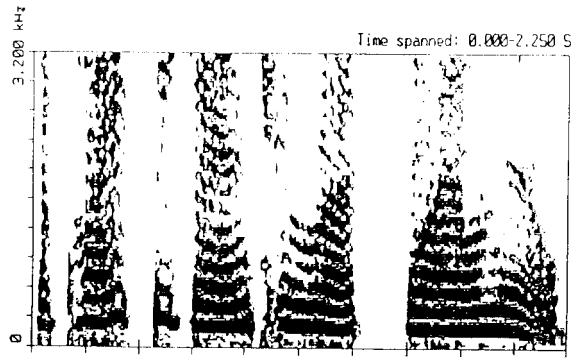

Figure 11. Spectrogram of Reconstructed Speech in Noiseless Channel for 4.8 kbits/s TDHS-Tree Coder : Sentence 1

Ⅳ. Remarks and Conclusions

The combination of time domain harmonic scaling with tree coding opens up new alternatives for medium-to-low rate speech coding. With various sampling rates and fractional rate trees, several alternatives exist for achieving a desired bit rate. For backward adaptive operation, the gain adaptation, short-term predictor adaptation, and long-term predictor adaptation algorithms must be carefully designed to achieve good ideal channel performance while retaining robustness to channel errors.

Our 6.4 kbits/s coder with a 4-4 deterministic code tree, a pole-zero filtered residual driven-adapted short-term predictor, and a long-term predictor adapted on a smoothed input, produces performance near that of 6 bit logPCM at 6400 samples/sec based on the results of paired comparison tests. The coder is also resilient in the face of independent bit errors up to an error probability of 10^{-2}.

The encoder/decoder pair is relatively complex since backward adaptation requires the implementation of the long and short term adaptive prediction algorithms at both the transmitter and receiver. However, using forward adaptation would imply a side information data rate of (at least) 1000 to 1500 bits/s[19, Chap. 12], thus requiring an equivalent reduction in the rate allocated to the tree structured codebook if the total rate is to be kept at 6400 bits/s. Such a forward adaptive coder may be viable at 8 kbits/s for some applications such as Internet voice transmission.

References

1. J-H. Chen and M. S. Rauchwerk, "An 8 kb/s Low-Delay CELP Speech Coder", *Conf. Rec.*, IEEE Global Teleco-

mmun. Conf., Phoenix, AZ, Dec 2-5, 1991, pp. 1894-1898.
2. J-H. Chen and M. S. Rauchwerk, "An 8 kb/s Low-Delay CELP coding of speech", in *Speech and Audio Coding for Wireless and Network Applications* (B. Atal, V. Cuperman, and A. Gersho, eds.), Ch. 4, Boston/Dordrecht/London: Kluwer Academic Publishers, 1993.
3. I. Gerson and M. A. Jasiuk, "Vector Sum Excited Linear Prediction(VSELP) Speech Coding at 8 kbits/s", in Proc. *IEEE Int. Conf. Acoust., Speech and Sig. Proc.*, Apr. 1990, pp. 461-464.
4. V. Iyenger and P. Kabal, "A low delay 16 kb/s speech coder", *IEEE Trans. Signal Processing*, vol. 39, pp. 1049-1057, May 1991.
5. J. D. Gibson and W-W. Chang, "Fractional rate multitree speech coding", *IEEE Trans. Commun.*, vol. 39, pp. 963-974, Jun 1991.
6. Y. C. Cheong, "Analysis and Design for Robust, Low Delay Tree Coding of Speech at 9.6 KBPS," Ph.D. dissertation, Dept. of Elec. Eng., Texas A&M Univ., Aug 1992.
7. H. C. Woo and J. D. Gibson, "Low Delay Tree Coding of Speech at 8 kbits/s", *IEEE Transactions on Speech and Audio Processing*, vol. 2, No. 3, July 1994.
8. D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals", *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. ASSP-27, pp. 121-133, Apr. 1979.
9. D. Malah, R. E. Crochiere, and R. V. Cox, "Performance of transform and subband coding systems combined with harmonic scaling of speech", *IEEE Trans. Acoust., Speech, and Sig. Proc.*, vol. ASSP-29, pp. 273-283, Apr. 1981.
10. D. Malah, "Combined time-domain harmonic compression and CVSD for 7.2 kbits/s transmission of speech signals", in *Proc. IEEE Int. Conf. Acoust. Speech, and Sig. Proc.*, pp. 504-507, Apr. 1980.
11. J. L. Melsa and A. K. Pande, "Mediumband speech encoding using time domain harmonic scaling and adaptive residual coding", in *Proc. IEEE Int. Conf. Acoust. Speech, and Sig. Proc.*, pp. 603-606, May 1981.
12. J. Yuan, C. S. Chen, and H. Zhou, "An ADM speech coding with time domain harmonic scaling", in *Advances in Speech Coding*, B. S. Atal, V. Cuperman, and A. Gersho, Eds. New York: Kluwer, 1990.
13. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
14. J. B. Anderson and J. B. Bodie, "Tree encoding of speech", *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 379-387, Jul 1975.
15. S. H. Nam and J. D. Gibson, "Analysis of the Smoothed Residual-Driven Algorithm for Speech Coders", *IEEE Transactions on Speech and Audio Processing*, vol. 2, No. 3, July 1994.
16. R. P. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders", *IEEE Trans. Aoust. Speech, and Sig. Proc.*, vol. ASSP-35, pp. 937-946, July 1987.
17. R. Pettigrew and V. Cuperman, "Backward pitch prediction for low delay speech coder", in *Conf. Rec.*, Global Telecomm. Conf., pp. 1247-1252, 1989.
18. J-H. Chen, R. V. Cox, Y-C. Lin, N. Jayant, and M. J. Melchner, "A low delay CELP coder for the CCITT 16kb/s speech coding standard", *IEEE J. Sel. Areas Commun.*, vol. 10, pp. 830-849, June 1992.
19. W. B. Kleijn and K. K. Paliwal, Eds. *Speech Coding and Synthesis*, Amsterdam, The Netherlands: Elsevier, 1995.

▲Insung Lee

Insung Lee received B.S. and M.S. degrees in Electronic Engineering from Yonsei University, Korea, in 1983, 1985m respecitvely, an Ph.D degree in Electrical Engineering from Texas A&M University, U.S.A. in 1992.

From 1986 to 1987, he was a research engineer at the Korea Telecomm. Research Conter, Korea. From 1989 to 1992, he was a graduate research assistant int he Dept. of Electrical Eng., Texas A&M Univ. U.S.A From 1993 to 1995, he was with the Signal Processing Section of Mobile Communication Division at Electronics and Telecommunication Resarch Institute(ETRI), Korea, as a senior member of technical staff. Since 1995, He has been with School of Electrical and Electronic Eng. Chungbuk National University, as an assistant professor.

His current research interests include speech coding, channel coding, mobile communications, adaptive filters.

▲구 본 응(Boneung Koo)

1975년 2월 : 서울대학교 공과대학 전자공학 학사
1977년 1월~1982년 7월 : 한국원자력연구소 연구원
1984년 12월 : 텍사스A&M대학 전자공학 석사
1988년 12월 : 텍사스A&M대학 전자공학 박사
1989년 3월~현재 : 경기대학교 전자공학과 부교수
※주관심분야 : 음성부호화, 잡음제거, 음성통신