A Comparative Performance Study of Speech Coders for Three-Way Conferencing in Digital Mobile Communication Networks

이동통신망에서 삼자회의를 위한 음성 부호화기의 성능에 관한 연구

M. S. Lee*, Y. K. Lee*, K. C. Kim*, H. S. Lee*, W. D. Cho**

이 미 숙*, 이 윤 근*, 김 기 철*, 이 황 수*, 조 위 덕**

ABSTRACT

In this paper, we evaluated the performance of vocoders for three-way conferencing using signal summation technique in digital mobile communication network. The signal summation technique yields natural mode of three-way conferencing, in which the mixed voice signal from two speakers are transmitted to a third person, though there has been no useful speech coding technique for the mixed voice signal yet. We established Qualcomm code excited linear prediction (QCELP), vector sum excited linear prediction (VSELP) and regular pulse excited-long term prediction (RPE-LTP) vocoders to provide three-way conferencing using signal summation technique. In addition, as the conventional speech quality measures are not applicable to the vocoders for mixed voice signals, we proposed two kinds of subjective quality measures. These are the sentence discrimination (SD) test and the modified degraded mean opinion score (MDMOS) test. The experimental results show that the output speech quality of the VSELP vocoder is superior to other two.

요 약

본 논문에서는 이동통신망에서 신호 가산방식을 이용한 삼자회의에서의 음성 부호화기 성능을 평가하였다. 두 사람의 섞인 목소리가 다른 회의 참가자에게 전달되는 신호 가산방식은 가장 자연스러운 삼자회의 방식이지만, 아직까지 두 사람의 섞인 목소리를 부호화할 수 있는 유용한 방법은 없다. 본 논문에서는 삼자회의에 신호 가산방식을 적용하기 위해 QCELP, VSELP, RPE-LTP 보코더를 구현하여 평가를 수행하였다. 또한, 두 화자의 목소리가 섞인 음성신호에 대한 부호화기의 성능평가를 위해 기존의 음질 평가법을 그대로 사용할 수 없으므로, 본 논문에서는 두 가지 주관적 평가법을 제안하였다. 제안된 방법은 문장 식별법(SD)과 수정된 DMOS(MDOMS) 방법이다. 실험결과에 의하면 VSELP 보코더의 출력음질이 다른 두 보코더에 비해 좋게 나타났다.

* Department of Information and Communications Engineering, KAIST

** Telecommunication Components Lab., KETI

접수일자 : 1994년 9월 8일

I. INTRODUCTION

In recent years, conventional analog mobile communication systems are rapidly transitioning toward the digital one to increase the channel capacity. The services provided in the mobile communication system are incoming call, vacant call, three-way conferencing and camp on. Among those services, the three-way conferencing has long been of interest to the military. And, nowadays more and more users want the three-way conferencing service. This service in the analog mobile communication system is achieved by using signal summation technique (bridging). Generally, signal summation and signal selection methods are used for the three-way conferencing.

However, there has been no reports on the performance of vocoders for three-way conferencing. Until now, the digital mobile communication has not been achieved in our country, but in the near future it will come into wide use. Thus, the three-way conferencing service will also be achieved in the digital mobile communication system. One narrowband technique currently in use is the signal selection : a speaker has a channel until he finishes or some one with a higher priority bumps him. This technique can be classified into three categories : voice controlled (VC), push-to-talk (PTT), and control signal selection (CSS)[8, 9]. However such techniques are cumbersome since most conference control is handled by interruptions and overlapping speakers, and those schemes present only one speaker to the listener. The signal summation in analog or wideband PCM signal provides natural mode of the three-way conferencing. However, this technique in the digital mobile communication system requires synthesis and reanalysis of the speech waveform in tandem, which makes it hard to represent the voices of multiple speakers without degradation of the speech quality[10].

If only the vocoded speech quality is guaranteed, the signal summation technique is the most useful conferencing technique. In this paper, we investigate whether the speech coders for digital mobile communication networks can model the mixed voice signals from two speakers successfully using the signal summation technique. The vocoding algorithms used for this purpose are the QCELP, the VSELP, and the RPE-LTP vocoders.

As the conventional speech quality measures are designed for a single speaker's voice signal, it is hard to compare the output speech quality of each vocoder for mixed voice signals. In the three-way conferencing, each conferee should listen the mixed voice signal from multiple speakers without any confusion. Thus, to evaluate the three-way conferencing capability, the ability of understanding the meaning of the mixed sentences from other two speakers should be measured to a certain extent.

Two kind of measures to evaluate the performance of vocoders for mixed sentences are proposed. These are the sentence discrimination test and the modified degraded mean opinion score (MDMOS) test. In the sentence discrimination test, the participants for evaluation listen to the mixed sentences spoken by different two speakers and then write down each sentence separately. The MDMOS test differs from the conventional DMOS test in that a pair of mixed sentences are compared as described in Section 3.

In Section 2, the conventional speech coders used in digital mobile communication networks are described briefly. The conventional quality measures for a single speaker's voice signal and the proposed quality measures for the mixed voice signals are described in Section 3. In Section 4, the assessment results are presented and discussed. Finally, conclusions are made in Section 5.

II. SPEECH CODERS FOR DIGITAL MOBILE COMMUNICATION NETWORKS

The digital mobile communication system provides many advantages over the conventional

analog one. For example, it increases channel capacity and improves noise immunity, and also offers the ability to use encryption. Thus much efforts have been made to develop high quality voice coders operating at relatively low bit rates.

Generally, there are three methods to code speech signal : waveform coding, voice coding (vocoding), and hybrid coding[2, 3]. Waveform coding algorithms can produce high quality speech but require relatively high transmission bit rates. Waveform coders include adaptive pulse code modulation (ADPCM), adaptive delta modulation (ADM), etc. Voice coding algorithms are usually based on the modeling of the human vocal tract characteristics using linear prediction technique. They can significantly reduce the bit rate, but their speech qualities are synthetic. They include formant vocoder, linear prediction coding (LPC) vocoder, etc. The hybrid coding algorithms are based on the time-varying excitation model by using both the waveform coding techniques in computing the excitation signals and vocoding techniques in estimating the vocal tract model parameters. The hybrid coders can produce high quality speech with a relatively low transmission bit rate. They include VSELP[4], CELP[8], RPE-LTP[5, 6], residual excited linear prediction (RELP), and improved multi-band excitation (IMBE) vocoders.

In this paper, we evaluate the output speech quality of the VSELP, the RPE-LTP and the QCELP vocoders for three-way conferencing. These vocoders are adopted as official standards for the digital mobile communication systems in many countries. The variable rate QCELP vocoder is selected as a standard for the code division multiple access (CDMA) digital mobile communication. Fig. 1 shows an encoder block diagram of the QCELP vocoder. It has a random codebook to quantize a residual signal using an analysis-by-synthesis method, and an adaptive codebook for the long term prediction. The vocoder operates with variable data rates (8, 4, 2, 1 kbps) based on the speech activity. The LPC coefficients, the pitch, and the codebook parameters are updated according to a selected data rate. The transmitting parameters are the line spectrum pair (LSP) frequency, the long term prediction lag and gain, and the stochastic codebook index and gain.

The 8 kbps VSELP vocoder is used as a standard for the time division multiple access (TDMA) digital communication in the United States. The VSELP encoder shown in Fig. 2 has an adaptive codebook for the long term prediction, and two deterministic codebooks for the excitation, and a gain codebook containing three jointly optimized gains for the two deterministic codebooks and the adaptive codebook. The closed-loop search method is used for an optimum codevector determination. As shown in Fig. 1, analog to digital (A/D) converted signal is filtered by a fourth order Chebyshev type II highpass filter. The LPC analysis using the fixed point covariance lattice algorithm (FLAT) is used to obtain the reflection coefficients. Before the LPC analysis, the spectral smoothing technique (SST) using a binomial window is applied to the covariance matrix. The transmitting parameters are the reflection coefficients, the frame energy, the long term prediction lag, two codewords for the excitation signal, and a gain codeword.

The 13 kbps RPE-LTP coder is selected as a European digital communication standard. The encoder block diagram of the RPE-LTP is shown in Fig. 3. The input speech signal is preprocessed to produce an offset-free signal, and preemphasized with a first order filter. After extracting the reflection coefficients of the short term analysis filter, they are transformed into the log area ratios for transmission. The long term prediction (LTP) lag and the gain are estimated and updated in the LTP analysis block. As a result of the regular pulse excitation (RPE) analysis, the long term residual samples are represented by one of four subsequences. Each subsequence contains 13 equi-distant samples of non-zero amplitude

while the remaining samples are equal to zero. The selected subsequence is identified by the RPE grid position. The transmitting parameters are the log area ratios, the LTP lag and gain, the RPE grid position, the RPE pulse, and the block amplitude.

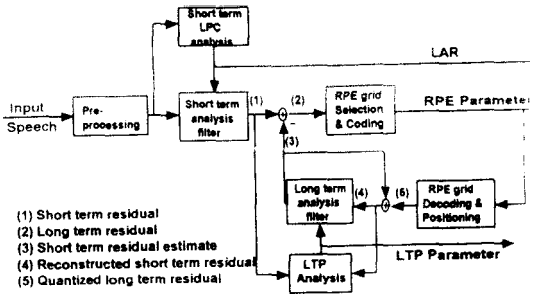

Fig. 3 RPE-LTP encoder block diagram.

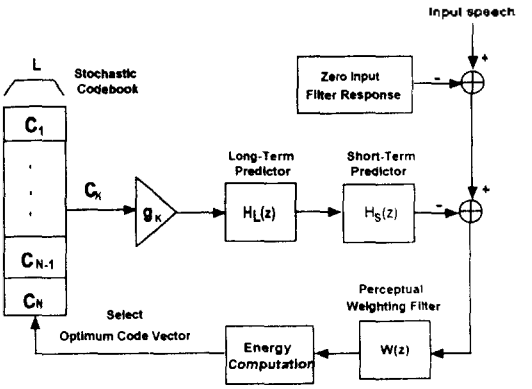

Fig. 1. QCELP encoder block diagram.

Ⅲ. SPEECH QUALITY ASSESSMENTS

One of the most difficult problems in speech coding is the assessment of the relative performance of different coding systems. In order to evaluate the output speech quality of the coding algorithms fairly, a variety of quality assessment techniques have been suggested. The quality assessment techniques generally fall into two classes : subjective measures and objective measures. In the subjective

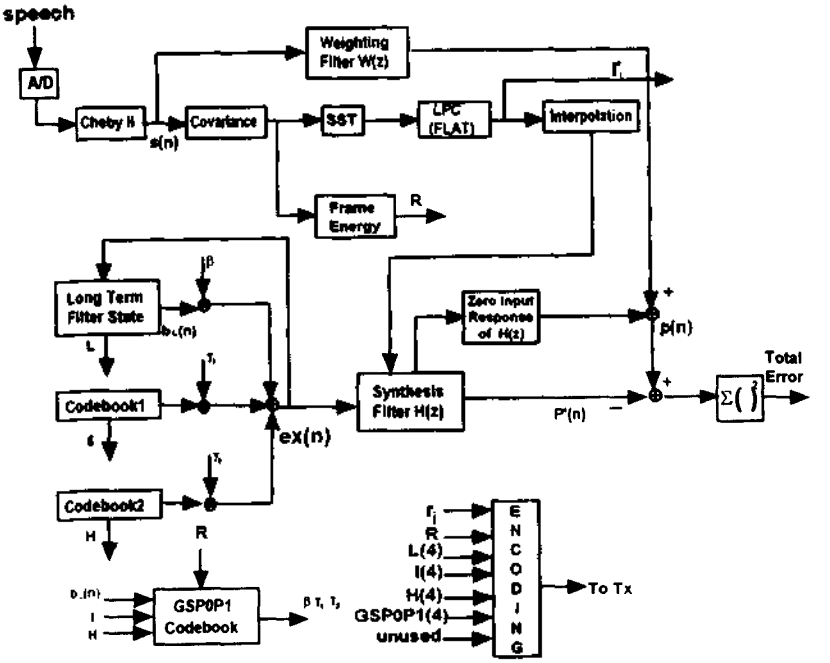

Fig. 2 VSELP encoder block diagram.

measures, a group of participants listen to the vocoded speech signals, and rank the quality of speech subjectively on the basis of a predetermined scale. These subjective measures are appropriate for the vocoders[3]. On the other hand, the objective quality measures are based on the mathematical comparison of the original speech signals with the coded ones. These measures are appropriate for the waveform coders[3].

In three-way conferencing system, the speech quality is an essential issue, because the speech quality of a coding method affects on the subjective judgement of the system's acceptability than the conference protocol or the control techniques. The conventional speech quality measures are not applicable to the vocoders for the mixed voice signals in three-way conferencing.

In the signal selection technique, a questionaire was designed to yield responses on four essentially different dimensions (for instance, each participant check the item corresponding to his opinion for the system on the questionaire) : (1) perceived difficulties of problems, (2) nature and quality of voices heard, (3) amount of effort required to use a system, and (4) perceived relative "goodness" of the system. However, only one conferee is in control at any time in the signal-selection technique.

Until now, there has been no speech quality measures to test the quality of vocoders using signal summation technique in three-way conferencing. In this paper, we propose two subjective speech quality measures. We have not established the overall three-way conferencing system, but evaluated vocoders using the signal summation technique for three-way conferencing.

The first measure proposed is the sentence discrimination test. To enable natural conferencing over voice communication systems, each conferee must comprehend the other two speaker's voices without any confusion. The sentence discrimination test measures the voice discrimination capability of a specific vocoding algorithm. The test procedure is shown in Fig. 4. The voice signals from two speakers are encoded and decoded respectively, and these two voice signals are added to be overlapped in time. Then this mixed sentence is vocoded. For the test, participants listen to the mixed voice signals and then write down two different sentences separately.

The test results are obtained by scoring each dictated sentence with the criteria given in table 1. That is, if a sentence is entirely correct (a sentence is comprised of three phrases), then we give one point. If two of the three phrases are correct, then we give 0.5 points. Otherwise, we give zero point. We assume that the meaning of a sentence could be conveyed correctly when at least two thirds of the sentence is heard to the listeners.

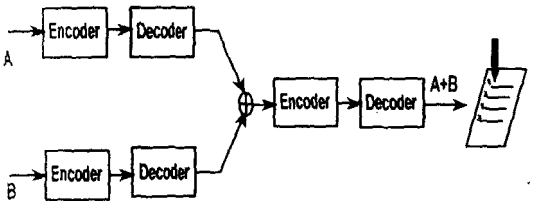

Fig. 4 Test procedure of the sentence discrimination.

Table 1. Scoring criteria for the sentence discrimination test.

Score/Sentence	Case
1	correct sentence
0.5	2 phrase correct, 1 phrase error
0	otherwise

The other proposed measure is the MDMOS test. The conventional vocoders are based on a single speaker's speech production model. To measure the performance of the vocoders for the mixed voice signal, the conventional degraded mean opinion score (DMOS) test is modified as follows. While the conventional DMOS test compares the original speech sentences with the

vocoded ones by rating the degradation according to the criteria in table 2. the MDMOS test uses a pair of speech sentences vocoded differently. The first one of each pair is obtained by vocoding each speech signal separately before summing them in the PCM domain. The second one is obtained by mixing two speech signals in the PCM domain before vocoding them. The difference in each pair is the order of coding and mixing of the speech signals. The remaining procedure of the MDMOS test is identical to that of the DMOS test. In the test, the higher the MDMOS scores, the better the vocoder models the mixed voice signals. If the quality of the second one of each pair which is vocoded after mixing the speech signals in the PCM domain, is not inferior to the first one we can say that the vocoder can model the mixed voice signals as well as a single speaker's one.

Table 2. Score table used in the MDMOS test.

Score	Description
5	Degradation is inaudible
4	Degradation is audible but not annoying
3	Degradation is slightly anoying
2	Degradation is annoying
1	Degradation is very annoying

Ⅳ. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the results of the mean opinion score (MOS) test for a single speaker's voice signal, the sentence discrimination test, and the MDMOS test for mixed voice signals are presented. The sentences spoken by two male and two female speakers are recorded in the laboratory with 8 kHz sampling rate and 16 bit linear quantization level. Ten persons between twenty and thirty years of age take part in the test as listeners. The listeners are not experts in speech quality test, and have not been trained for this test through feedback trials.

4.1 MOS Test for a Single Speaker

For the MOS test, we use a database consisted of two sentences with the duration of 6 and 10 seconds, respectively. The vocoded speech samples are obtained by using the three vocoding algorithms. The results of the test are shown in Fig. 5.

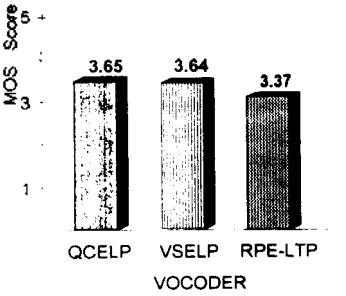

Fig. 5 Results of the MOS test for three vocoders.

In the results, we should consider the transmission rate of each vocoder. The average transmission rate of the QCELP is 5 kbps for the first sentence and 6.7 kbps for the second sentence. The VSELP and the RPE-LTP have the fixed transmission rates of 8 kbps and 13 kbps, respectively. Fig. 5 shows that the QCELP and the VSELP vocoders have comparable speech quality, while the performance of the RPE-LTP is degraded significantly.

4.2 Test Results for Mixed Voice Signals

Ten sentences composed of three phrases are used as a database in the sentence discrimination test and the MDMOS test. Two sentences are mixed as if two speakers talk simultaneously. During the summation, we scale the amplitude of the mixed voice signals to avoid an overflow. For the fair comparison, the transmission rate of the QCELP vocoder is fixed to 8 kbps.

4.2.1 Sentence Discrimination Test

The results of the sentence discrimination test are shown in Fig. 6. Each result is an averaged percentage of the scores obtained from table 1.

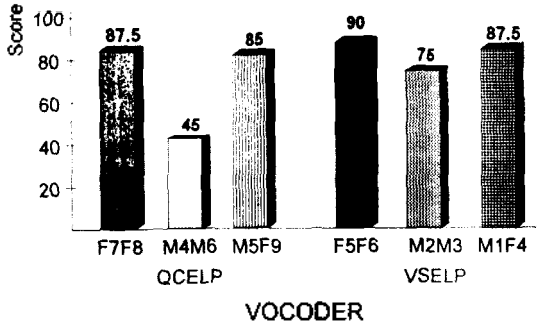

Fig. 6 The results of the sentence discrimination test.

In Fig. 6, we omitted the results of the RPE-LTP vocoder since its score was very low. In Fig. 6, F and M represent female and male speakers, respectively. For instance, M5F9 represents the mixed voice of the 5th sentence spoken by a male speaker and the 9th sentence spoken by a female speaker. During the test, as the listeners hear the same pair of sentence repeatedly, they may become familiar with the voice or the sentences. Thus MOS score for a pair of the sentences heard later may be higher than the one heard before. To avoid this problem, different pairs of the sentences for the same speakers are used in each test. As shown in Fig. 6, the QCELP and the VSELP vocoders discriminate between the sentences well. The low scores for the data spoken by two male speakers (M4M6, M2M3) may be caused by their similar voice tones. In the practical three-way conferencing environment, since each conferee's talk does not fall exactly on other's talk, the score is expected higher than these results.

4.2.2 MDMOS Test

The pairs of sentences for the MDMOS tests are obtained as follows. The first one of each pair is obtained by encoding and decoding two sentences separately before mixing them as shown in Fig. 7(a), while the other is obtained by mixing the two sentences before vocoding them as shown in Fig. 7(b).

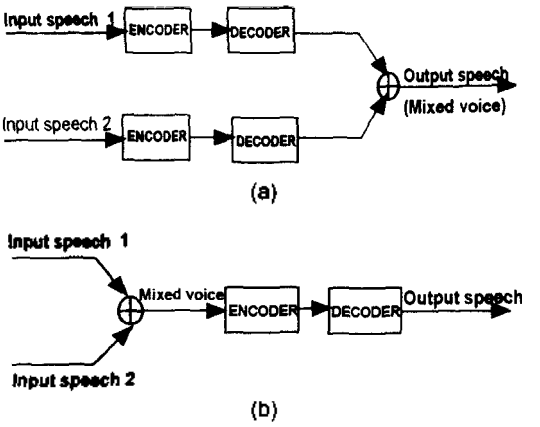

Fig. 7 Sentence generation for the MDMOS test : (a) sentence generated by mixing the vocoded speech signals, (b) sentence generated by vocoding the mixed speech signals.

Fig. 8 shows the results of the MDMOS tests. The participants listen to the two sentences sequentially, then score the amount of degradation of the sentence obtained by vocoding the mixed

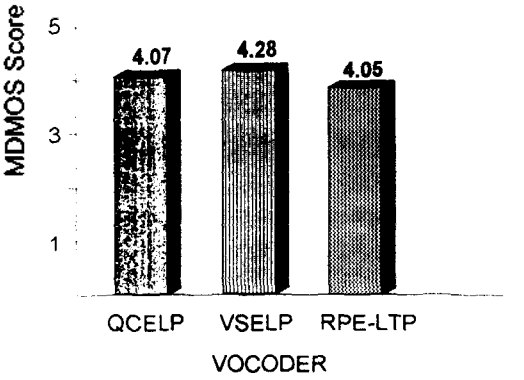

Fig. 8 The results of the MDMOS tests for mixed voice signals.

sentence compared with the one obtained by mixing two vocoded signals according to the criteria in table 2. In Fig. 8 the VSELP shows the best score. That is, the VSELP vocoder represents the mixed voice signals well compared with the other vocoders. The results of the MDMOS test are consistent with those of the sentence discrimination test.

V. CONCLUSIONS

In recent years, the conventional analog mobile communication system is rapidly transitioning toward the digital one. The services provided in the mobile communication system are incoming call, vacant call, three-way conferencing and camp on. Among those services, more and more users want the three-way conferencing service. Generally, signal summation and signal selection method are used for the three-way conferencing.

The signal selection technique is cumbersome since most conference control is handled by interruption and overlapping speakers, and it allows only one speaker's talk to the listeners. The signal summation technique yields more natural mode of three-way conferencing, though it causes degradation in speech quality because of tandemming and the insufficiency in representing the voices of multiple speakers.

In this paper, we evaluate the ability of the vocoders representing the voice signals of multiple speakers, and investigate the possibility of three-way conferencing using signal summation technique in the digital mobile communication system. The vocoding algorithms evaluated are the QCELP, the VSELP, and the RPE-LTP. Also, two kind of subjective quality measures for the mixed voice signals are proposed : the sentence discrimination test and the modified DMOS (MDMOS) test.

From the performance tests using two proposed measures, we obtained consistent results showing that the performance of the VSELP is superior to the other vocoders to represent the voice signals of multiple speakers, and the performance of the QCELP is slightly inferior to the VSELP, while the performance of the RPE-LTP is degraded significantly.

During the tests, we mixed the two sentences to be overlapped in time. However, in actual environments, it is unusual that the two speakers talk simultaneously for a few seconds of period. Thus the score may be higher in actual three-way conferencing environments. As a result, we conclude that the VSELP and the QCELP vocoders can be used in three-way conferencing by using signal summation technique, although they cause some speech quality degradation.

REFERENCES

1. J. D. Tardelli, P. D. Gatewood, E. W. Kreamer, and P. A. La Follette, "The benefits of multi-speaker conferencing and the design of conference bridge control algorithm," Proc. ICASSP93, pp. 435-438, 1993.
2. C. Wheddon and R. Linggard, Speech and Language Processing, Chapman and Hall, pp. 29-62, 1990.
3. P. E. Papamichalis, Practical Approachs to Speech Coding, Prentice-Hall Inc., Englewood Cilffs, New Jersey, pp. 17-149, 1987.
4. Motorola Inc. , Digital Signal Processor Operations Austin, Texas, "Principles of vector-sum excited linear predictive(VSELP) speech coder and Its implementation on the DSP56156".
5. P. Kroon, E. F. Deprettere, and R. J. Sluytser, "Regular-pulse excitation-a novel approach to effective and efficient multipulse coding of speech," IEEE Trans. on ASSP, Vol. 34, No. 5, pp. 1054-1063, Oct. 1986.
6. E. F. Deprettere and P. Kroon, "Regular excitation reduction for effective and efficient LP-coding of speech," Proc. ICASSP85, pp. 25. 8. 1-25. 8. 4, 1985.
7. Telecommunication Industry Association, TIA/EIA Interim standard, "Speech service option standard for wideband spread spectrum digital cellular system," 1994.
8. M. R. Schroeder and B. S. Atal, "Code excited lin-

ear prediction (CELP) : high-quality speech at very low bit rates," Proc. ICASSP85, pp. 25. 1. 1 25. 1. 4. 1985.
9. T. Champion, "Multi-speaker conferencing over narrowband channels," MILCOM 91 Conference Record, IEEE Military Communications Conference, pp. 51. 6. 1-51. 6. 4., 1991.
10. J. W. Forgie, C. E. Feehrer, and P. L. Weene, "Voice conferencing technology program : final report," Lincoln Laboratory, MIT, Lexington, MA, Mar. 1979.

▲We-Duck Cho

1977년 3월~1981년 2월 : Sogang University EE(BS)
1981년 3월~1983년 2월 : KAIST EE(MS)
1983년 3월~1987년 2월 : KAIST EE(Ph.D)
1983년 3월~1990년 3월 : Goldstar Electric Co.
1990년 4월~1991년 10월 : KAITECH HDTV-Group
1991년 1월~현재 : KETI Communication Team Head
R&D field : Digital Communication System
Digital signal Processing Application
Digital Mobile Communication Algorithm Design

▲Hwang-Soo Lee
Professor,
Department of Information and Communications Engineering, KAIST.
See Vol.6 No.3.

▲Ki Chul Kim : See Vol.13 No.1E

▲Yun-Keun Lee

Yun-Keun Lee was born in Korea Febrary 18, 1964. He received the B.S. degree in EE from Seoul national university in 1986. He got the M.S. degree in EE from KAIST in 1988. He is a senior engineer in Goldstar, and courrently enrolled in Ph.D. degree of KAIST. His research area includes the speech signal processing.

▲Mi-Suk Lee

Mi-Suk Lee was born in Korean on March 15, 1968. She received the B.S. degree(1991) and the M.S. degree(1993) in electronics engineering from Ho Seo university. She is currently enrolled in a Ph.D degree of KAIST. Her major research area include the topics related to speech signal processing.