VQ와 GMM을 이용한 문맥독립 화자인식기의 성능 비교*

Performance comparison of Text-Independent Speaker Recognizer Using VQ and GMM

김 성 종 · 정 훈** · 정 익 주***
Seong-Jong Kim · Ik-Joo Chung · Hoon Chung

ABSTRACT

This paper was focused on realizing the text-independent speaker recognizer using the VQ and GMM algorithm and studying the characteristics of the speaker recognizers that adopt these two algorithms. Because it was difficult ascertain the effect two algorithms have on the speaker recognizer theoretically, we performed the recognition experiments using various parameters and, as the result of the experiments, we could show that GMM algorithm had better recognition performance than VQ algorithm as following. The GMM showed better performance with small training data, and it also showed just a little difference of recognition rate as the kind of feature vectors and the length of input data vary. The GMM showed good recognition performance than the VQ on the whole.

Keywords : VQ, GMM, LPC, MFCC

I. 서 론

본 논문에서는 두 가지의 알고리즘 즉, VQ(Vector Quantization)와 GMM(Gaussian Mixture Model)이 화자인식기에 적용되었을 때, 이들 두 화자인식기의 특징을 다양한 실험을 통하여 알아내는데 주안점을 두었다. 이것을 알아내기 위해 다양한 파라미터를 이용하여 인식실험을 하였다. 이 실험을 위해 우선 두 알고리즘을 이용한 화자인식기를 구현하고 이 두 인식기에 파라미터를 적용하여 실험을 하였다. 여기에 사용된 파라미터는 음성특징벡터인 선형예측계수(LPC)와 멜 주파수 켑스트럼 계수(MFCC), 훈련데이터의 길이(약 40, 약 80초), 코드북의 크기(VQ화자인식기의 경우), mixture 크기(GMM 화자인식기

* 이 논문은 정보통신부의 정보통신 우수시범학교 지원사업에 의해 수행된 것입니다.
** 강원대학교 대학원 전자공학과
*** 강원대학교 공과대학 전기전자정보통신공학부

의 경우), 그리고 인식테스트에 사용되는 입력음성의 길이(3음절, 5음절) 등이다. 이 파라미터들을 이용한 실험결과를 토대로 하여 화자인식기에 사용된 두 알고리즘(VQ, GMM)이 어떤 특징을 가지고 있는지를 알아내었다.

본 논문에서 구현한 화자인식기는 화자의 입력음성을 이용하여 주어진 모집단내에서 그 화자가 누구인가를 찾아내는 화자식별이며, 또한 입력되는 음성신호가 미리 정해진 문장 또는 단어 등으로 제한받지 않는 문맥 독립이다. 본 논문에서는 이것을 "문맥독립 화자인식기"라 부르며, 줄여서 "화자인식기"로 칭하였다.

본 논문은 2장에서, 본 화자인식기의 구현에 대하여 서술하였고, 3장에서는 2장에서 구현된 화자인식기를 이용한 인식실험 및 결과를 서술하였으며, 4장 결론에서는 실험결과를 바탕으로 하여 두 화자인식기의 특징과 성능을 비교 분석하였다.

II. 화자인식기 구현

본 화자인식기는 벡터 양자화를 이용한 화자인식기(VQ 화자인식기)[1,4]와 GMM을 이용한 화자인식기(GMM 화자인식기)[3,4]로 나누어지며 다음의 <그림 2.1>과 같은 구조를 갖는다.

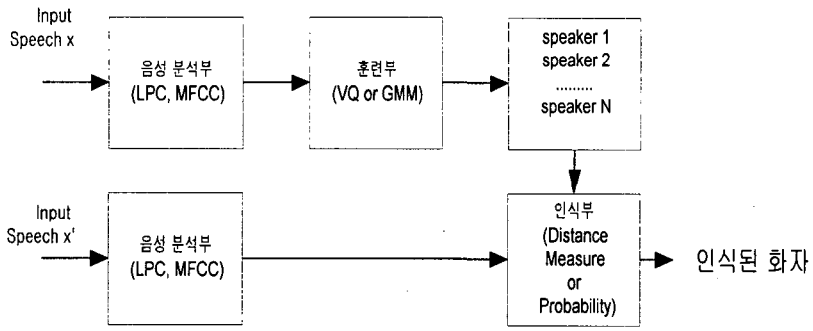

그림 1. 화자인식기의 구조

2.1 음성분석부

VQ와 GMM 화자인식기에 공통적인 부분으로서 본 논문에서는 가장 많이 사용되고 있는 선형예측계수[2][5][6]와 멜 주파수 켑스트럼 계수를 이용하였다.

2.2 VQ 화자인식기

훈련부와 인식부로 구성되어 있으며, 훈련부는 음성분석부에서 구한 특징벡터(LPC, MFCC)를 이용하여 각 화자에 대한 코드북을 만드는 벡터 양자화를 하게 되고 인식부에서는 훈련부에서 만든 코드북을 이용하여 화자를 인식하게 된다.

2.2.1 훈련부

그림 <2.2>에서 나타낸 것처럼 K-means 와 이진분리 알고리즘[2]을 이용하여 벡터양자화를 행하였고 각 화자에 대한 코드북을 만들게 된다.

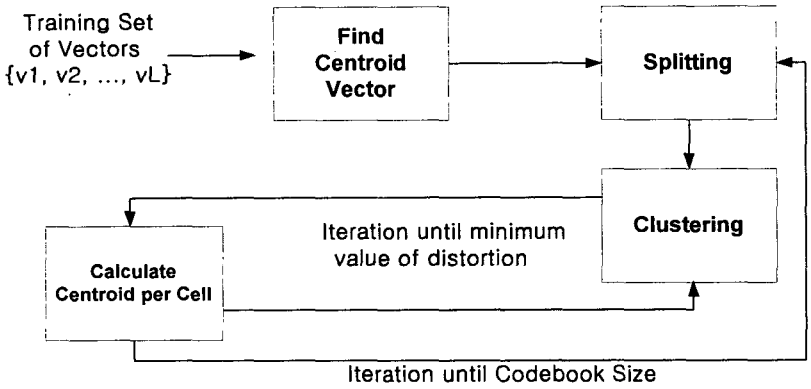

그림 2. VQ로 코드북을 만드는 블록도

2.2.2 인식부

인식부분은 입력된 음성데이터를 훈련과정에서 만들어둔 각 화자에 대한 코드북(Codebook 1, Codebook 2, ..., Codebook N)과 비교하여 스펙트럼 거리(spectral distance)가 가장 가까운 것을 선택하여 그 코드북에 해당하는 화자를 인식으로 결정한다.

2.3 GMM 화자인식기

VQ 화자인식기와 마찬가지로 훈련부와 인식부로 구성되어 있으며, 훈련부에서는 각 화자의 GMM을 만들고 인식부에서는 이 모델을 이용하여 화자를 인식한다.

2.3.1 GMM 훈련부

여기에서는 화자의 음성 특징벡터들의 분포를 가장 잘 나타낼 수 있는 GMM 파라미터 즉, λ를 추정한다. 여기서 λ는 다음의 (1)식처럼 mixture weight p_i, mean vector $\vec{\mu_i}$, covariance matrix $\sum i$ 으로 구성된다.

$$\lambda=\{p_i\ ,\ \vec{\mu_i}\ ,\ \sum i\}\quad i=1,\ldots,M \tag{1}$$

이 λ를 추정하는 방법이 몇 가지가 있지만 본 논문에서는 EM(Expectation Maximization) algorithm을 사용하였는데, 기본개념은 $P(X|\lambda')\geq P(X|\lambda)$가 되는 새로운 모델 λ'를 정해진 threshold에 도달 할 때까지 아래 식 (2)(3)(4)를 이용하여 반복과정을 통하여 찾는 것이다. 이 EM algorithm을 적용할 때, 초기 모델이 필요하게 되는데, 본 논문에서는 적용할 모델차수(mixture size)만큼의 셀 블록을 VQ를 이용하여 구하고 각 셀 당

mean과 variance를 구하여 이것을 초기 모델로 정하였다.

Mixture Weights :

$$p_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x_t}, \lambda) \tag{2}$$

Means :

$$\vec{\mu_i} = \frac{\sum_{t=1}^{T} p(i \mid \vec{x_t}, \lambda)\vec{x_t}}{\sum_{t=1}^{T} p(i \mid \vec{x_t}, \lambda)} \tag{3}$$

Variances :

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x_t}, \lambda)x_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x_t}, \lambda)} - \mu_i^2 \tag{4}$$

여기서, posteriori probability는

$$p(i \mid \vec{x_t}, \lambda) = \frac{p_i b_i(\vec{x_t})}{\sum_{k=1}^{M} p_k b_k(\vec{x_t})} \tag{5}$$

이고, component density는

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\sum i|^{1/2}} \times \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu_i})' \sum i^{-1} (\vec{x} - \vec{\mu_i})\right\} \tag{6}$$

이다.

2.3.2 GMM 인식부

훈련부에서 만든 각 화자에 대한 GMM, 즉 λ_1, λ_2, ..., λ_S을 이용하여 입력된 음성에 대한 posteriori probability를 구하여 이중 가장 큰 확률을 가진 speaker model을 찾는 것이다. 즉,

$$\hat{S} = \arg \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x_t} \mid \lambda_k) \tag{7}$$

III. 실험 및 결과

3.1 실험

실험은 본 화자인식기의 성능을 평가하기 위한 것으로서, 앞에서 언급한 음성의 특징 파라미터인 선형예측계수와 멜 주파수 켑스트럼 계수, 훈련데이터의 양, 코드북의 크기, 알고리즘(VQ & GMM) 등에 따라 인식률을 비교해 봄으로서 본 화자인식기에 영향을 미치는 파라미터와 더불어 두 화자인식기(VQ & GMM) 간의 성능 비교분석을 중점적으로 다루었다.

본 실험에는 원광대학교 PBW 음성데이터 내에 있는 남자화자 37명을 이용하였으며, 음성특징벡터로 사용될 각 화자의 음성데이터는 16 KHz로 샘플링 된 것이므로 한 프레임 길이는 25ms를 적용하기 위하여 400샘플/frame을 취하였다. VQ 화자인식기의 경우 코드북의 크기를 32, 64, 128, 256으로 정하여 각각에 대한 인식률을 실험하였고, GMM 화자인식기의 경우는 Mixture의 크기를 8, 16, 32로 변화시키면서 인식률을 실험하였다. 특징벡터로 사용되는 선형예측계수는 10차와 14차를, 멜 주파수 켑스트럼 계수는 14차를 적용하였고, 각 화자의 훈련 데이터의 양은 약 40초 분량과 약 80초 분량으로 정하였다.
본 실험의 결과에 사용된 인식률의 단위는 백분율(%)로 나타내며 식은 다음과 같다.

$$\text{인식률}(\%) = \frac{\text{정(正)인식 횟수}}{\text{입력 음성 데이터의 수}} \times 100 \tag{8}$$

각각의 화자에 대한 인식률은 위의 식을 이용하여 구하고, 전체인식률은 37명의 화자에 대한 인식률의 평균을 취하였다. 즉, 다음 식과 같다.

$$\text{전체 인식률}(\%) = \frac{\text{각 화자의 인식률의 총합}}{\text{화자의 수(37명)}} \times 100 \tag{9}$$

훈련에 사용되는 단어 세트는 될 수 있으면 모든 음가를 포함할 수 있도록 즉, 자음과 모음(이중모음포함)의 모든 음가가 포함되도록 설정된 것이며 훈련에 사용되지 않은 단어 세트들은 인식실험에서의 입력음성으로 사용하였다.

3.2 결과

실험 결과는 VQ와 GMM 화자인식기의 인식률 결과를 비교할 수 있도록 표와 그래프로 작성하였다.

표 1. 훈련데이터: 약 40초, LPC차수: 10차일 경우의 인식률 비교

VQ 화자인식기			GMM 화자인식기		
코드북 크기	입력 음성의 길이		mixture 수	입력 음성의 길이	
	3음절	5~8음절		3음절	5~8음절
32	88.6	89.1	8	88.4	90.5
64	89.3	92.3	16	92.7	93.4
128	90.9	92.9	32	93.1	96.0

표 2. 훈련데이터: 약 40초, LPC차수: 14차일 경우의 인식률 비교

VQ 화자인식기			GMM 화자인식기		
코드북 크기	입력 음성의 길이		mixture 수	입력 음성의 길이	
	3음절	5~8음절		3음절	5~8음절
32	88.0	91.3	8	92.0	95.8
64	90.9	93.2	16	96.8	97.6
128	92.9	93.8	32	98.1	98.3

표 3. 훈련데이터: 약 40초, MFCC의 차수: 14차일 경우의 인식률 비교

VQ 화자인식기			GMM 화자인식기		
코드북 크기	입력 음성의 길이		mixture 수	입력 음성의 길이	
	3음절	5~8음절		3음절	5~8음절
32	89.7	91.0	8	96.2	97.5
64	92.7	94.8	16	98.4	97.8
128	95.0	96.1	32	98.8	99.2

표 4. 훈련데이터: 약 80초, LPC차수: 10차일 경우의 인식률 비교

VQ 화자인식기			GMM 화자인식기		
코드북 크기	입력 음성의 길이		mixture 수	입력 음성의 길이	
	3음절	5~8음절		3음절	5~8음절
64	92.6	94.8	8	88.2	90.5
128	92.3	95.4	16	93.1	94.6
256	94.4	96.3	32	96.0	97.0

표 5. 훈련데이터: 약 80초, LPC차수: 14차일 경우의 인식률 비교

VQ 화자인식기			GMM 화자인식기		
코드북 크기	입력 음성의 길이		mixture 수	입력 음성의 길이	
	3음절	5~8음절		3음절	5~8음절
64	94.1	95.1	8	95.0	94.8
128	95.0	95.7	16	97.3	97.5
256	95.5	96.1	32	98.3	99.0

표 6. 훈련데이터: 약 80초, MFCC의 차수: 14차일 경우의 인식률 비교

VQ 화자인식기			GMM 화자인식기		
코드북 크기	입력 음성의 길이		mixture 수	입력 음성의 길이	
	3음절	5~8음절		3음절	5~8음절
64	95.8	97.1	8	95.7	95.4
128	96.8	97.8	16	98.1	97.5
256	97.6	98.5	32	99.2	98.5

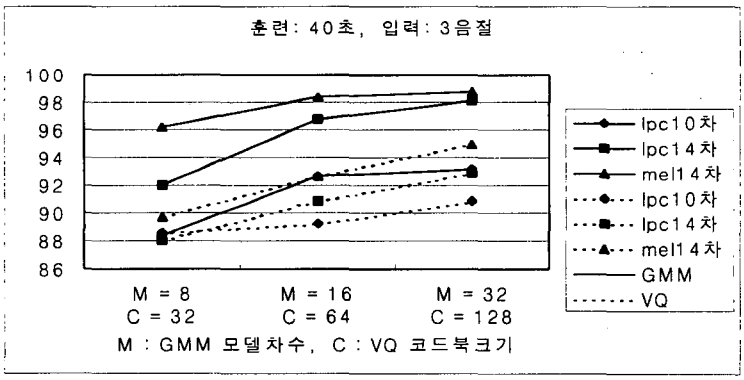

그림 3. 훈련데이터: 약40초, 입력음성: 3음절일 경우의 인식률 비교 그래프

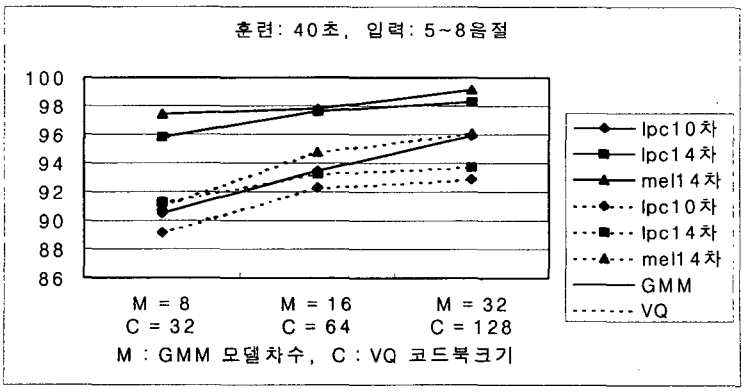

그림 4. 훈련데이터: 약40초, 입력음성: 5~8음절일 경우의 인식률 비교 그래프

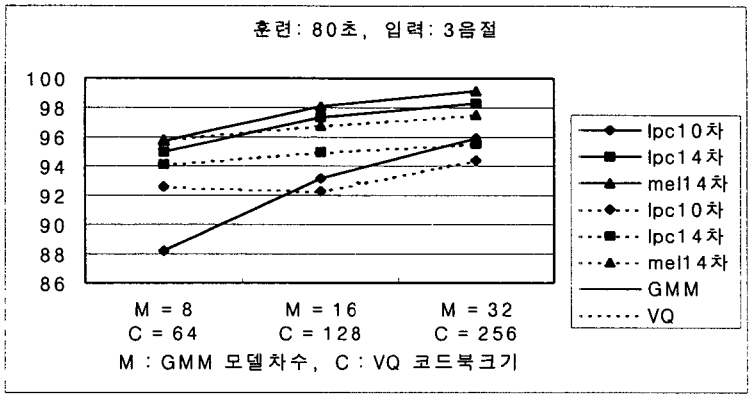

그림 5. 훈련데이터: 약80초, 입력음성: 3음절일 경우의 인식률 비교 그래프

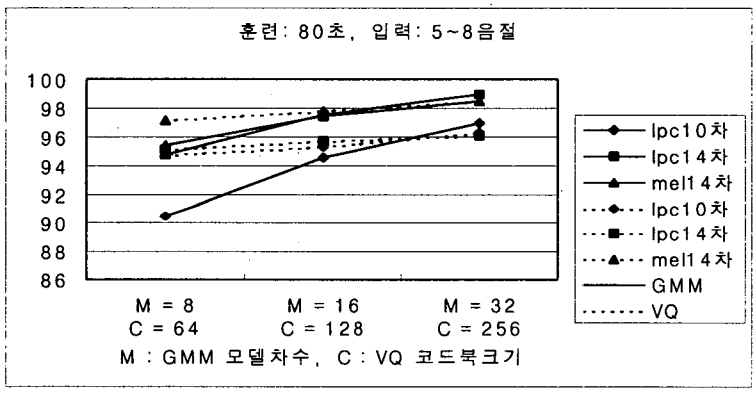

그림 6. 훈련데이터: 약 80초, 입력음성: 5~8음절일 경우의 인식률 비교 그래프

실험 결과를 종합해보면, VQ 화자인식기의 경우 선형예측계수와 멜 주파수 켑스트럼 계수의 경우를 비교해볼 때, 우선 두 경우 모두 다음과 같은 파라미터에 의해 인식률의 변화를 볼 수 있었다. 우선 훈련데이터 량의 증가(40초 분량의 음성데이터로 훈련한 경우와 1분 20초 분량의 음성데이터로 훈련한 경우)에 따라 인식률의 향상을 보였으며, 코드북의 크기(VQ)와 모델차수(GMM)가 커짐에 따라 인식률이 또한 높아졌다. 다음으로는 인식에 사용되는 입력 데이터의 길이가 길어짐에 따라서도 인식률의 향상을 보임을 알 수 있었다. 또한 두 계수의 인식률을 비교했을 때, 멜 주파수 켑스트럼 계수의 경우가 선형예측계수에 비해 좀 더 높은 인식률을 보였다. 한편, GMM 화자인식기의 경우에는 훈련데이터의 량과 모델차수가 적절하게 선택된 경우에는 두 계수의 인식률이 VQ 화자인식기에 비해 그 차이가 현저하게 감소하였다. 즉, GMM 화자인식기의 경우는 알고리즘 내에서 주어진 파라미터(LPC, MFCC)를 이용하여 특정 화자의 음성 특징을 모델링하는 능력이 월등함을 보인다고 할 수 있다. 또한, 표와 그래프에서 나타난 것처럼 VQ 화자인식기에

비하여 적은량의 훈련데이터로도 좋은 인식률을 보인다는 것을 알 수 있다. 그리고 VQ에 비하여 입력음성의 길이에 대하여도 강인함을 알 수 있다.

본 실험 결과에서 중요한 파라미터중의 하나는 VQ 코드북의 크기와 GMM 모델차수인데, VQ 화자인식기의 경우 코드북의 크기는 훈련에 사용되는 음성데이터 량에 따라 제한되며, 더욱이 코드북의 크기는 어느 정도의 크기(256~512)이상이 되면 인식률이 포화(saturation)되기 때문에 코드북 크기와 훈련 데이터 량간의 적절한 절충이 이루어져야 하며 본 VQ 화자인식기의 경우 40초 분량의 훈련 데이터의 경우 코드북 크기가 128이 적절하였으며 80초 분량의 경우는 256이 훌륭한 인식률의 결과를 얻을 수 있었다. GMM 화자인식기의 모델 차수(Mixture Size)의 경우는 차수가 작게되면 화자의 변별력이 떨어지고 너무 크면 파라미터가 증가하게되고 또한 계산량도 증가하게 된다. 본 GMM 화자인식기의 경우는 16, 32정도가 적절한 모델차수로 나타났다.

IV. 결 론

본 실험 결과에서 나타난 두 화자인식기의 특징은 요약하면 다음과 같다. VQ 화자인식기는 훈련데이터의 양이 많을수록 좋은 인식률을 나타내며, 화자의 음성 특징 파라미터의 종류와 차수에 대해서도 민감한 반응을 보였다. 즉 LPC보다는 MFCC가 더 좋은 인식성능을 보였으며, 또한 차수가 높을수록 좋은 성능을 보였다. 또한 입력되는 음성 길이의 증가에 따라 좋은 인식률을 보였다. 이에 반하여, GMM 화자인식기는 훈련데이터의 양, 화자의 음성 특징 파라미터의 종류와 차수, 그리고 입력음성의 길이에 대하여 VQ와 마찬가지로 인식률의 향상을 가져왔지만 향상의 정도가 VQ만큼 크게 나타나지는 않았다. 특히, 특징벡터의 종류(LPC, MFCC)에 따른 인식률과 훈련데이터의 증가에 따른 인식률의 향상은 VQ에 비하여 현저하게 줄어드는 것을 알 수 있었다. 이것은 곧 GMM 화자인식기의 경우에는 사용된 특징벡터의 종류와 훈련데이터의 길이에 대해 어느 정도 강인하다는 것을 말해주는 것이며, 따라서 보다 적은 양의 훈련데이터를 가지고도 좋은 인식성능을 낼 수 있다는 것이다. 또한 성능 면에 있어서 VQ 화자인식기에 비하여 대체적으로 좋은 인식률을 보임을 알 수 있었다.

본 논문의 실험으로 증명된 VQ 화자인식기에 대한 GMM 화자인식기의 특징은 다음과 같이 요약될 수 있다. 첫째, 특징벡터(LPC, MFCC)의 종류에 대하여 덜 민감하다. 즉, LPC를 적용했을 때와 MFCC를 적용했을 때의 인식성능이 VQ의 경우보다 적게 차이가 난다. 둘째, 입력음성(테스트 음성)의 길이에 있어서 VQ에 비하여 짧은 경우에도 좋은 인식성능을 보인다. 셋째, 약 40초의 훈련데이터로도 VQ 화자인식기에서의 약 80초 훈련데이터를 이용한 인식 성능을 보인다. 다시 말하면, VQ보다 적은 양의 훈련 데이터로도 좋은 성능을 발휘한다고 할 수 있다.

참 고 문 헌

[1] Tomoko Matsui & Sadaoki Furui, 1994. "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's", *IEEE Transactions on Speech and Audio Processing, VOL. 2, NO. 3.*

[2] Lawrence Rabiner & Biing-Hwang Juang, 1993. *Fundamentals of Speech Recognition,* Prentice Hall.

[3] Douglas A. Reynolds & Richard C. Rose, 1995. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing, VOL. 3, NO. 1.*

[4] Konstantin P.MARKOV & Seiichi NAKAGAWA, 1997. "Text-Independent Speaker Identification Utilizing Likelihood Normalization Technique", *IEEE TRANS. INF&SYST. VOL.E80-D. NO.5.*

[5] Shuzo Saito & Kazuo Nakata, 1985. *Fundamentals of Speech Signal Processing,* Academic Press.

[6] A.M.Kondoz, 1994. *Digital Speech Coding for Low Bit Rate Communications Systems.* Wiley.

접수일자: 2000. 5. 06.
게재결정: 2000. 6. 11.

▲ 김 성 종
강원도 춘천시 효자2동
강원대학교 대학원 전자공학과
Tel: (0361)250-7274(302호) (O), 263-4345 (H)
e-mail: deniro@mirae.kangwon.ac.kr

▲ 정 훈
강원도 춘천시 효자2동
강원대학교 대학원 전자공학과
Tel: (0361)250-6322
e-mail: hchung@mirae.kangwon.ac.kr

▲ 정익주
강원도 춘천시 효자2동
강원대학교 공과대학 전기전자정보통신공학부
Tel: (0361)250-6322
e-mail: ijchung@cc.kangwon.ac.kr