제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

히스토그램 기반의 Over-estimation을 이용한 잡음환경에서의 음성인식

권영욱, 김형순
부산대학교 전자공학과

Speech Recognition in Noisy Environments using Histogram-based Over-estimation

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

ABSTRACT

In the speech recognition under the noisy environments, reducing the mismatch introduced between training and testing environments is an important issue, and spectral subtraction is widely used technique because of its simplicity and relatively good performance in noisy environments. In this paper, we introduced histogram method as a reliable noise estimation approach for spectral subtraction. To deal with the problem of residual noise after spectral subtraction, we proposed a new over-estimation technique based on distribution characteristics of histogram used for noise estimation. Since the proposed technique decides the degree of over-estimation adaptively according to the measured noise distribution, it can cope with the SNR variations effectively in compared with the conventional over-estimation technique.

Ⅰ. 서 론

잡음환경에서 음성인식의 성능향상을 위해 전처리과정을 통해 잡음을 제거하는 음질개선 방식들 중에서 현재 가장 널리 사용되는 대표적인 방법은 스펙트럼 차감법이다[1][2]. 이 방법은 입력된 음성에서 잡음 스펙트럼을 추정하여 이를 빼주는 것으로, 계산이 간단하고 잡음환경에서 비교적 우수한 성능을 나타내고 있다. 스펙트럼 차감법의 성공적인 적용을 위해서는 신뢰도 높은 잡음 스펙드럼 추정이 필수적인 요소이다. 이를 위해 음성/비음성 구간의 자동검출 결과에 따라 비음성구간으로 판단된 구간에서의 스펙트럼들의 평균을 잡음 스펙트럼의 추정치로 하는 것이 일반적인 방법이지만, 실제로 잡음환경에서는 음성/비음성 구간의 자동검출 자체가 신뢰도 높게 수행되기 어렵다. 특히 음성 스펙트럼의 왜곡이 심한 낮은 SNR에 대해서 잡음을 정확히 추정하는데는 어려움이 있다. 이에 따라 현실적으로 입력음성의 초기 몇 프레임을 비음성 구간으로 가정하여 잡음 스펙트럼을 추정하는 방법이 사용되지만, 이 경우에도 기반이 되는 가정이 항상 성립하지는 않으며 시변 잡음환경 대처할 수 없다는 문제점이 있다.

스펙트럼 차감법이 지니는 두 번째 문제점은 잡음 스펙트럼의 평균을 빼주기 때문에 낮은 SNR에 대해서는 스펙트럼 차감법 이후에도 잔여 잡음이 많이 남게 된다는 점이다. 이 문제의 해결을 위해 추정된 잡음레벨을 기반으로 over-estimation factor 및 flooring factor를 사용하는 방법이 도입되어 어느 정도의 성능향상이 이루어지고 있다. 그러나, 이 경우에도 over-estimation factor를 높게 할 경우에는 일부 음성부분도 손상을 받게되고, 반면에 over-estimation factor를 낮게 할 경우에는 잡음부분이 충분히 제거되지 않는 문제점이 남는다.

본 논문에서는 스펙트럼 차감법을 적용하기 위한 신뢰도 높은 잡음 스펙트럼 추정방법으로 히스토그램 처리방법을 도입하였다[3]. Hirsch에 의해 제안된 히스토그램 처리방법은 음성이 아닌 구간의 검출이 필요 없으며 SNR의 의존도가 적은 장점이 있다. 특히 이 방법은 시간적으로 서서히 변화하는 잡음의 특성에 대해서도 적용이 가능하다[4]. 그러나 히스토그램 처리방법으로 신뢰도 높은 잡음 스펙트럼의 평균값을 추정해 내더라도 스펙트럼 차감법을 적용했을 때의 잔여잡음의 문제는 여전히 남아 있다. 따라서 본 논문에서는 기존의 히스토그램 처리방법에 의한 잡음추정에서 히스토그램의 분포특성을 고려한 히스토그램 기반의 over-estimation 적용방식을 제안한다. 이는 히스토그램의 분포특성에 따

른 잡음분포의 레벨을 보다 충실히 반영함으로써 기존의 스펙트럼 차감법에 비해서 잡음음성의 SNR에 대한 영향이 적은 장점이 있다. 실제로 유색 및 자동차 소음 환경에서의 인식실험 결과 기존의 over-estimation factor를 적용한 경우보다 인식성능이 개선되었다.

Ⅱ. 히스토그램 기반의 스펙트럼 차감법

실제 환경에서의 음성신호에는 다양한 종류의 부가잡음 및 왜곡이 존재한다. 그 중에서 서로 다른 마이크 특성이나 전송선로 특성에 의한 채널왜곡을 무시한다면, 대부분의 문제는 입력음성에 더해지는 형태의 배경잡음으로 설명할 수 있다. 이 경우, 잡음이 섞인 음성신호 $y(m)$은 다음 식과 같이 표현된다.

$$y(m) = x(m) + n(m) \tag{1}$$

여기서, $x(m)$은 잡음이 섞이지 않은 원래의 음성신호이고 $n(m)$은 부가잡음이다. $n(m)$은 정적이거나 $x(m)$에 비해 매우 천천히 변화한다고 가정한다.

$Y(f)$, $X(f)$ 그리고 $N(f)$를 신호 $y(m)$, $x(m)$ 그리고 $n(m)$ 각각의 단구간 전력 스펙트럼 밀도(Power Spectral Density)라 하면, 신호와 잡음은 서로 상관성이 없으므로 다음과 같은 관계식이 성립한다.

$$Y(f) = X(f) + N(f) \tag{2}$$

여기서 f는 subband를 나타내며 $N(f)$는 부가잡음 $n(m)$의 스펙트럼 특성을 나타낸다. 식 (2)에서 잡음성분 $N(f)$를 제거하기 위해서는 잡음음성의 스펙트럼 $Y(f)$로부터 $N(f)$를 추정해야 한다. 그리고 추정된 잡음을 잡음음성에서 빼 줌으로써 잡음을 제거할 수가 있다. 일반적으로 다음 식과 같은 방법이 널리 사용된다.

$$|\hat{X}(f)| = \begin{cases} |Y(f)| - \alpha|\hat{N}(f)| & \text{if } |Y(f)| - \alpha|\hat{N}(f)| > \beta|\hat{N}(f)| \\ \beta|\hat{N}(f)| & otherwise \end{cases} \tag{3}$$

여기서 α는 over-estimation factor이고 β는 flooring factor를 나타낸다.

특히 낮은 SNR의 경우 잡음 자체의 프레임간 편차가 매우 큰 특성을 가질 때, 추정된 잡음의 평균레벨을 빼주는 것만으로는 잡음이 충분히 제거되지 않는다. 이러한 잔여 잡음을 제거하기 위해서 추정된 잡음레벨을 α배만큼 올려주는 over-estimation 처리를 한다. 그리고 스펙트럼 차감법 수행시 단순한 반파정류로 인해 발생하는 musical tone 형태의 잡음을 없애기 위하여 추정된 잡음 스펙트럼을 감쇠($\beta \ll 1$)시켜 이용하게 된다. 이때, 일부는 입력된 잡음음성을 감쇠시켜 사용하기도 한다[3].

이러한 스펙트럼 차감법은 비교적 연산이 간단하면서도 잡음환경에서 상당한 효과가 있기 때문에 잡음환경에서의 음성인식에 많이 사용된다.

이와 같은 스펙트럼 차감법에서는 먼저 신뢰성 있는 잡음을 추정해야 하며 본 논문에서는 히스토그램 처리방법을 도입하였다. Hirsch에 의해 제안된 히스토그램 처리방법은 특정 주파수 대역(subband)에서의 잡음에 대한 스펙트럼 크기의 통계적인 특성을 이용하여 잡음의 스펙트럼을 추정하는 방법으로, 다음과 같은 관찰결과에 근거를 두고 있다[4].

(1) 잡음이 포함된 잡음 음성신호의 SNR이 낮을수록 스펙트럼 크기 밀도의 분포가 스펙트럼 크기의 값이 큰 값으로 분포하게 되며, 반면에 SNR이 높을수록 잡음음성신호의 스펙트럼 크기 밀도분포는 진폭 스펙트럼의 값이 작은 값으로 분포하게 된다

(2) 잡음 음성신호의 SNR이 낮을수록 스펙트럼 크기 밀도의 분포에서 스펙트럼 크기의 분산도 큰 값으로 분포(broad distribution)한다.

히스토그램 처리방법에서는 이러한 사실들에 근거하여 각각의 주파수 대역에 대해서 스펙트럼 크기의 분포 밀도함수의 값이 최대로 되는 스펙트럼 크기를 해당 주파수 대역에서의 잡음레벨이라 판정한다[3][4].

Ⅲ. 히스토그램 기반의 over-estimation 방법

히스토그램 처리방법을 통해 잡음의 평균 스펙트럼을 보다 신뢰성 있게 추정한다고 하더라도, 잡음 평균을 빼주는 스펙트럼 차감법을 적용할 경우 잔여잡음의 문제점이 해결되지 않는다. 따라서, 식 (3)에서와 같이 적절한 over-estimation factor α를 도입하여 잡음평균 스펙트럼보다 과도하게 빼 줌으로써 잔여 잡음의 문제를 완화시키는 방법이 사용될 필요가 있다. 이때 over-estimation factor를 선정함에 있어서, 잡음이 많은 낮은 SNR을 기준으로 하면 하여 선정하게 되면 잡음 자체의 분산이 크기 때문에 over-estimation factor를 높여주어야 잔여 잡음이 줄어들게 된다. 그 대신 음성구간에서는 오히려 음성의 스펙트럼을 왜곡시키게 된다. 반면에 높은 SNR을 기준으로 하여 over-estimation factor를 낮게 선정할 경우는 낮은 SNR의 잡음구간에서 스펙트럼 차감법 이후에도 잡음이 처리되지 않고 남게 된다.

이러한 문제를 해결하기 위한 방법으로 스펙트럼 크기의 통계적인 특성의 관찰결과에 근거하여 잡음이 부가된 잡음음성에서 그림 1과 같이 각각의 주파수 대역별 스펙트럼 크기에 대하여 두 개의 가우시안으로 모델

화하고 잡음분포의 레벨을 결정하는 방식이 검토될 수 있다. 그림에서 기존의 히스토그램 기반의 스펙트럼 차감법에서는 두 개의 분포 중에서 잡음영역 가우시안 분포의 최대값 근처, 즉 히스토그램의 최빈값에 해당하는 스펙트림 크기를 추정하고자 하는 평균적인 잡음레벨로 결정한다.

스펙트럼 차감법에서 over-estimation을 적용할 때 잡음과 음성의 분포특성을 기반으로 하여 다음과 같은 검토가 가능하다. Over-estimation을 히스토그램 및 가우시안 분포에서 잡음과 음성의 경계(threshold)보다 낮게 설정하게 되면 스펙트럼 차감법에서 제거되지 않는 잔여 잡음이 많게 되며, 반면에 경계보다 높게 설정할 경우에는 스펙트럼 차감법에서 잔여 잡음은 대부분 제거되나 음성의 일부가 함께 제거되어 왜곡이 발생하게 된다. 따라서 음성의 왜곡을 최소화하는 범위에서 잡음을 최대한으로 제거하기 위해서는 두 개의 가우시안 분포가 만나는 경계에서 over-estimation을 결정하는 것이 바람직하다고 할 수 있겠다.

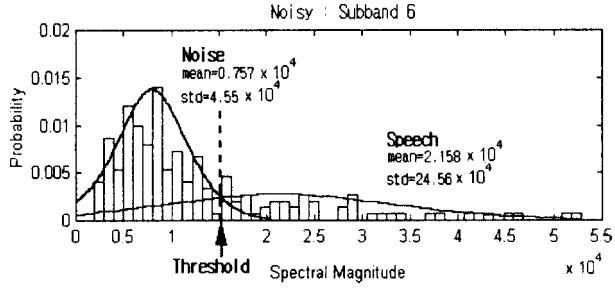

그림 1. 두 개의 가우시안 분포로 모델링한 5dB SNR의 잡음음성에 대한 히스토그램

이와 같이 잡음과 음성의 두 가우시안 분포특성에서 경계를 이용한 over-estimation 결정방식은 입력음성의 SNR에 관계없이 잡음이 분포하는 정도를 보다 충실히 반영하는 잡음레벨을 결정할 수가 있다. 그러나 이 방법은 가우시안 분포의 모델에 따른 처리시간이 많이 소요되어 실시간 처리의 구현에는 문제가 있다.

따라서 본 논문에서는 그 대신에 히스토그램의 최대값의 γ배인 $\gamma \cdot H_{max}$에 해당하는 스펙트럼 크기를 그 대역에서의 잡음레벨로 결정하는 히스토그램 기반의 over- estimation 방식을 제안한다. 이와 같이 잡음의 히스토그램 분포특성을 고려하여 over-estimation의 정도를 결정하게 되면, 잡음음성의 SNR이 달라지더라도 잡음과 음성분포의 분산에 의한 경계를 비교적 잘 표현할 수 있다.

그림 2에서 히스토그램의 최대값 H_{max}에 해당하는 스펙트럼 크기의 최빈값 $\hat{N}(f)$가 기존의 히스토그램 처리방법에서 결정하는 잡음레벨을 나타낸다. 그리고 본 논문에서 제안한 개선된 잡음레벨 추정방식은 그림에서 히스토그램의 최대값의 우측의 $\gamma \cdot H_{max}$에 해당하는 스펙트럼 크기 $\hat{N}_{OE}(f)$를 추정하고자 하는 잡음레벨로 판정하는 것이다. 이때 γ는 0과 1 사이의 값을 가지며, γ값이 작아질수록 over-estimation을 많이 해 주는 효과를 가진다. 이 γ를 본 논문에서는 히스토그램 기반의 over-estimation 방법에서의 over-estimation factor라 부르기로 한다.

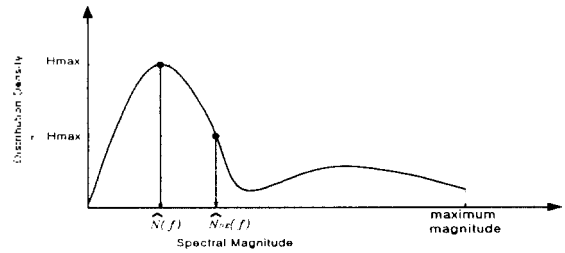

그림 2. 히스토그램 기반의 over-estimation 방식

히스토그램 기반의 over-estimation 방식에 의한 스펙트럼 차감법은 히스토그램 기반의 스펙트럼 차감법에서 over-estimation factor α를 사용하는 대신에 다음 식과 같이 표현될 수 있다.

$$\hat{X}(f) = \begin{cases} Y(f) - \hat{N}_{OE}(f) & \text{if } Y(f) - \hat{N}_{OE}(f) > \beta \cdot \hat{N}_{OE}(f) \\ \beta \cdot \hat{N}_{OE}(f) & otherwise \end{cases} \quad (4)$$

여기서 $\hat{N}_{OE}(f)$는 히스토그램 기반의 over-estimation 방식에서 추정한 잡음레벨을 나타내며, 기존의 스펙트림 차감법에서의 $\alpha \cdot \hat{N}(f)$에 대응되는 값이다. 그리고 β는 flooring factor를 나타내는 것으로 기존의 스펙트럼 차감법에서 사용한 것과 동일하다.

그림 3은 히스토그램 기반의 over-estimation에 따른 잡음레벨 추정치와 기존의 잡음레벨 추정치의 비교를 나타내었다. 이 그림은 잡음음성의 8번째 필터뱅크 출력에 대한 잡음레벨을 나타낸 것이다. 기존의 스펙트림 차감법에서 over-estimation factor α가 1.3일 때와, 본 논문에서 제안하는 히스토그램 기반의 over-estimation 방식에서의 over-estimation factor γ를 0.3으로 적용한 경우를 함께 나타낸 것이다. 이때, 기존의 방식과 제안된 방식에서의 over-estimation factor의 결정은 각각의 방법에서 구한 필터뱅크 출력의 잡음레벨에서 대수영역

평균이 동일하도록 조정한 것이다.

그림 3에서 기존의 히스토그램 처리방법에서 보다 히스토그램 기반의 over-estimation 방식이 잡음의 영역에서 peak에 가까운 레벨을 잘 나타내고 있음을 보여준다. 반면에 음성구간에서는 오히려 추정한 잡음레벨이 떨어지는 것을 볼 수 있다. 이는 스펙트럼 차감법의 적용 후에 배경잡음의 영역에서는 가능한 한 모든 잡음이 제거되어야 하며, 음성구간에서는 음성의 에너지들을 그대로 보존시키는 것이 음성인식의 성능을 향상시킬 수가 있다는 점에서 바람직한 결과이다.

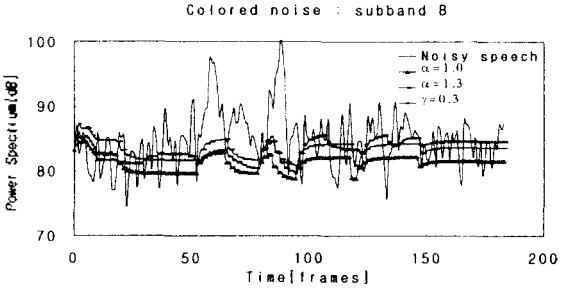

그림 3. 히스토그램 기반의 over-estimation에 따른 잡음 레벨 추정치와 기존의 잡음레벨 추정치의 비교

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[5]. 각 단어는 자신 및 다음 상태로만 천이를 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 데이터베이스 중에서 고립단어 형태의 음성 데이터만을 이용하여 화자독립으로 수행하였다[6]. 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.

기존의 초기 프레임 평균방법에 비해 히스토그램 처리방법에 의한 스펙트럼 차감법이 보다 우수함을 본 논문의 선행연구에서 확인한 바 있다[4]. 이를 기반으로 본 논문에서는 3장에서 설명한 잡음 추정방식에서 히스토그램 및 히스토그램 기반의 over-estimation 방식의 각각을 적용한 스펙트럼 차감법을 수행하고, 이에 대한 인식실험을 수행하였다. 각각의 방식에서 over-estimation factor에 대한 처리는 추정한 잡음레벨을 이용하여 스펙트럼 차감법으로 잡음처리를 할 때, 기존의 히스토그램 처리방법에서는 over- estimation factor를 실험적으로 일정한 상수로 결정하였다. 반면에 제안된 히스토그램 기반의 over-estimation 방식은 스펙트럼 차감법에서 일정한 over-estimation factor를 사용하는 대신에 히스토그램의 분포특성에서 분포의 최대값의 γ배에 해당하는 스펙트럼의 크기를 잡음레벨의 추정치로 결정하게 된다. 본 논문에서 제안한 이러한 히스토그램 기반의 over-estimation 방식에 따른 인식실험 결과를 기존의 방식과 비교하기 위하여 표에 함께 나타내었다.

표 1은 자동차 소음이 부가된 잡음음성에서 기존의 히스토그램 처리방법 및 히스토그램 기반의 over-estimation 방식에서의 인식실험 결과를 나타낸 것이다. 표에서 기존의 히스토그램 처리방법에 의한 결과는 over-estimation factor α가 1.5인 경우에서 다른 α값에 비해 우수한 성능을 나타내고 있으며, 히스토그램 기반의 over-estimation 방식에서는 히스토그램의 최대값의 30%에 해당하는 γ가 0.3일 때의 잡음레벨에서 가장 우수한 성능을 내고 있으며, 낮은 SNR일수록 인식률의 개선이 현저하다.

표 1. 기존의 over-estimation 및 히스토그램 기반의 over-estimation 방식을 이용한 자동차 소음환경에서의 인식결과 비교

Over-estimation Technique	Over-estimation factor	Accuracy[%]						
		Clean	30dB	20dB	10dB	5dB	0dB	Average
NO	NO	99.4	98.4	97.6	93.9	85.8	67.8	90.5
Standard Over-estimation	$\alpha=1.0$	98.5	98.2	97.6	95.8	92.4	83.3	94.3
	$\alpha=1.2$	98.5	98.2	98.5	95.8	91.2	81.2	93.9
	$\alpha=1.5$	98.5	98.5	97.5	98.8	97.9	85.8	96.2
	$\alpha=2.0$	98.2	97.8	97.0	87.6	72.4	52.4	84.2
Histogram-based Over-estimation	$\gamma=0.7$	98.5	98.5	98.5	98.8	97.6	92.4	97.4
	$\gamma=0.5$	98.5	98.2	98.5	98.2	97.6	91.2	97.0
	$\gamma=0.3$	98.8	98.5	98.2	98.2	97.3	93.9	97.5
	$\gamma=0.1$	98.5	98.2	97.9	97.0	93.0	80.6	94.2

표 1에서 히스토그램 기반의 over-estimation 방식에서는 γ가 0.3일 때가 Clean 환경에서 0dB SNR까지의 평균 인식률이 97.5%로 가장 우수하며, 이는 기존 방식에서 가장 우수한 α가 1.5에서의 96.2%보다도 우수한 결과이다. 특히 히스토그램 기반의 over-estimation 방식에서는 $\gamma=0.5$ 및 $\gamma=0.7$의 경우에도 기존 방식의 가장 우수한 경우보다 평균 인식률이 높아 γ값의 변화에

따른 민감도는 크지 않은 것으로 판단된다. 다만 히스토그램 최대값의 10%인 γ가 0.1인 경우는 평균 인식률이 94.2%로 떨어지는 결과를 보인다. 이는 상대적으로 과도한 over-estimation을 사용하기 때문에 잡음구간의 잡음은 대부분 제거되는 반면, 음성구간에서 음성의 왜곡이 크게 발생하기 때문이다.

표 1의 결과에서 잡음의 레벨을 보다 신뢰성 있게 추정하는 히스토그램 기반의 over-estimation 방식을 적용한 스펙트럼 차감법이 기존의 over-estimation에 비해 인식성능이 향상됨을 볼 수 있다. 특히, 잡음의 영향이 심할수록 over-estimation에 비해서 히스토그램 기반의 over-estimation 방식에서 보다 향상된 인식성능을 얻을 수 있었다.

V. 결 론

본 논문에서는 스펙트럼 차감법을 적용하기 위한 신뢰도 높은 잡음 추정방법으로 히스토그램 처리방법을 도입하고, 스펙트럼 추정에 사용된 히스토그램의 분포 특성을 고려한 새로운 over-estimation 방식을 제안하였다. 제안된 방법은 over-estimation을 얼마만큼 적용할 것인가 하는 것을 잡음분포의 특성에 따라 적응적으로 결정함으로써, 기존의 경직된 over-estimation 방식에 비해 SNR 변화에 효과적으로 대처할 수 있는 장점을 가진다.

본 논문에서는 제안된 방식들의 성능 평가를 위하여 실제 자동차 소음환경에 대하여 HMM 기반의 화자독립 고립단어 인식실험을 수행하였다. Clean 환경에서 0dB SNR까지의 다양한 여건에서의 인식 실험결과, 기존의 over-estimation 방식에서 가장 우수한 경우의 평균 인식률이 96.2%였는데 반하여, 히스토그램 기반의 over-estimation 방식을 도입함으로써 평균 인식률이 97.5%로 개선되었다. 특히 SNR이 0dB일 경우 기존 방식의 인식률이 최고 85.8%인데 반하여 제안된 방식의 최고 인식률은 93.9%로서 인식오류의 57%가 줄어든 효과를 얻었다.

---

* 본 논문에서는 한국전자통신연구원이 구축한 무서명 음성 데이터베이스의 일부를 사용하였습니다.

참 고 문 헌

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans., Acoust., Speech Signal Processing, Vol. ASSP-27, No. 2, pp.113-120, April 1979.

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

[3] H. G. Hirsch and C. Ehricher, "Noise estimation techiques for robust speech recognition", in Proc. IEEE ICASSP-95, pp.153-156, May 1995.

[4] 권영욱, 김형순, "히스토그램 처리방법에 의한 잡음 스펙트럼 추정을 이용한 잡음환경에서의 음성인식", 한국음향학회논문집, 제16권 5호, pp.68-75, 1997년 7월.

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

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