1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

스펙트럼 차감법과 잡음 마스킹의 hybrid 방식을 이용한 잡음환경에서의 음성인식

권영욱, 김형순
부산대학교 전자공학과

Speech Recognition in the Noisy Environments using Hybrid Method of Spectral Subtraction and Noise Masking

Young Uk Kwon, Hyung Soon Kim
Dept. of Electronics Eng. Pusan National University

요 약

잡음환경에서의 음성인식 성능향상을 위하여 본 논문에서는 스펙트럼 차감법 이후에 남아 있는 잔여 잡음으로 인한 mismatch를 극복하는 수단으로 기존의 스펙트럼 차감법에서의 flooring factor를 사용하는 대신에 target 잡음레벨을 이용하여 잡음 마스킹을 적용하는 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식을 사용한다. 이 방법은 낮은 SNR에서 개선되지 않는 기존의 잡음 마스킹이 가지는 약점을 극복하고 동시에 스펙트럼 차감법에서의 잔여 잡음 문제를 완화시킬 수가 있었다. 특히 시간/주파수 영역 smoothing을 적용함으로써 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식의 적용 이후에도 여전히 남아 있는 일부 잡음을 추가적으로 감소시켰으며, 더욱 향상된 인식성능을 얻을 수 있었다.

Ⅰ. 서 론

잡음환경에서 음성인식의 성능향상을 위해 신치리과정을 통해 잡음을 제거하는 음질개선 방식들 중에서 현재 가장 널리 사용되는 대표적인 방법은 스펙트럼 차감법이다[1][2]. 스펙트럼 차감법은 잡음 스펙트럼의 평균을 빼주기 때문에 잡음의 분산이 큰 낮은 SNR에 대해서는 스펙트럼 차감법 이후에도 잔여 잡음이 많아지는 문제가 있다. 이 문제의 해결을 위해 추정된 잡음레벨을 기반으로 over-estimation factor 및 flooring factor를 주로 사용한다. 이 경우에도 어느 정도의 성능향상이 이루어지지만, over-estimation factor를 높게 할 경우에는 일부 음성부분도 손상을 받게 되고, 반면에 over-estimation factor을 낮게 할 경우에는 잡음부분이 충분히 제거되지 않는 문제점이 남는다. 이를 극복하기 위하여 본 논문에서는 잡음 마스킹 방법을 도입하여 기존의 스펙트럼차감법에서의 flooring factor를 사용하는 대신에 target 잡음레벨을 이용하여 잡음 마스킹을 적용하는 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식을 사용한다.

잡음 마스킹 방법은 필터뱅크 출력 에너지 레벨을 target 잡음레벨에 따라 마스킹 함으로써 환경적인 차이를 감소시키는 방식으로, 음성신호에 잡음이 공존시에 학습과 인식시의 환경이 일치될 때는 보다 높은 인식성능을 얻을 수 있다. 실제로 이 방식은 그 자체만으로도 자동차 소음 등의 잡음환경에 대해 인식성능을 개선시킨다[3][4]. 그러나 잡음 마스킹 방법은 SNR이 매우 낮거나 대역별로 심한 변화를 나타내는 잡음에 대해서는 실제 잡음레벨이 target 잡음레벨보다 높아지게 되므로 성능개선 효과를 얻을 수 없다[5].

스펙트럼 차감법과 잡음레벨 마스킹의 hybrid 방식은 스펙트럼 차감법 적용 이후에 남아 있는 잔여 잡음으로 인한 mismatch를 극복하는 수단으로 잡음 마스킹을 사용함으로써, 낮은 SNR에서 개선되지 않는 기존의 잡음 마스킹이 가지는 약점을 극복하고 동시에 스펙트럼 차감법에서의 잔여 잡음 문제를 완화시켰다. 특히 잔여 잡음이 target 잡음레벨에 의해 마스킹이 이루어지지만, 일부 잡음들이 여전히 남아 있어 이들에 의한 mismatch가 잔존한다. 따라서 본 논문에서는 시간/주파수 영역 smoothing을 적용하여 마스킹 이후에도 남아 있는 잔여 잡음을 추가적으로 감소시켰으며, 결과적으로 인식성능이 향상되었다.

Ⅱ. 잡음 마스킹

잡음환경에서 이미 구축한 인식기와 실제 인식대상 환경에 대한 mismatch를 줄이기 위한 방법으로 잡음 마스킹 방식이 있다. 이 방식은 식 (1)과 같이 멜-스케일 BPF의 출력값을 미리 정한 target 잡음레벨과 비교하여 높은 레벨의 값을 취하는 것이다. 즉, target 잡음레벨을 이용한 잡음 마스킹 방법을 사용하게 된다[3].

$$\hat{X}(f) = \max(Y(f),\ TH) \tag{1}$$

여기서 TH는 target 잡음레벨을 나타낸다.

이러한 잡음 마스킹 방식은 계산이 매우 간단한 방식으로 훈련용 깨끗한 음성의 낮은 에너지 값을 미리 정해진 실제 잡음환경의 target 잡음레벨과 유사한 레벨로 올려줌으로 훈련용 음성과 실제 잡음환경에서의 인식대상 음성과

의 불일치를 줄일 수 있다.

잡음 마스킹 방식의 실제적인 문제는 환경잡음에 적응된 target 잡음레벨을 정하는 것이며, 인식대상 음성의 환경이 어느 정도인가에 따라 결정되는 target 잡음레벨이 너무 낮으면 잡음에 대한 마스킹 효과가 작아지고 반면에 너무 높으면 음성신호 자체의 심한 왜곡을 초래하게 된다. 즉, 낮은 SNR인 경우에는 잡음음성의 잡음레벨이 target 잡음레벨보다 높은 값을 가지므로 이러한 잡음음성에서는 잡음 마스킹 알고리즘이 효과적으로 적용되지 않으므로, 인식의 결과는 현저히 저하된다.

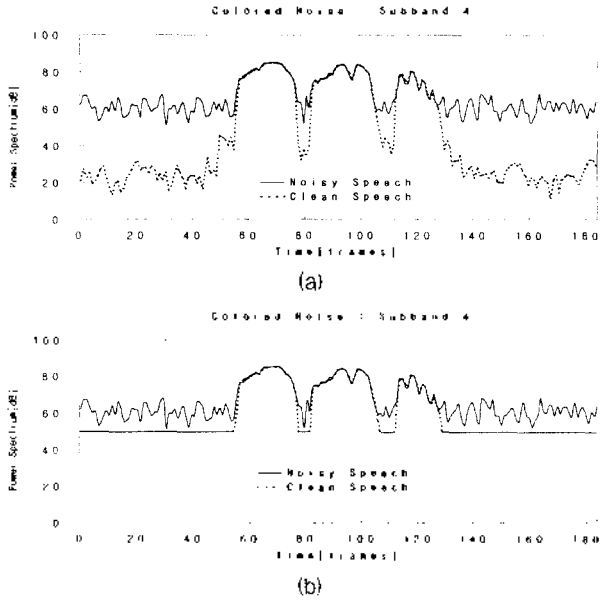

그림 1. 음성 "재무관리실"에 대한 잡음 마스킹의 예 (5dB 유색잡음)
(a) 원 음성 및 잡음음성에 대한 4번째 멜-주파수 필터뱅크의 출력
(b) Target 잡음레벨을 50dB로 마스킹을 한 결과

그림 1은 잡음이 없는 음성과 5dB 유색잡음이 부가된 잡음음성에 대한 4번째 멜-스케일 필터뱅크 출력 에너지 및 target 잡음레벨 50dB에 대한 마스킹 결과를 나타내고 있다. 그림 1의 (a)에서 잡음이 없는 경우와 잡음음성간의 동적 범위가 30dB 이상의 차이를 나타내고 있다. 이러한 불일치를 줄이기 위해서 잡음이 없는 원 음성을 대상으로 여러 적절한 target 잡음레벨로 마스킹을 함으로 잡음음성의 동적 잡음레벨과의 차이를 줄여야 한다. (b)는 (a)와 동일한 출력에서 target 잡음레벨을 50dB로 하여 잡음 마스킹을 수행한 이후의 결과를 나타낸 것이다. 그림에서 잡음 마스킹 이후는 깨끗한 음성에서의 에너지 레벨은 target 잡음레벨로 마스킹이 잘 이루어지고 있으나, 잡음음성에 대한 잡음/비음성 구간에서의 마스킹은 거의 이루어지지 않으며 target 잡음레벨과는 여전히 차이를 나타내고 있다. target 잡음레벨 범위보다 높은 잡음레벨을 가지는 낮은 SNR의 잡음음성에서는 마스킹 알고리즘이 적용되지 못함을 나타낸다.

Ⅲ. 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식

앞 절에서 설명한 바와 같이 target 잡음레벨을 이용한 잡음 마스킹 방식은 낮은 SNR인 경우에는 잡음음성의 잡음레벨이 target 잡음레벨보다 높기 때문에 잡음 마스킹 알고리즘이 효과적으로 적용되지 않는 경우가 있었다. 이러한 문제를 해결하기 위하여 입력된 잡음음성에 대해 음질개선의 방법을 먼저 수행한 다음 잡음 마스킹을 적용하면 입력 음성의 동적 범위를 보다 효과적으로 일치시킬 수가 있다. 따라서 높은 잡음레벨을 갖는 낮은 SNR의 잡음음성인 경우에서도 마스킹이 가능하여 인식기의 성능을 향상시킬 수가 있다.

식 (2)에 나타낸 것과 같이 현재 입력에 대해서 먼저 스펙트럼 차감법을 수행하여 음질을 개선시키고, 미리 설정한 target 잡음레벨을 기준으로 잡음 마스킹을 수행한다. 이때, 스펙트럼 차감법을 위해서는 본 논문의 선행연구에서 설명한 히스토그램 기반의 over-estimation 방식에 의해 잡음을 추정하였다[6]. 스펙트럼 차감법을 적용한 잡음 마스킹 방식은 다음 식과 같다.

$$\hat{X}(f) = \max(\max(Y(f) - \hat{N}_{OE}(f),\ \beta\,\hat{N}_{OE}(f)),\ TH) \quad (2)$$
$$\because \max(Y(f) - \hat{N}_{OE}(f),\ TH)$$

여기서 $\hat{N}_{OE}(f)$는 히스토그램 기반의 over-estimation 방식에서 추정한 잡음레벨을 나타내며, β는 flooring factor 그리고 TH는 target 잡음레벨을 각각 나타낸다. 기존의 스펙트럼 차감법에서 사용하는 flooring은 추정한 잡음을 β($\ll 1$)배 감쇄시켜 사용한다. 이는 flooring factor를 0으로 할 때 전폭왜곡의 문제를 해결하기 위한 것이며, flooring과 target 잡음레벨은 $\beta\,\hat{N}_{OE}(f) < TH$의 관계가 만족한다. 따라서 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식은 스펙트럼 차감법 이후에 target 잡음레벨 TH와의 비교만으로 처리가 된다.

이 방법은 음성의 스펙트럼이 추정된 잡음의 에너지 레벨보다 낮은 경우에 미리 설정한 target 잡음레벨로 낮은 에너지 레벨을 갖는 음성 스펙트럼 레벨을 올려주는 것이다. 이러한 처리는 대부분이 비음성 구간에서 이루어지므로 높은 에너지 레벨을 갖는 음성구간에서는 에너지 레벨이 그대로 유지되므로 잡음이 없는 음성과 잡음음성간의 동적 범위를 상당히 줄일 수가 있다.

스펙트럼 차감법에서 flooring factor에 의해 flooring된 결과는 잡음의 특성에 따라 차이를 나타내어 여전히 mismatch의 원인으로 남게 된다. 반면에 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식에서는 일정한 target 잡음레벨로 대치되기 때문에 환경 차이에 의한 영향이 크게 감소된다.

그림 2는 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식을 적용한 예를 나타낸 것이다. (a)는 그림 1의 (a)와 동일한 출력에서 스펙트럼 차감법을 수행한 결과이며, (b)는 그림 2(a)의 결과에서 잡음 마스킹을 적용한 것이다. 그림 1(b)에 비해서 그림 2(b)에서 잡음음성에 대한 잡음 마스킹이 잘 이루어짐을 볼 수 있다.

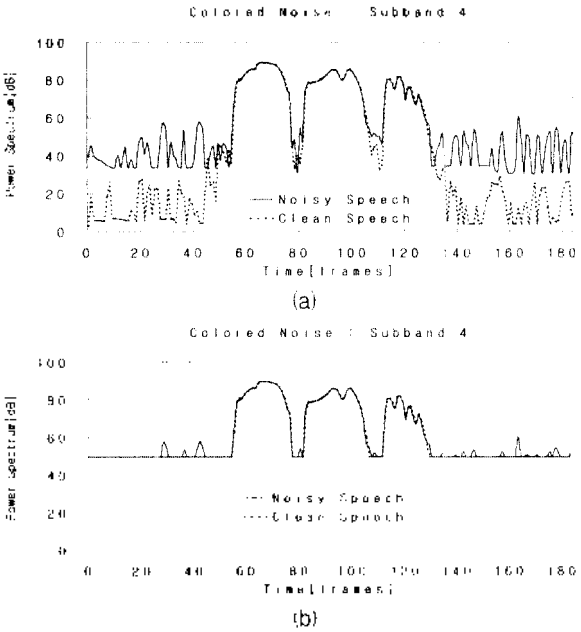

그림 2. 스펙트럼 차감법과 잡음 마스킹의 hybrid 방법의 적용 예(5dB 유색잡음)
(a) 그림 1(a)에 대한 스펙트럼 차감법 이후의 결과
(b) 스펙트럼 차감법 이후에 잡음 마스킹을 적용한 결과 (50dB의 target 잡음레벨)

Hybrid 방식을 적용함으로써 스펙트럼 차감법 이후에 남아있는 잡음 에너지들의 차이를 보다 줄일 수가 있다. 이때 target 잡음레벨보다 높은 레벨의 낮은 SNR의 잡음환경에서도 마스킹의 효과가 뚜렷하다. 그림 2의 결과는 깨끗한 음성에서의 잡음레벨과 잡음음성에서의 잡음레벨이 거의 일치하여 서로 다른 잡음환경에서 대해서도 mismatch를 대부분 줄일 수 있음을 보여주고 있다.

그리고 스펙트럼 차감법에서 flooring은 잡음의 프레임간 편차가 크기 때문에 빈번하게 되며 대부분이 음성구간에 비해 상대적으로 에너지 레벨이 낮은 잡음구간에서 이루어진다. 그러나 이러한 잡음구간에서도 간혹 스펙트럼 차감법에서 flooring 값으로 대체되지 않는 높은 에너지를 나타내는 경우가 있다. 그림 2에서 스펙트럼 차감법을 수행한 이후에도 target 잡음레벨에 의해 마스킹이 이루어지지 않는 잡음구간의 일부 피크들을 볼 수가 있다. 이러한 잔여 잡음처리를 위해서 스펙트럼 차감법 이후에 시간/주파수 영역 smoothing 처리방식을 도입하였다. 본 논문에서는 현재 프레임을 기준으로 전후의 프레임에 대해 시간이동 평균 및 메디안 smoothing을 적용하였다. 이 경우 대부분 잡음구간에서 스펙트럼 차감법 이후에 남아있는 높은 에너지 레벨의 피크들을 현저히 줄일 수가 있으며, 이후의 마스킹 처리에 대한 효율을 보다 높일 수 있다. 즉, 식 (3)과 같이 시간영역 메디안 smoothing 처리를 하는 것이다.

$$\hat{X}_t(f_k) = MED(\hat{X}_{t-1}(f_k),\ \hat{X}_t(f_k),\ \hat{X}_{t+1}(f_k)) \quad (3)$$

여기서 $\hat{X}_t(f_k)$는 k번째 필터뱅크 출력에서 t번째 프레임에서의 smoothing 결과를 나타내며 $MED(a, b, c)$는 a, b, c 의 중앙값(median)을 의미한다.

그리고 스펙트럼 영역에서 스펙트럼 차감법 이후의 결과는 스펙트럼 차감법에서 flooring factor로 대체된 스펙트럼의 경우 flooring이 없는 인접 대역에 비해서 에너지 레벨이 급격히 변화되는 결과를 나타낸다. 이러한 스펙트럼 포락선의 모양을 보상하기 위해 식 (4)와 같이 인접한 신후 대역의 스펙트럼 에너지의 값으로 메디안 smoothing을 한다. 따라서 flooring으로 인해 에너지 레벨이 현저히 떨어지는 것을 보상할 수가 있다.

$$\hat{X}_t(f_k) = MED(\hat{X}_{t-1}(f_{k-1}),\ \hat{X}_t(f_k),\ \hat{X}_{t+1}(f_{k+1})) \quad (4)$$

여기서 $\hat{X}_t(f_k)$는 t번째 프레임, k번째 필터뱅크에서 smoothing된 결과를 나타낸다.

Ⅳ. 인식실험 결과 및 고찰

본 논문에서의 음성인식 시스템은 12차의 MFCC 및 12차의 델타 MFCC 계수들을 특징 파라메터로 사용하였으며, 상용화된 HMM 인식도구인 HTK1.5를 이용하여 훈련 및 인식을 수행하였다[7]. 각 단어는 자신 및 다음 상태로만 전이를 허용하는 left-to-right 연속 HMM으로 모델링 하였으며, 상태수는 단어내의 음소당 2개로 하고 상태당 mixture의 갯수는 2개로 정하였다.

인식실험은 22개의 부서명을 대상으로 한 한국전자통신연구원의 부서명 음성 데이타베이스 중에서 고립단어 형태의 음성 데이터만을 이용하여 화자독립으로 수행하였다[8]. 22개의 부서명을 50인 각 1회 발성한 것 중에서 35명의 잡음이 섞이지 않은 음성을 모델형성을 위한 학습용으로 사용하였으며 나머지 15명의 음성을 인식대상으로 사용하여 각각의 잡음레벨에 따라 인식실험을 하였다.

표 1은 유색잡음이 부가된 잡음음성에서 각각의 SNR에 대한 인식 결과를 나타낸 것이며, 다양한 target 잡음레벨을 설정하여 마스킹을 수행한 후의 결과를 나타낸 것이다.

표 1. 잡음 마스킹에 의한 잡음 환경에서의 인식 결과

Pre-processing Technique	Noise Estimation	Target Noise Level (THI)	Accuracy[%]							
			Clear	30dB	20dB	10dB	5dB	0dB	-5dB	Average
Noise masking	NO	NO	99.4	97.0	95.2	84.5	71.8	56.7	28.8	76.2
		50dB	98.5	98.2	95.2	76.7	46.7	18.2	9.7	63.3
		55dB	95.5	82.1	77.3	46.1	28.8	16.7	11.5	51.1
		60dB	96.4	76.7	12.4	13.9	19.4	14.2	11.8	35.0
		65dB	95.8	93.3	20.6	4.9	4.9	5.5	6.4	33.1
		70dB	99.1	76.7	12.4	13.9	19.4	14.2	11.8	35.4
		75dB	96.1	96.1	81.8	10.6	5.5	3.9	4.2	42.6
		80dB	89.7	90.0	69.4	7.3	4.9	4.6	4.6	38.6

표에서 SNR이 낮아짐에 따라 인식율이 급격히 떨어지는

것을 볼 수 있다. 이는 유색잡음의 특성상 부가된 잡음의 각 필터뱅크 출력값이 대역에 따라 심한 차이를 나타내기 때문이다. 잡음 마스킹은 target 잡음레벨을 일정한 레벨로 하여 마스킹을 수행하기 때문에 target 잡음레벨보다 높게 나타나는 경우에는 마스킹이 이루어지지 않는다. 이때는 미리 설정한 target 잡음레벨을 올려주어야 한다. 이 경우에는 잡음구간의 잡음은 target 잡음레벨에 의해 대부분 마스킹이 이루어지지만, 음성구간에서 낮은 에너지 분포를 가지는 음성이 target 잡음레벨에 의해 대체됨으로써 음성 자체의 왜곡이 발생한다. 각각의 target 잡음레벨에 대해서도 낮은 SNR일수록 인식성능이 현저히 떨어지는 것을 볼 수 있다. 이는 낮은 SNR일수록 일부 대역을 제외한 대부분의 대역에서 target 잡음레벨에 의해 마스킹이 이루어지지 않는 높은 에너지 레벨을 유지하기 때문이다.

표 2는 본 논문에서 수행한 히스토그램 기반의 스펙트럼 차감법 및 잡음 마스킹을 이용한 잡음환경에서의 유색 잡음환경에 대한 인식결과를 요약한 것이다. 즉, 히스토그램 기반의 스펙트럼 차감법(SS, α=1.0), 히스토그램 기반의 over-estimation 방식(SS, γ=0.5), 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식(SS+NM), 그리고 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식에서 시간/주파수 영역 메디안 smoothing을 적용(SS+TFS+NM)한 각각의 인식결과를 clean을 포함하여 -5dB까지의 다양한 SNR에 대하여 나타낸 것이다.

표 2. 스펙트럼 차감법 및 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식을 이용한 잡음환경에서의 인식결과

Preprocessing Technique	Accuracy[%]							
	Clean	30dB	20dB	10dB	5dB	0dB	-5dB	Average
NO	99.4	97.0	95.2	84.5	71.8	56.7	28.8	76.2
SS (α=1.0)	98.8	98.8	98.3	96.7	94.2	85.2	62.1	90.6
SS (γ=0.5)	98.5	98.2	97.9	96.4	93.1	85.2	67.6	91.0
SS+NM (γ=0.1)	97.9	97.9	98.8	97.6	94.1	80.1	51.1	92.1
SS+TFS+NM (γ=0.1)	99.1	99.1	97.9	97.3	93.5	90.0	78.2	93.6

표 2에서 잡음처리를 하지 않은 경우의 76.2%의 평균 인식률에 비해 히스토그램 기반의 스펙트럼 차감법의 경우가 90.6%의 평균 인식률로 현저한 인식률 향상을 보여주고 있으며, 히스토그램 기반의 over-estimation 방식에서 91.0%의 평균 인식률로 더욱 향상된 성능을 나타내고 있다. 그리고 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식을 적용한 경우에서 92.4%의 평균 인식률로 보다 우수한 성능을 보여주고 있다. 특히 스펙트럼 차감법 후에 시간/주파수 영역 메디안 smoothing 처리를 수행한 경우에서 93.6%의 평균 인식률을 나타내어 가장 우수한 결과를 보여주고 있다.

V. 결 론

본 논문에서는 스펙트럼 차감법 적용 이후에 남아 있는 잔여 잡음으로 인한 mismatch를 극복하는 방법으로 잡음 마스킹을 사용함으로써, 잡음 마스킹이 가지는 약점을 극복하고 동시에 스펙트럼 차감법에서의 잔여 잡음 문제를 완화시킬 수 있었다. 특히 시간/주파수 영역 smoothing을 적용함으로써, 스펙트럼 차감법과 잡음 마스킹의 hybrid 방식 이후에도 어느 정도 남아 있는 잔여 잡음을 추가적으로 감소시켰으며 결과적으로 인식성능이 더욱 향상되었다.

시불변 유색잡음 환경에 대하여 HMM 기반의 화자독립 고립단어 인식실험을 하였으며, Clean 환경에서 -5dB SNR까지의 다양한 여건에서의 인식 실험결과, 기존의 히스토그램 기반의 over-estimation 방식에서의 평균 인식률이 91%였는데 반하여, 스펙트럼 차감법과 잡음레벨 마스킹의 hybrid 방식을 통해 92%의 평균인식률이 얻어졌으며, 여기에 시간/주파수 영역 smoothing 처리를 추가함으로써 최종적으로 평균인식률이 94%로 향상되었다. 이 결과는 히스토그램 기반의 스펙트럼 차감법을 적용한 경우에 비해서도 30% 정도의 인식 오류가 감소된 것이며, 잡음처리를 하지 않은 경우에 비해서는 인식오류가 73%나 감소된 결과이다.

* 본 논문에서는 한국전자통신연구원이 구축한 부서명 음성 데이터베이스의 일부를 사용하였습니다.

참 고 문 헌

[1] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, John Wiley & Sons Ltd, USA, 1996.

[2] A. Varga, R. Moore, J. Bridle, K. Ponting, and M. Russell, "Noise compensation algorithms for use with hidden Markov model based speech recognition," In Proc. IEEE ICASSP-88, pp.481-484, 1988.

[3] H. I. Jung, K. J. Shim and H. S. Kim, "Modified SNR-normalization technique for robust speech recognition," The Journal of the Acoustical Society of Korea, vol.16, no.3E, pp.14-18, Dec. 1997.

[4] T. Claes and D. Van Compernolle, "SNR-normalization for robust speech recognition," In Proc. IEEE ICASSP-96, vol.1, pp.331-334, May 1996.

[5] T. Claes, F. Xie and D. Van Compernolle, "Spectral estimation and normalisation for robust speech recognition," In Proc. Inter-national Conference on Spoken Language Processing, vol.4, pp.1997-2000, Oct. 1996.

[6] 권영욱, 김형순, "히스토그램 기반의 over-estimation을 이용한 잡음환경에서의 음성인식", 제15회 음성통신 및 신호처리 워크샵 논문집, pp.262-266, 1998년 8월.

[7] S. J. Young et al., *HTK : Hidden Markov Model Toolkit V1.5*, Entropic Research Laboratory, Inc., 1993.

[8] 이영직 외, "ETRI의 음성 데이터베이스 구축 현황",제12회 음성통신 및 신호처리 워크샵 논문집, pp.265-267, 1995년 6월.