한국음향학회지 제24권 제4호 pp. 224~229 (2005)

잡음환경에서의 음성인식을 위한 켑스트럼의 확률분포 정규화 기법

Cepstrum PDF Normalization Method for Speech Recognition in Noise Environment

석 용 호*, 최 승 호**, 이 황 수***
(Yong Ho Suk*, Seung Ho Choi**, Hwang-Soo Lee***)

*(주)엠큐브웍스, **서울산업대학교 전자정보공학과, ***한국과학기술원 전자전산학과
(접수일자: 2005년 3월 7일; 수정일자: 2005년 4월 4일; 채택일자: 2005년 4월 8일)

본 논문에서는 부가잡음 환경에서의 강인한 음성인식을 위해 켑스트럼의 확률밀도 (pdf) 정규화 기법을 제안한다. 기존의 방법들은 켑스트럼의 평균 및 분산 등 주로 1, 2차 통계치 만을 정규화 하지만 제안한 방법은 깨끗한 음성과 잡음이 부가된 음성의 켑스트럼의 pdf를 동일하게 함으로써 켑스트럼의 통계치를 완벽하게 정규화 한다. 목표 pdf로는 다양한 확률분포를 고려하기 위하여 일반 (generalized) 가우시안 분포를 선택하였다. 또한 인식시 계산량을 감축하기 위하여 표참조방법 (table lookup method)를 개발하였다. 화자독립 고립단어 인식 실험에서 제안된 기법이 기존 방법들보다 우수한 성능을 보였으며, 특히 잡음이 심한 환경에서 성능향상이 두드러졌다.

핵심용어: 음성인식, 켑스트럼, 정규화, 확률밀도함수
투고분야: 음성처리 분야 (2.5)

In this paper, we propose a novel cepstrum normalization method which normalizes the probability density function (pdf) of cepstrum for robust speech recognition in additive noise environments. While the conventional methods normalize the first- and/or second-order statistics such as the mean and/or variance of the cepstrum, the proposed method fully normalizes the statistics of cepstrum by making the pdfs of clean and noisy cepstrum identical to each other. For the target pdf, the generalized Gaussian distribution is selected to consider various densities. In recognition phase, we devise a table lookup method to save computational costs. From the speaker-independent isolated-word recognition experiments, we show that the proposed method gives improved performance compared with that of the conventional methods, especially in heavy noise environments.

Keywords: Speech Recognition, Cepstrum, Normalization, Pdf
ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

최근 들어 음성인식 기술은 실험실 데모 수준을 벗어나 실생활에 적용, 상용화되고 있다. 그러나 현재의 음성인식 시스템은 제한된 환경에서는 비교적 좋은 성능을 보이나 이것이 실제 인식환경에 적용된다면 성능이 급격히 저하될 수 있다. 이것은 음성인식을 수행하는 실제 환경이 주변 소음, 발성 거리, 마이크 특성, 채널 왜곡 및 화자의 변이 등 인식 성능을 저하시키는 요소들을 수반하기 때문이다[1]. 부가적인 잡음은 음성신호를 오염시키며 음성을 표현하는 특징벡터를 변화시킨다. 그래서 특징벡터의 통계적 특성의 변이를 유발한다. 예를 들어, 백색 잡음은 스펙트럼의 포락선 정보를 표현하는 켑스트럼과 같은 특징벡터의 동적 범위 (또는 분산)를 감소시킨다. 실제 인식기가 사용될 경우와 유사한 조건에서 학습된 시스템은 좋은 성능을 보인다. 그래서 음성인식 기술 개발자들은 학습과 인식 환경의 불일치에 의한 인식

책임저자: 최 승 호 (shchoi@snut.ac.kr)
서울특별시 노원구 공릉2동 172번지
서울산업대학교 전자정보공학과
(전화: 02-970-6461 팩스: 02-979-7903)

성능의 저하를 최소화하기 위한 노력을 기울여 왔다. 본 연구는 잡음환경에서의 강인한 음성인식을 위한 특징추출에 관한 것이다.

잡음 환경에서의 음성인식을 위한 방법으로서 Cambridge 대학에서 제안한 병렬 모델 결합 (Parallel Model Combination, PMC) 방식[2]과 같은 HMM 모델 영역에서의 보상 방식이 있다. 그리고, 인식환경의 변화를 보상하기 위한 특징 파라미터 영역에서의 처리 기법들이 있다. 켑스트럼과 같은 특징 파라미터를 정규화하기 위한 가장 간단한 방법으로서 켑스트럼 벡터의 차수별로 통계적 평균치를 차감하는 방법인 켑스트럼 평균 정규화 (Cepstral Mean Normalization, CMN) 기법이 있으며, Codeword-Dependent Cepstral Normalization (CDCN), RATZ라고 일컬어지는 multi-vaRiate gAussian-based cepsTrum normaliZation 기법 등도 이에 관한 것이다[3, 4]. 최근에는 다중 모델을 이용한 특징 보상 기법이 제안되었다[5]. 이와 같이 켑스트럼의 평균 및 분산 등을 정규화하여 통계적인 불일치를 줄이기 위한 방법들이 개발되었으나 이것은 확률밀도함수 (probability density function, pdf)를 부분적으로만 정규화 하는 것이다[6].

본 연구는 잡음에 의해 특징 파라미터가 왜곡이 되었을 때 이를 보상하여 인식성능의 저하를 최소화하기 위해 확률밀도를 정규화 하는 새로운 기법을 제안한다. 이것은 켑스트럼의 통계적 특성을 완벽하게 정규화 하는 켑스트럼 pdf 정규화 기법 (cepstrum pdf normalization, CPN)이다. CPN은 깨끗한 켑스트럼과 오염된 켑스트럼의 pdf들을 목적 pdf에 정규화하여 이들의 통계적 특성을 동일하게 한다. 다양한 분포를 고려하기 위하여 일반 가우시안 분포 (generalized Gaussian distribution, GGD)를 목적 pdf로 사용하였다. 또한 CPN의 계산량 부담을 줄이기 위해 표참조방식 (table lookup method)를 개발하였다. 2장에서 기존의 켑스트럼 1, 2차 통계치 정규화 기법을, 3장에서는 본 연구에서 제안한 CPN 기법을 설명하고, 4장에서는 음성인식 실험 및 고찰, 마지막으로 제5장에서 결론을 맺는다.

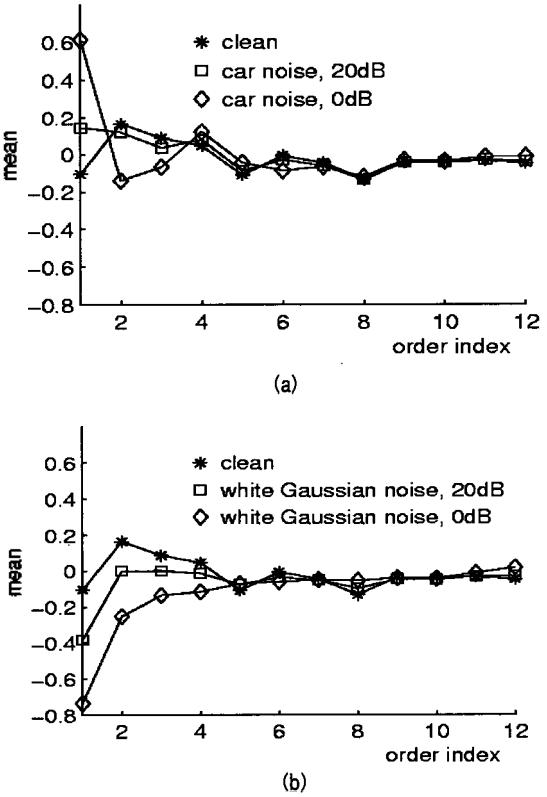
II. 켑스트럼의 평균 및 분산 정규화

깨끗한 켑스트럼과 잡음 섞인 켑스트럼 간의 통계적 차이를 예시하기 위하여 2400개의 발성음으로 구성된 음성데이터를 사용하여 다양한 잡음환경에서 켑스트럼의 평균치과 분산치를 계산하였다. 그림 2.1과 그림 2.2에서 보이는 것과 같이 평균의 변이와 분산의 감소와 같은 통계적 변이가 켑스트럼의 차수와 잡음의 종류 및 SNR 등에 따라 차이를 보인다. 특히, 백색잡음이 첨가되었을 경우 분산의 감소가 심하게 나타난다. 그림 2.3에는 발성음 /청와대/에 대한 3차 켑스트럼 계수의 열을 보였다. 그림에서 보듯이 SNR이 감소함에 따라 켑스트럼의 분산이 더욱 감소함을 알 수 있다. 그러나 다양한 잡음환경에서의 발성음에 대한 특징 파라미터 열은 깨끗한 켑스트럼의 그것과 전체적인 모양이 유사함을 알 수 있다. 이와 같은 관찰 결과로부터 각각의 발성음에 대해 분산치를 정규화 한다면 잡음환경에서의 음성인식의 성능이 향상될 수 있음을 기대할 수 있다.

켑스트럼의 분산 정규화 기법을 다음 식과 같이 가중함수를 사용하여 표현하자.

$$x_k'(t) = w_k x_k(t) \tag{1}$$

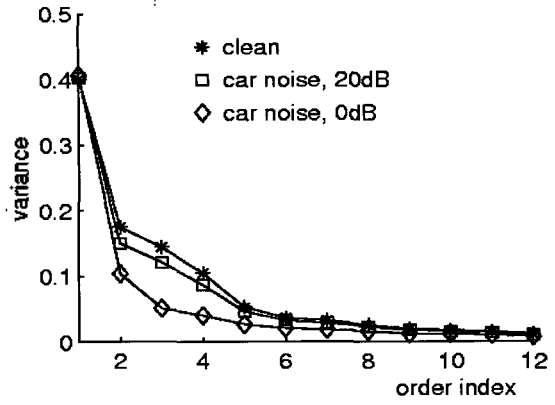
여기에서 k, x, x'은 각각 켑스트럼 차수, 정규화 이



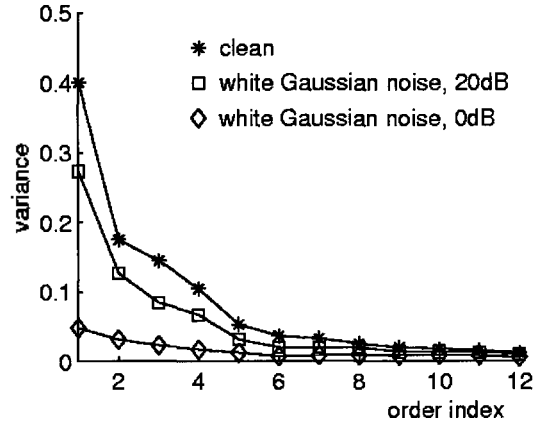
(a) 100km/h의 차량주행소음 첨가 (b) 백색잡음 첨가

그림 1. 켑스트럼의 평균치

Fig. 1. Mean value of cepstrum.



(a)



(b)

(a) 100km/h의 차량주행소음 첨가, (b) 백색잡음 첨가

그림 2. 켑스트럼의 분산치

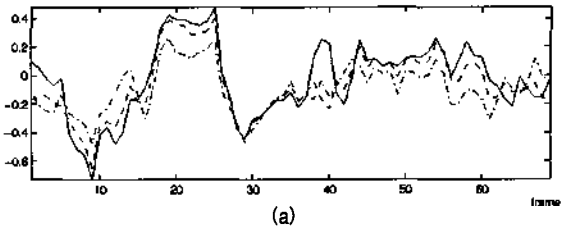
Fig. 2. Variance value of cepstrum.

전 켑스트럼 계수, 정규화 이후 켑스트럼 계수를 의미한다. 기존의 켑스트럼 분산 정규화 기법은 각각의 발성음의 켑스트럼에 대해 일종의 가중함수를 두는 것이 아니라 전체 학습데이터로부터 가중함수를 구한다[7]. 또는 켑스트럼 리프터링 (cepstrum liftering)과 같은 기법은 미리 결정된 가중함수를 갖고 각 차수에서 가중치들이 가해진다[8, 9]. 이러한 기존의 방법들은 모든 학습 데이터 및 모든 인식 데이터에 대해 일률적으로 고정된 가중함수들을 가진다.

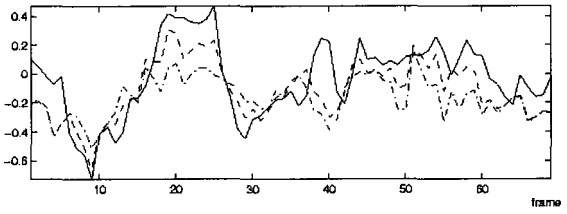
마이크 및 전화선과 같은 채널과 부가잡음 등은 켑스트럼의 평균의 변이를 초래한다. 이러한 켑스트럼 평균의 변이를 보상하기 위해 켑스트럼 평균 정규화 기법 (CMN)이 널리 사용되고 있으며, 다음 식으로 주어진다.

$$x_{CMN} = x - \mu_x \tag{2}$$

여기에서 μ_x는 한 개의 발성음으로부터 구한 켑스트럼의 평균벡터이다. 또한, 잡음환경에서의 켑스트럼의 분산의 감소를 보상하기 위해 켑스트럼의 분산을 정규화하는 기법들이 최근에 연구되고 있다[6]. 이러한 연구에서는 켑스트럼 분산 정규화 기법 (Cepstral Mean and Variance Normalization, CMVN)을 다음과 같이 정의한다.



(a)



(b)

(a) 100km/h의 차량주행소음 첨가 (b) 백색잡음이 첨가

(solid: clean, dashed: 10 dB SNR, dashed and dotted: 0 dB SNR)

그림 3. 발성음 /청와대/에 대한 3차 켑스트럼 계수의 열

Fig. 3. A sequence of third cepstral coefficient for an utterance /Cheong-Wa-Dae/.

$$x_{CMVN} = V_x^{-1/2} (x - \mu_x) \tag{3}$$

여기에서, V_x는 한 개의 발성음으로부터 구한 켑스트럼의 공분산 행렬이다.

III. 켑스트럼 pdf 정규화

본 연구에서 제안한 CPN 기법은 입력 켑스트럼을 원하는 분포를 가지도록 변환한다. 여기에서 임의의 랜덤변수를 원하는 분포를 가지는 또 다른 랜덤 변수로 변환하기위해 순서통계 (order statistics)의 기대치를 이용하는 득점기록 과정 (scoring procedure)를 도입하였다[10].

켑스트럼 벡터열 $\{x(t), t=1,2,\cdots,N\}$에 대해서 CPN 기법을 사용해서 정규화한 벡터열을 $\{x_{CPN}(t), t=1,2,\cdots,N\}$이라고 하자. 그리고 켑스트럼 벡터가 차수간에 독립적이라는 가정 하에 벡터를 차수별로 정규화

하며 $x(t)$와 같이 스칼라 표기를 사용한다. $x_{CPN}(t)$는 목표 pdf의 순서통계의 기대치를 사용한 득점기록 과정에 의해 다음 식과 같이 주어진다.

$$x_{CPN}(t) = E[z_{r(t)}] = \int_{-\infty}^{\infty} y f_{r(t)}(y)\, dy \tag{4}$$

여기서 z_r은 목표 pdf의 r번째 순서통계이고 $r(t)$는 켑스트럼 $x(t)$의 rank이고 $f_r(\cdot)$은 $z(r)$의 pdf이다. 순서통계의 기대치에 대한 정확한 해를 구하는 것이 어떤 pdf에 대해서는 가능하지만, rank $r(t)$ 또는 샘플 사이즈 N이 증가하면 정확한 해를 구하는 것이 어렵다. David와 Johnson은 [11]에서 식 (3.1)에 대해 $O(N^{-3})$의 에러로 근사화 하는 방법을 다음과 같이 제안하였다.

$$E[z_r] = Q(p_r) + \frac{p_r q_r}{2(N+2)} \frac{d^2 Q(p_r)}{dp^2} + \frac{p_r q_r}{(N+2)^2} \{ \frac{1}{3}(q_r - p_r) \frac{d^3 Q(p_r)}{dp^3} + \frac{1}{8} p_r q_r \frac{d^4 Q(p_r)}{dp^4} \} \tag{5}$$

여기서 $Q(\cdot)$은 목적 랜덤변수 z의 누적분포함수(Cumulative Distribution Function, CDF)의 역함수이며, $p_r = \frac{r}{N+1}$이고 $q_r = 1 - p_r$이다. 이때 $Q(\cdot)$를 구하기 위해서 수치해석을 이용한다.

목적 pdf로 선택된 GGD는 [12]에서 정의된 바와 같이 두 개의 상수인 분산 σ^2과 decay 파라미터 k (>0)를 파라미터로 하여 다음 식과 같이 대칭적 (symmetric)이고 unimodal한 밀도함수이다.

$$f(x;k) = \frac{k}{2A(k)\Gamma(\frac{1}{k})} e^{-|x|/A(k)^k} \tag{6}$$

여기서 $A(k) = [\sigma^2 \frac{\Gamma(\frac{1}{k})}{\Gamma(\frac{3}{k})}]^{\frac{1}{2}}$이고 $\Gamma(\cdot)$는 Gamma 함수이다. 본 연구에서는 GGD 분포의 형태를 결정하는 파라미터 중에서 분산 σ^2는 1로 고정하였으며, decay 파라미터 k를 변화시키며 실험을 수행하였다. 이때, k=1이고 k=2인 경우, GGD는 각각 Laplacian과 Gaussian 분포가 된다[12].

Decay 파라미터 k를 가지는 GGD의 CDF의 역함수를 $Q = Q(p;k)$ 라고 하고 이것이 수치적분에 의해 얻어진다고 하자. 식 (3.2)에서 필요한 $Q = Q(p;k)$의 미분값들은 다음과 같이 유도된다.

$$\frac{d^2 Q(p;k)}{dp^2} = \frac{\eta_k \cdot sign(Q) \cdot |Q|^{k-1}}{f(Q;k)^2} \tag{7}$$

$$\frac{d^3 Q(p;k)}{dp^3} = \frac{\eta_k \cdot |Q|^{k-2}}{f(Q;k)^3} (k-1) + 2\eta_k |Q|^k \tag{8}$$

$$\frac{d^4 Q(p;k)}{dp^4} = \frac{\eta_k \cdot sign(Q) \cdot |Q|^{k-3}}{f(Q;k)^4} (k-1)(k-2) + 7\eta_k |Q|^k + 6\eta_k^2 |Q|^{2k} \tag{9}$$

여기서 $\eta_k = \frac{k}{A(k)^k}$이다. 식 (3.4), (3.5), (3.6)을 식 (3.2)에 대입하여 decay 파라미터 k를 가지는 평균치의 근사치를 구할 수 있다.

위와 같은 수치적 방법과 더불어 본 연구에서는 표참조 방식 (table lookup)을 고안하였다. 샘플 사이즈가 각각 N_R이고 N_C인 r번째 순서통계의 평균값들을 $S_R(r)$과 $S_C(r)$이라고 하자. 첫 번째, 계산량을 줄이기 위해 프레임의 길이가 N_R인 기준표 $T_R = S_R(r)$, $r = 1, 2, \cdots, N_R$을 만든다. 그리고 기준표 T_R을 사용하여 현재 발성음의 표 $T_C = S_C(r)$, $r = 1, 2, \cdots, N_C$를 다음 식과 같이 근사화 한다.

$$S_C(r) = S_R(1 + \left| \frac{r-1}{N_C - 1} (N_R - 1) \right|), \quad r = 1, 2, \cdots, N_C \tag{10}$$

여기서 N_C는 현재 발성음의 프레임 수이고 $|\cdot|$는 버림 (rounding off) 연산자이다. 현재 테이블 T_C를 사용하여 정규화된 켑스트럼 열 $x_{CPN}(t)$은 다음의 식으로 구한다.

$$x_{CPN}(t) = S_C(r(t)), \quad t = 1, 2, \cdots, N_C \tag{11}$$

IV. 음성인식실험 및 고찰

음성인식 실험은 음성학적으로 균형잡힌 100개의 고립단어로 이루어진 데이터베이스를 사용하고 잡음 환경 실험을 위해 다양한 SNR의 백색가우시안잡음 (white

표 1. 백색가우시안잡음 환경에서의 인식률 (%) (A: 수치적분방식, B: 표참조방식)
Table 1. Recognition results (%) in the white Gaussian noise environment. (A: numerical method, B: table lookup method)

알고리즘		SNR(dB)				
		20	10	5	0	-5
CMN		75.1	21.4	8.4	4.3	3.1
CMVN		97.8	90.3	70.3	41.1	12.4
CPN (k=1.0)	A	96.9	89.4	72.4	45.4	22.0
	B	97.2	89.3	73.7	45.8	23.1
CPN (k=1.5)	A	97.8	88.8	74.0	48.1	23.9
	B	97.9	91.4	76.9	51.3	26.8
CPN (k=2.0)	A	97.9	90.9	75.3	47.6	24.9
	B	97.7	90.8	75.3	47.9	25.4
CPN (k=2.5)	A	97.9	90.9	73.6	47.3	23.5
	B	97.6	90.8	73.6	48.1	24.1
CPN (k=3.0)	A	97.8	90.9	73.2	46.4	23.5
	B	97.8	90.7	73.2	46.8	24.1

표 2. 차량주행소음 (100km/h) 환경에서의 인식률 (%) (A: 수치적분방식, B: 표참조방식)
Table 2. Recognition results (%) in the car noise environment. (A: numerical method, B: table lookup method)

알고리즘		SNR(dB)				
		20	10	5	0	-5
CMN		94.3	69.0	38.6	15.1	10.1
CMVN		97.6	93.2	84.9	62.6	31.4
CPN (k=1.0)	A	97.6	93.9	88.1	71.6	40.2
	B	97.6	93.8	87.6	71.4	39.3
CPN (k=1.5)	A	98.1	94.4	89.8	74.6	44.1
	B	98.1	95.2	91.4	77.9	46.9
CPN (k=2.0)	A	98.4	94.3	89.2	74.1	44.9
	B	97.8	94.4	89.1	73.2	44.2
CPN (k=2.5)	A	97.8	94.1	88.4	73.1	43.6
	B	97.8	94.3	88.6	72.7	43.3
CPN (k=3.0)	A	98.1	93.6	88.3	71.3	42.2
	B	97.6	93.7	88.2	70.8	42.4

Gaussian noise)과 차량주행소음을 부가하였다. 훈련을 위해 12명의 화자로부터 수집된 2,400개의 발성음을 사용하고 인식테스트를 위해 8명의 화자로부터 수집된 1,600개의 발성음을 사용하였다. 각 발성음은 일반사무실 환경에서 녹음되었고 표본화와 양자화 비율은 8kHz와 16bit이다. 특징파라미터는 12차 MFCC (Mel-Frequency Cepstral Coefficients) 벡터이고 음향모델은 반연속 HMM (semicontinuous HMM)을 사용하였다. 목적 pdf로서 사용한 GGD는 분산이 1이고 decay 파라미터를 달리하면서 실험을 하였다 (k=1,1.5,2,2.5,3). 인식테스트시 표참조방식에서는 프레임 수 $N_R=100$ 이다.

표 3.1과 표 3.2에서 보듯이 제안한 CPN 기법은 기존의 켑스트럼 정규화 기법보다 우수한 성능을 보였으며, $k=1.5$에서 가장 성능이 좋다. 표참조방식이 사용되었을 때 인식결과는 수치적분 (numerical integration) 방법의 결과와 유사함을 알 수 있었다. 특히, CPN 기법은 SNR이 낮은 경우에 유용함을 알 수 있다.

V. 결 론

본 연구에서는 잡음환경에서의 음성인식을 위해 켑스트럼의 확률분포를 정규화하는 새로운 기법 (CPN)을 제안하였다. 켑스트럼의 평균이나 분산 등을 정규화하는 즉, 부분적으로 확률분포를 정규화 하는 기존의 방법들과는 달리 CPN은 켑스트럼의 확률밀도함수 (pdf)를 완벽하게 정규화 하는 것이다. 고립단어 인식실험 결과, CPN이 기존의 방법들에 비해 우수한 성능을 보였으며 특히, SNR이 낮은 경우에 매우 유용함을 알 수 있었다. 또한, 계산량의 감축을 위해 표참조방식을 개발하였다.

참 고 문 헌

1. J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*, (Kluwer Academic Publishers, 1996).
2. M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," IEEE Trans. on Speech and Audio Process., **4** (5), 352-259, Sep. 1996.
3. A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, (Kluwer Academic Polishers, Boston, MA, 1993).
4. P. J. Moreno, B. Raj, E. Gouvea, and R. M. Stern, "Multivariate-Gaussian-based cepstral normalization for robust speech recognition," in Proc. ICASSP, 137-140, May 1995.
5. 김우일, 고한석, "시변 잡음에 대처하기 위한 다중 모델을 이용한 PCMM 기반 특징 보상 기법," 한국음향학회지, **23** (6), 473-480, Aug., 2004.
6. O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in Proc. ICASSP, 733-736, 1998.
7. Y. Tohkura, "A weighted cepstral distance measure for speech recognition," IEEE Trans. on Acoust. Speech and Signal Process., **35** (10), 1414-1422, Oct. 1987.
8. M. R. Schroeder, "Direct (nonrecursive) relations between cepstrum and predictor coefficients," IEEE Trans. on Acoust.

Speech and Signal Process., **29** (2), 297-301, Apr. 1981.
9. J. C. Junqua and H. Wakita, "A comparative study of cepstral lifters and distance measures for all pole models of speech in noise," in Proc. of ICASSP, 476-479, May 1989.
10 H. A. David, *Order statistics*, (John Wiley & Sons, NY, 1981).
11 F. N. David and N. L. Johnson, "Statistical treatment of censored data, Part I. fundamental formulae," Biometrika, 41, 228-240, 1956.
12 S. A. Kassam, *Signal detection in non-Gaussian noise*, (Springer-Verlag, NY, 1988).

저자 약력

• 석 용 호 (Yong Ho Suk)

1992년: 한양대학교 공과대학 전자공학과 (공학사)
1994년: 한국과학기술원 전기및전자공학과 (공학석사)
2000년: 한국과학기술원 전기및전자공학과 (공학박사)
2000년~2001년: (주) 에스원
2001년~현재: (주) 엠큐브웍스
※주관심분야: 음성신호처리, Multimedia on Mobile Environments, Digital Multimedia Broadcasting, Digital Video Broadcasting

• 최 승 호 (Seung Ho Choi)

1991년: 한양대학교 공과대학 전자공학과 (공학사)
1993년: 한국과학기술원 전기및전자공학과 (공학석사)
1999년: 한국과학기술원 전기및전자공학과 (공학박사)
1996년~2002년: 삼성종합기술원 전문연구원
2002년~현재: 서울산업대학교 전자정보공학과 조교수
※주관심분야: 음성인식, 음성코딩, 멀티미디어신호처리, 디지털통신

• 이 황 수 (Hwang-Soo Lee)

한국음향학회지 제20권 제2호 참조.