개선된 PMC 동적 파라미터 보상에 의한 강인한 음성인식

정규준°, 조훈영, 오영환
한국과학기술원 전자전산학과 전산학전공

Improved compensation of dynamic parameter in PMC for robust speech recognition

Gue-Jun Jung°, Hoon-Young Cho, Yung-Hwan Oh
Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology
E-mail: {sylph, hycho, yhoh}@bulsai.kaist.ac.kr

요 약

본 논문에서는 모델 기반의 잡음 보상 방법인 PMC (Parallel Model Combination)의 동적 파라미터의 보상에 관해 논한다. PMC는 연속 HMM의 모델 단계에서 음성 모델과 잡음 모델을 조합하여 학습 환경과 평가 환경간의 차이를 보상한다. 기존 PMC의 경우, 동적 파라미터 보상 시에 음성과 잡음 모델의 평균 벡터 및 공분산 행렬에 대해 많은 양의 계산이 요구된다. 본 논문에서는 공분산의 수축-확대 방법을 동적 파라미터 보상 과정에 적용하고 짧은 구간에서 음성과 잡음의 변화가 적다는 가정 하에 보상식을 단순화함으로써, 연산량을 줄이면서도 인식 성능을 향상시킬 수 있는 방법을 제안한다.

100단어 규모의 고립단어 인식시스템에 대해 백색 및 헬리콥터 잡음을 SNR 0, 5, 10 dB로 가산한 3000개의 평가 자료로 실험한 결과, 기존 방법에 비해 인식 성능 개선 및 보상 속도 향상을 얻음으로서 제안한 방식의 유효성을 검증할 수 있었다.

1. 서 론

현재 음성 인식 기술은 연구 단계를 지나 실용화 단계에 접어들고 있다. 그러나 음성 인식 기술을 실용화하기 위해서는 학습 환경과 사용 환경의 불일치에 의해 발생하는 음성 인식 시스템의 성능 저하 문제를 해결해야 한다. 일반적으로 불일치 요인을 음성신호의 스펙트럼 영역에서 가산적인 배경 잡음과 음성의 켑스트럼 영역에서 가산적인 채널 왜곡으로 구분한다. 이러한 불일치 요인을 제거하기 위한 기존의 연구 방법은 크게 세가지 방법으로 구분할 수 있다. 첫 번째는 MFCC (Mel Frequency Cepstral Coefficient), EIH (Ensemble Interval Histogram), SMC (Short-time Modified Coherence), PLP (Perceptual Linear Prediction) 분석과 같이 음성에서 본질적으로 잡음에 강인한 특징을 추출하는 방법이다 [4]. 두 번째는 스펙트럼 차감법, 상태기반 음질개선(state-based speech enhancement) 등과 같이 잡음이 섞여있는 음성에서 깨끗한 음성을 얻기 위한 전처리 방법이다 [9][10]. 세 번째는 PMC, speech and noise decomposition, MLLR (Maximum Likelihood Linear Regression)과 같이 음성 인식 모델이 학습된 환경을 사용 환경과 비슷하도록 수정하는 방법이다 [3].

본 논문에서는 음성 인식 모델을 보상하는 연구 방식 중 하나인 PMC를 살펴보고, PMC의 보상 방법 중 동적 파라미터를 좀 더 효율적으로 보상하여 인식 성능을 향상시키는 방법을 제안한다. 제안한 방법은 공분산 수축-확대 방법 [6]을 동적 파라미터 보상 과정에 적용하고 짧은 구간에서 음성과 잡음의 변화가 적다는 가정 하에 보상식을 단순화함으로써, 연산량을 줄이면서도 인식 성능을 향상시킨다.

본 논문의 구성은 다음과 같다. 2장에서는 일반적인 PMC에 관해 논의한다. 3장에서는 동적 파라미터 보상을 위해 기존에 취해진 방법을 살펴보고 개선점에 관해 설명한다. 4장에서는 실험 환경 및 제안된 방법에 따른 실험 결과를 기술하고, 5장에서 결론을 맺는다.

2. Parallel Model Combination

PMC는 깨끗한 음성 모델과 잡음 모델을 조합하여 인식 모델을 환경에 적응시킨다 [1][2]. 잡음이 혼합된 음성을 가장 잘 인식할 수 있는 방법은 동일한 잡음 환경에서 자료를 수집하고 인식기를 재학습시키는 것이다. 그러나 이러한 방법은 실용적이지 못하다. 만약 음성 모델이 학습 자료의 통계적 특성을 잘 가지고 있다면 그림 1과 같은 모델 파라미터 보상으로 동일한 효과를 얻을 수 있다.

음성과 잡음 신호는 선형 스펙트럼 영역에서 가산적으로 이루어지기 때문에 각 모델의 파라미터를 선형 스펙트럼 영역으로 변환하여 조합한다. 조합할 때 기준이 되는 식을 불일치 함수라 하며, 이는 다음의 가정에 기초한다 [2].

1) 음성과 배경 잡음은 상호 독립적이다.

2) 음성과 배경 잡음은 시간 영역에서 가산적이다.
3) 단독 다변량 가우스 모델(single multivariate Gaussian model)로 음성과 배경 잡음 정보를 충분히 알 수 있다.
4) 잡음 첨가 후에도 프레임 및 HMM 모델의 상태 배열은 유지된다.

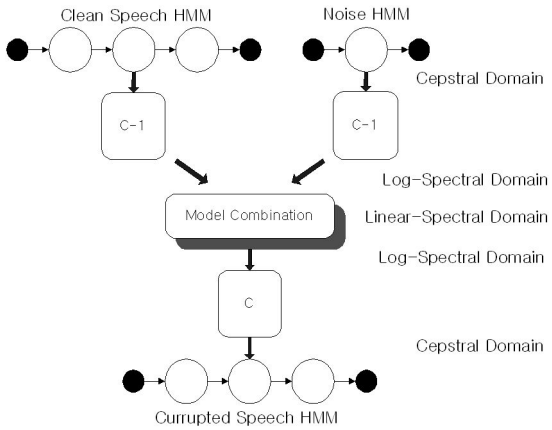

그림 1. Basic PMC process

2.1. 정적 파라미터의 불일치 함수

음성과 배경 잡음의 정적 요소는 로그 스펙트럼 영역에서 다음의 불일치 함수에 근거하여 혼합된다.

$$\begin{aligned} O^l(\tau) &= F(S_i^l(\tau), N_i^l(\tau)) \\ &= \log(g\ \exp(S_i^l(\tau)) + \exp(N_i^l(\tau))) \end{aligned} \quad (1)$$

식에서 g는 음성과 잡음 크기 비를 조절하며, τ는 프레임 인덱스이다. S와 N은 각각 깨끗한 음성과 잡음을 위첨자 l은 로그 스펙트럼 영역을 의미한다. 식 (1)을 기준으로 다음과 같이 모델의 정적 파라미터가 보상된다.

$$\widehat{\mu_i^l} = E[F(S_i^l, N_i^l)] \quad (2)$$

$$\widehat{\Sigma_{ij}^l} = E[F(S_i^l, N_i^l)F(S_j^l, N_j^l)] - \widehat{\mu_i^l}\,\widehat{\mu_j^l} \quad (3)$$

식에서 $\widehat{\mu}$과 $\widehat{\Sigma}$는 각각 보상된 Gaussian 분포의 평균과 공분산을 의미한다.

2.2. 동적 파라미터의 불일치 함수

동적 파라미터는 전후 프레임 간의 상호 관련성을 모델링한 것으로 배경 잡음이 일정할 경우 좋은 성능을 얻을 수 있다. 일반적으로 전후 프레임 간의 차를 이용하는 차분(delta), 가속(acceleration) 파라미터가 많이 사용되며 이들의 정의는 다음과 같다.

$$\Delta O^c(\tau) = O^c(\tau + w) - O^c(\tau - w) \quad (4)$$

$$\Delta^2 O^c(\tau) = \Delta O^c(\tau + w_a) - \Delta O^c(\tau - w_a) \quad (5)$$

식에서 위첨자 c는 켑스트럼을 의미하며 w, w_a는 파라미터 계산에 사용하는 프레임 간격을 의미한다. 동적 파라미터는 다음 불일치 함수를 기초로 혼합된다 [2].

$$\begin{aligned} \Delta O_i^l(\tau) &= F^{\Delta}(S_i^l(\tau - w), N_i^l(\tau - w), \Delta S_i^l(\tau), \Delta N_i^l(\tau)) \\ &= \log(\exp(\Delta S_i^l(\tau) + S_i^l(\tau - w) + g^l) \\ &\quad + \exp(\Delta N_i^l(\tau) + N_i^l(\tau - w))) \\ &\quad - \log(\exp(S_i^l(\tau - w) + g^l) + \exp(N_i^l(\tau - w))) \end{aligned} \quad (6)$$

$$\begin{aligned} \Delta^2 O_i^l(\tau) &= F^{\Delta^2}(S_i^l(\tau), N_i^l(\tau), \Delta S_i^l(\tau - w), \\ &\quad \Delta N_i^l(\tau - w), \Delta^2 S_i^l(\tau), \Delta^2 N_i^l(\tau)) \\ &= \log(\exp(\Delta^2 S_i^l(\tau) + 2(S_i^l(\tau) - O_i^l(\tau))) \\ &\quad + \exp(\Delta^2 N_i^l(\tau) + 2(N_i^l(\tau) - O_i^l(\tau))) \\ &\quad + \exp(\Delta^2 N_i^l(\tau) + \Delta N_i^l(\tau - w) - \Delta S_i^l(\tau - w) \\ &\quad + S_i^l(\tau) + N_i^l(\tau) - 2O_i^l(\tau)) \\ &\quad + \exp(\Delta^2 S_i^l(\tau) + \Delta S_i^l(\tau - w) - \Delta N_i^l(\tau - w) \\ &\quad + S_i^l(\tau) + N_i^l(\tau) - 2O_i^l(\tau))) \end{aligned} \quad (7)$$

$g^l = \log(g)$이며, 다른 변수들은 식 (1)과 동일하다. 동적 파라미터는 식 (6), (7)을 식 (2), (3)과 같이 계산하여 얻을 수 있지만 이를 정확히 계산할 수 있는 닫힌 형태(closed form)의 해법이 없기 때문에 값을 얻기가 매우 어렵다 [3].

3. 동적 파라미터 보상

3.1. 연속시간 근사법

앞장에서 기술한 바와 같이 동적 파라미터를 식 (6), (7)에 근거하여 보상하기는 매우 어렵다. 대신 연속 시간 근사(continuous time approximation)를 이용한 보상법이 제안되었다 [5]. 이 방법은 동적 파라미터 추정시 식 (6), (7) 대신 정적 파라미터 불일치 함수인 식 (1)을 시간에 대해 일, 이차 미분한 식을 이용한다.

$$\Delta O_i^l \approx \frac{\delta O_i^l}{\delta t} = \alpha \frac{\delta S_i^l}{\delta t} + \beta \frac{\delta N_i^l}{\delta t} \quad (8)$$

$$\Delta^2 O_i^l \approx \frac{\delta^2 O_i^l}{\delta t^2} = \alpha \frac{\delta^2 S_i^l}{\delta t^2} + \beta \frac{\delta^2 N_i^l}{\delta t^2} + \alpha\beta\left(\frac{\delta S_i^l}{\delta t} - \frac{\delta N_i^l}{\delta t}\right)^2 \quad (9)$$

$$\left(\alpha = \frac{\exp(S_i^l)}{\exp(S_i^l) + \exp(N_i^l)},\ \beta = \frac{\exp(N_i^l)}{\exp(S_i^l) + \exp(N_i^l)}\right)$$

식 (8), (9)를 이용하여 다음과 같은 식을 얻을 수 있다 [5].

$$\Delta \widehat{\mu_i^l} = E[\Delta O_i^l] \approx \alpha' \Delta \mu_i^l + \beta' \Delta \widehat{\mu_i^l} \quad (10)$$

$$\begin{aligned} \Delta^2 \widehat{\mu_i^l} = E[\Delta^2 O_i^l] &\approx \alpha' \Delta^2 \mu_i^l + \beta' \Delta^2 \widehat{\mu_i^l} + \alpha'\beta'[(\Delta \mu_i^l)^2 \\ &\quad + \Delta \Sigma_{ii}^l + (\Delta \widehat{\mu_i^l})^2 + \Delta \widehat{\Sigma_{ii}^l} - 2\Delta \mu_i^l \Delta \widehat{\mu_i^l}] \end{aligned} \quad (11)$$

$$\left(\alpha' = \frac{\exp(\mu_i^l)}{\exp(\mu_i^l) + \exp(\widehat{\mu_i^l})},\ \beta' = \frac{\exp(\widehat{\mu_i^l})}{\exp(\mu_i^l) + \exp(\widehat{\mu_i^l})}\right)$$

연속 시간 근사법을 이용하여 공분산을 보상할 수도 있지만 공분산의 보상은 많은 계산량에도 불구하고 성능 향상을 가져오지 않는다 [5]. 연속 시간 근사법은 음성과 배경 잡음의 공분산이 작다는 기본 가정이 필요하지만 실제의 경우 공분산이 작지 않기 때문에 이 방법으로 가속 파라미터 보상할 경우 오히려 인식률이 저하되기도 한다.

3.2. 공분산 수축-확대를 적용한 근사법

본 논문에서는 연속 시간 근사법의 가정을 지키면서 효율적인 보상식을 얻기 위해 공분산 수축-확대(contraction and expansion of covariance) 방법을 동적 파라미터를 보상에 적용한다.

공분산 수축-확대 방법은 정적 파라미터의 영역 변환시 공분산의 영향을 줄이기 위해 제안된 방법으로 그림 2에서 나타낸 바와 같은 순서로 진행되며 수축-확대 정도는 상수 γ에 의해 결정된다. γ는 평균과 분산을 보상할 때 로그 스펙트럼 영역에서의 공분산 영향을 효과적으로 조절할 수 있는 값이다. 또한 이 방법을 통해 보상된 모델들은 자료의 식별력이 향상되어 낮은 SNR에서도 인식률 향상을 가져온다 [6].

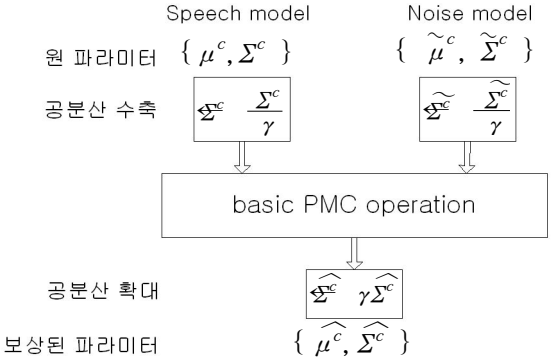

그림 2. 공분산 수축-확대 과정

이 과정을 동적 파라미터에 적용하면 식 (11)은 식 (12)와 같이 바뀐다. 식에서 γ값을 적절히 크게 하면 공분산의 영향을 줄일 수 있다.

$$\Delta^2 \widehat{\mu_i^l} = E[\Delta^2 O_i^l] \approx \alpha' \Delta^2 \mu_i^l + \beta' \Delta^2 \widetilde{\mu_i^l} + \alpha'\beta'[(\Delta\mu_i^l)^2 + (\Delta \widetilde{\mu_i^l})^2 + \frac{\Delta\Sigma_{ii}^l}{\gamma} + \frac{\Delta \widetilde{\Sigma_{ii}^l}}{\gamma} - 2\Delta\mu_i^l \Delta \widetilde{\mu_i^l}] \quad (12)$$

3.3. 빠른 동적 파라미터 근사법

공분산 수축-확대 방법에서 γ는 1보다 큰 상수를 이용한다. 만약 γ값이 매우 큰 상수일 경우 공분산은 매우 작은 값이 되며 이것은 다음과 같이 근사된다 [6].

$$0 \approx \left|\frac{\Delta\Sigma_{ij}^l}{\gamma}\right| \ll 1 \ , \ 0 \approx \left|\frac{\Delta\widetilde{\Sigma_{ij}^l}}{\gamma}\right| \ll 1$$

이러한 가정을 적용할 경우 식 (12)는 식 (13)과 같이 간소화된 형태로 표현된다.

$$\Delta^2 \widehat{\mu_i^l} = E[\Delta^2 O_i^l] \approx \alpha' \Delta^2 \mu_i^l + \beta' \Delta^2 \widetilde{\mu_i^l} + \alpha'\beta'[(\Delta\mu_i^l)^2 + (\Delta \widetilde{\mu_i^l})^2 - 2\Delta\mu_i^l \Delta \widetilde{\mu_i^l}] \quad (13)$$

이와 같이 보상할 경우 영역 변환에 많은 연산량이 요구되는 공분산을 계산할 필요가 없으므로 보상 시간이 빨라진다. 또한 잡음이 일정한 형태로 유지된다는 잡음 가정을 적용하면 잡음 모델의 차분 파라미터와 가속 파라미터의 평균을 0으로 근사된다. 이 가정에 의해 식 (10)과 식 (13)을 다음과 같이 간소화할 수 있다.

$$\Delta \widehat{\mu_i^l} = E[\Delta O_i^l] \approx \alpha' \Delta\mu_i^l \quad (14)$$

$$\Delta^2 \widehat{\mu_i^l} = E[\Delta^2 O_i^l] \approx \alpha' \Delta^2 \mu_i^l + \alpha'\beta'(\Delta\mu_i^l)^2 \quad (15)$$

3.4. 개선된 동적 파라미터 근사법

식 (10)과 (13)에서 α', β'은 각각 0과 1사이의 상수이며 $\alpha' + \beta' = 1$ 이다. 짧은 구간의 음성과 잡음의 변화는 크지 않다는 가정을 적용하면 다음과 같이 근사할 수 있다.

$$\alpha'\beta'[(\Delta\mu_i^l)^2 + (\Delta \widetilde{\mu_i^l})^2 - 2\Delta\mu_i^l \Delta \widetilde{\mu_i^l}] \approx 0$$

이 근사를 식 (13)에 적용하면 동적 파라미터는 다음과 같은 형태로 보상할 수 있다.

$$\Delta \widehat{\mu_i^l} = E[\Delta O_i^l] \approx \alpha' \Delta\mu_i^l + \beta' \Delta \widetilde{\mu_i^l} \quad (10)$$

$$\Delta^2 \widehat{\mu_i^l} = E[\Delta^2 O_i^l] \approx \alpha' \Delta^2 \mu_i^l + \beta' \Delta^2 \widetilde{\mu_i^l} \quad (16)$$

4. 실험 및 결과

실험에 사용된 음성 자료는 국어 공학 연구 센터에서 만든 PBW452 단어 중 100단어를 사용하였고 [7], 잡음 자료는 NoiseX 92에 있는 자료 중 백색 잡음과 헬리콥터 잡음을 이용하였다 [8]. 학습 자료는 남, 여 각각 50명의 깨끗한 음성을 이용하였으며, 평가 자료는 학습에 포함되지 않은 남, 여 각각 18명, 12명에 대한 3000개의 단어 발성을 이용하였다. 평가 자료는 SNR 0, 5, 10dB 3단계로 잡음을 혼합하였으며 잡음 모델 생성을 위해 각 SNR 당 20초 정도의 잡음을 준비하였다. 인식 모델은 HTK 3.0을 이용하여 생성하였으며 597개의 triphone 모델로 이루어졌다. 각 HMM은 3개의 상태로 구성되어 있으며 각 상태는 1개의 Gaussian분포를 가진다. 잡음 모델은 1개의 상태로 구성되며 1개의 Gaussian분포로 이루어졌다. 음성 분석 구간은 20ms이고, 10ms씩 이동하며 해밍창을 사용하였다. 특징 벡터로는 0차를 포함한 13차 MFCC, 13차 차분 파라미터, 13차 가속 파라미터를 사용하였다.

인식 모델의 기본 성능을 알아보기 위해 모델 생성 후 각 SNR별 인식률을 측정해 보았으며 결과는 표 1과 같았다.

표 1. 인식기의 SNR별 인식률(%)

	clean	10dB	5dB	0dB
백색잡음	99.83	46.63	21.87	2.63
헬리콥터잡음	99.83	68.30	23.16	2.26

제안된 동적 파라미터 보상의 유효성을 검증하기 위한 실험 결과는 표 2와 같았다.

표 2. 동적 파라미터 보상 시 인식률(%)

백색잡음	10 dB	5 dB	0 dB
S	88.83	63.43	27.33
S+D	92.73	66.53	41.07
S+D+A	81.80	60.63	31.83
S+D+CEA	89.83	77.66	50.43
S+D+FA	89.83	77.86	50.77
S+ND+NA	88.87	74.03	42.23
S+D+PA	94.43	84.83	56.07

헬리콥터잡음	10 dB	5 dB	0 dB
S	93.60	69.47	23.33
S+D	95.70	79.50	33.87
S+D+A	78.60	57.07	17.80
S+D+CEA	92.43	78.00	37.53
S+D+FA	92.47	78.03	37.60
S+ND+NA	89.83	71.20	33.96
S+D+PA	96.53	78.03	39.33

표 2에서 S는 정적 파라미터 보상, D는 차분 파라미터 보상, A는 연속 시간 근사법에 의한 기존 가속 파라미터 보상, CEA는 3.2절의 공분산 수축-확대를 이용한 가속 파라미터 보상(γ =100), FA와 N은 각각 3.3절의 빠른 가속 파라미터 보상과 잡음 가정이 포함된 파라미터 보상, PA는 3.4절에서 제안한 방법을 이용한 가속 파라미터 보상을 의미한다.

제안한 방법에서 연산량 감소 효과를 확인하기 위해 보상 시간을 측정하였으며 결과는 표 3과 같다. 측정된 값은 전체 파라미터 보상 시간과 정적 파라미터 보상시간의 차이며 Pentium III-500 CPU, 128M RAM, windows 2000 환경에서 측정하였다.

표 3. 1791개의 Gaussian 분포의 동적 파라미터 보상에 소비된 시간 비교

Dim	기존 방법 (S+D+A)	빠른 보상방법 (S+D+FA)	제안한 방법 (S+D+PA)
39차	21.73 s	2.51 s	2.48 s

연속 시간 근사법에 의한 기존 방법은 차분 파라미터 보상의 경우 인식 성능의 향상을 가져오지만, 가속 파라미터 보상은 오히려 인식률을 저하시켰다. 그러나 공분산 수축-확대 방법을 적용하였을 때는 인식률 저하를 막을 수 있었으며 SNR이 낮을 경우도 성능이 향상되었다. 빠른 보상 방법을 적용할 경우, 연산량을 감소시키면서도 인식률을 유지함을 알 수 있었으며 잡음 가정도 유효함을 알 수 있었다. 또한 제안한 보상 방법은 보상 시간 면에서 빠른 보상 방법과 비슷하면서도 인식 성능은 향상되었다.

5. 결 론

본 논문에서는 PMC의 파라미터 보상 방법 중 동적 파라미터 보상에 관해 살펴보았다. 기존 방법의 경우 연속 시간 근사법을 사용하여 동적 파라미터를 보상한다. 이 방법은 공분산의 값이 작다는 점을 기본 가정으로 하고 있으나 실제의 경우 공분산 값이 작지 않아 가속 파라미터를 보상할 경우 오히려 인식률이 저하되었다. 이 문제를 해결하기 위해 공분산 수축-확대 방법을 동적 파라미터 보상에 적용하였으며 이를 바탕으로 동적 파라미터 보상 방법을 제안하였다. 제안한 방법은 동적 파라미터 보상을 빠르게 하면서도 높은 인식률을 얻음을 실험을 통해 확인하였다.

향후에는 온라인 상에서 배경 잡음 변화를 정확히 인식하는 방법 및 실시간 모델 보상에 관한 연구가 필요할 것이다.

6. 참 고 문 헌

[1] M. J .F. Gales, S. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication, vol 12, pp. 231-240, 1993

[2] M. J. F. Gales, S. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEEE Transactions on Speech and Audio Processing, vol 4, pp. 352-359, 1996

[3] Y. Gong, "Speech recognition in noisy environments: A survey," Speech Communication, vol. 16, pp. 261-291, 1995

[4] J. C. Junqua, J. P. Haton, "Robustness in Automatic Speech Recognition: Fundamentals and Applications," Kluwer Academic Publishers, 1996

[5] R. A. Gopinath, M. J. F. Gales, P. S. Gopalakrishnan, S. Balakrishnan-Aiyer, M. A. Picheny, "Robust speech recognition in noise-performance of the IBM continuous speech recognizer on the ARPA noise spoke task," Proc. ARPA Workshop on Spoken Language Systems Technology, pp. 127-130, 1995

[6] T. H. Hwang, H. C. Wang, "A fast algorithm for parallel model combination for noisy speech recognition," Computer Speech and Language, vol 14, pp. 81-100, 2000

[7] 이용주, "음성데이터베이스의 현황 및 과제," 제13회 음성통신 및 신호처리워크샵, 13권, pp. 279-278, 1996

[8] A. P. Varga, H. J. M Steenken, M. Tomlinson, D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Technical report, DRA Speech Rearch Uint, 1992

[9] J. S. Lim, A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech,", Proceedings IEEE, 67:1586-1604, 1979

[10] Y. Ephraim, D. Malah, B. H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech", Proc. ICASSP'88, pp 533-536, 1992