히스토그램 처리방법에 의한 잡음 스펙트럼 추정을 이용한 잡음환경에서의 음성인식

Speech Recognition in Noisy Environments using the Noise Spectrum Estimation based on the Histogram Technique

권 영 욱*, 김 형 순*
(Young-Uk Kwon*, Hyung-Soon Kim*)

요 약

스펙트럼 차감법은 잡음이 더해진 환경에서의 음성인식에 널리 사용되는 전처리 방법이지만, 이를 위해서는 잡음의 스펙트럼을 잘 추정할 필요가 있다. 본 논문에서는 잡음 스펙트럼의 추정방법으로 히스토그램 처리방법을 사용한다. 이 방법은 음성/비음성 구간의 구분을 할 필요가 없으며 서서히 변화하는 잡음의 스펙트럼도 추정할 수 있다는 점에서 여타의 잡음 추정방법에 비해 장점을 지닌다. 다양한 SNR 조건하에서 유색 가우시안 잡음 및 실제 자동차 소음을 부가시킨 음성에 대해 화자독립 고립단어 인식 실험을 수행한 결과, 히스토그램 처리방법에 기반을 둔 스펙트럼 차감법의 인식성능이 초기 비음성구간의 스펙트럼 평균을 이용한 기존의 잡음 스펙트럼 추정방법에 비해 우수한 성능을 나타내었다.

ABSTRACT

Spectral subtraction is widely-used preprocessing technique for speech recognition in additive noise environments, but it requires a good estimate of the noise power spectrum. In this paper, we employ the histogram technique for the estimation of noise spectrum. This technique has advantages over other noise estimation methods in that it does not requires speech/non-speech detection and can estimate slowly-varying noise spectra. According to the speaker-independent isolated word recognition in both colored Gaussian and car noise environments under various SNR conditions. Histogram-technique-based spectral subtraction method yields superier performence to the one with conventional noise estimation method using the spectral average of initial frames during non-speech period.

I. 서 론

잡음환경에서 음성인식 시스템의 성능을 향상시키는 것은 음성인식의 실용화를 위해 매우 중요한 과제이다. 실제로 잡음이 없는 여건에서는 매우 우수한 성능을 나타내는 음성인식 시스템들이 잡음환경에서는 급격한 성능저하를 초래한다.[1] 이는 훈련과정 및 인식과정사이의 환경차이에 기인한 것으로서, 인식환경과 동일한 잡음환경에서 훈련과정을 수행할 수 있다면 어느 정도 극복이 가능하지만 인식시의 환경을 미리 예측할 수 없으므로 이 방법은 현실적이지 못하다. 이에 따라 잡음환경에서 음성인식 시스템의 성능을 향상시키기 위한 다양한 방법들이 개발되어 왔다. 잡음환경에서의 음성인식을 위한 처리방식으로는 전처리과정을 통해 잡음을 제거하는 음질개선(speech enhancement) 방식과 청각기관의 모델 등에 근거한 잡음에 강인한 음성특징 추출방식, 잡음에 강인한 거리측정법, 그리고 이미 만들어진 모델을 토대로 환경에 따라 파라메터를 보상하는 모델적응화 기법 등이 있다.[1-2] 이들 방식 중에서 음질개선방식은 음성인식 시스템의 전처리과정에서 잡음을 제거함으로 기존의 인식 시스템의 구조를 변화시키지 않고 처리 가능하다는 장점이 있다. 음질개선 방식의 구체적인 방법으로는 스펙트럼 차감법(spectral subtraction), Wiener 필터링 그리고 MMSE(Minimum Mean Square Error estimation) 방법 등을 들 수 있으며, 그 중에서도 스펙트럼 차감법이 대표적인 방법으로 알려져 있다.[3-5]

스펙트럼 차감법은 잡음이 섞인 음성의 스펙트럼으로

*부산대학교 전자공학과
접수일자: 1997년 4월 1일

부터 미리 추정한 잡음 스펙트럼을 빼 줌으로써 잡음을 제거하는 방법이다. 이를 위해서는 잡음의 스펙트럼을 추정할 필요가 있으며, 음성/비음성 구간의 자동검출 결과에 따라 비음성구간으로 판단된 구간에서의 스펙트럼들의 평균을 잡음 스펙트럼의 추정치로 간주한다. 그러나, 실제로 잡음환경에서는 음성/비음성 구간의 자동검출 자체가 신뢰도 높게 수행되기 어렵기 때문에 입력음성의 초기 몇 프레임을 비음성 구간으로 가정하여 잡음 스펙트럼을 추정하는 방법이 종종 사용된다. Hirsch는 음성/비음성 구간의 구분없이 잡음 스펙트럼을 효과적으로 추정하는 히스토그램 처리 방법을 제안하였다.[6-7] 본 논문에서는 잡음환경에 강인하다고 알려진 인지선형예측(Perceptual Linear Prediction(PLP)) 분석방법[8] 및 스펙트럼 차감법에 기반을 둔 잡음환경에서의 음성인식 시스템을 구현하고 잡음 스펙트럼의 추정방법으로서 기존의 방법과 히스토그램 처리 방법의 성능을 비교하였다. 모의 발생된 유색 가우시안 잡음과 실제 자동차 주행잡음을 대상으로 한 화자독립 음성인식 실험 결과, 히스토그램 처리방법이 기존의 비음성구간의 스펙트럼 평균치에 의한 방법보다 인식성능면에서 우수함을 확인하였다.

Ⅱ. 잡음환경에서의 전처리과정

2.1 Perceptually Linear Prediction(PLP)

인지선형예측(PLP) 분석방법은 기존에 음성인식에 널리 사용되어 온 선형예측 방법의 변형된 형태로서, 심리음향학적 지식에 기반을 둔 청각 스펙트럼 특성을 all-pole 모델로 표현한 것이다. PLP 방법의 처리과정은 그림 1과 같으며, 기존의 선형예측 방법과의 차이점은 다음과 같다.

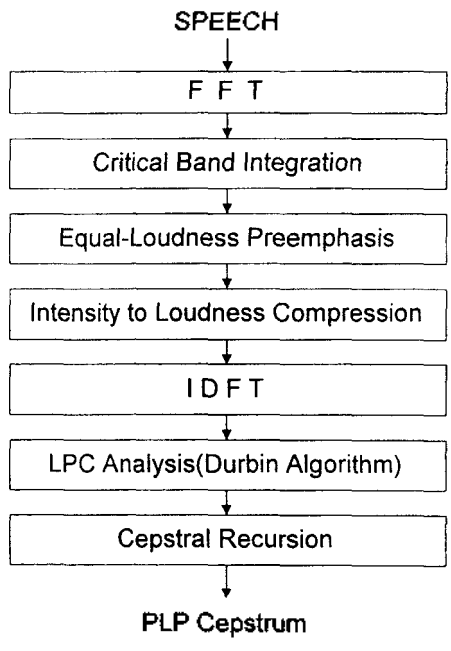


그림 1. PLP 처리의 블럭도
Fig. 1. Block diagram of PLP processing.

1) Critical-band 필터특성에 의한 스펙트럼 분석
2) Equal-loudness curve 특성을 반영한 preemphasis
3) Cubic root 형태의 intensity-loudness 변환

청각특성의 스펙트럼을 얻기 위해서 17개의 critical-band 대역필터의 출력들이 사용되는데, 이들 각각의 중심주파수는 bark 영역에서 등간격으로 구성된다. Brak 단위의 주파수[z]와 Hz 단위의 주파수[f] 사이의 변환식은 다음과 같이 표현된다.

$$z = 6 \ln\left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right] \quad (1)$$

k번째 critical-band 필터의 중심주파수는 $z_k = 0.9994k$로 주어진다. 각각의 critical-band 필터는 20msec 크기의 Hamming 창을 씌운 다음 FFT를 수행하여 구한 단구간 전력 스펙트럼 $P(w)$로 부터 가중합산 과정을 통해 구현될 수 있다. 이 때 k번째 critical-band 필터에 사용되는 가중함수 $C_k(w)$는 다음과 같이 주어진다.

$$C_k(w) = \begin{cases} 10^{1.0(z - z_k + 0.5)} & \text{for } z \le z_k - 0.5 \\ 1 & \text{for } z_k - 0.5 < z < z_k + 0.5 \\ 10^{-2.5(z - z_k - 0.5)} & \text{for } z \ge z_k + 0.5 \end{cases} \quad (2)$$

이들 필터들은 중심주파수보다 낮은 주파수에서는 +10dB/bark, 그리고 중심주파수보다 높은 주파수에서는 −25dB/bark의 기울기를 가지는 비대칭 형태이다. 또한, Equal-loudness 곡선은 다음 식과 같이 근사화 될 수 있다.

$$\mathrm{E}(w) = 1.151 \sqrt{\frac{(w^2 + 144 \times 10^4)\, w^2}{(w^2 + 16 \times 10^4)\,(w^2 + 961 \times 10^4)}} \quad (3)$$

따라서, k번째 critical-band 대역필터의 equal-loudness 가중 출력은 다음과 같이 얻을 수가 있다.

$$\mathrm{F}_k = \mathrm{E}(w_k) \int_0^{\pi} C_k(w)\, P(w)\, dw, \quad k = 1, 2, \cdots, 17 \quad (4)$$

심리음향적 처리의 마지막 단계인 intensity-loudness[8] 변환과정은 다음과 같다.

$$\mathrm{Q}(w_k) = [\mathrm{F}_k]^{1/3}, \qquad k = 1, 2, \cdots, 17 \quad (5)$$

상기의 모든 처리과정의 결과로 청각특성 스펙트럼의 이산 표현이 얻어지며, 이 결과를 inverse DFT(IDFT) 하면 청각특성이 반영된 자기상관함수가 얻어진다. 그 다음 과정은 일반적인 선형예측 분석방법과 동일하다. 이들은 all-pole 모델 함수로 주어진다.

2.2 잡음환경의 모델과 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 배경잡음으

로 설명할 수 있다. 이 경우, 잡음이 섞인 음성신호 $y(m)$은 다음과 같이 표현된다.

$$y(m) = x(m) + n(m) \tag{6}$$

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

$Y(w)$, $X(w)$ 그리고 $N(w)$를 신호 $y(m)$, $x(m)$ 그리고 $n(m)$ 각각의 단구간 전력스펙트럼 밀도(Power Spectral Density)라 하면, 신호와 잡음은 서로 상관성이 없으므로 각각의 주파수 대역 w_k에 대해 다음과 같은 관계식이 성립한다.

$$Y(\omega_k) = X(\omega_k) + N(\omega_k) \tag{7}$$

여기서 w_k는 k번째의 subband를 나타낸다.

스펙트럼 차감법은 잡음이 섞인 신호에서 잡음을 억제하거나 제거하는 목적으로 사용되는 기법들 중의 하나이다. 원래 이 방법은 잡음환경에서 통화품질의 개선을 목적으로 처음 제안되었지만, 최근에는 음성인식의 응용에 많이 사용되고 있다. 이 방법은 음성 스펙트럼을 구하기 위해 각각의 주파수 대역에서 잡음음성의 에너지로부터 추정된 잡음의 에너지를 빼 주는 기법이다.

특정 프레임에서 잡음음성의 전력스펙트럼밀도 $\hat{Y}(w)$가 구해지고 잡음의 전력 스펙트럼밀도 $\hat{N}(w)$가 추정되면, 식 (7)로부터 음성의 전력스펙트럼 밀도는 다음과 같이 추정할 수 있다.

$$\hat{X}(w_k) = \hat{Y}(w_k) - \hat{N}(w_k) \tag{8}$$

전력 스펙트럼은 음의 값을 가질 수 없으므로 만약 이러한 추정값 $\hat{X}(w_k)$의 결과가 음이 될 경우 식 (8)은 다음과 같이 반파정류를 수행하는 것으로 수정하여 처리한다.

$$\hat{X}(w_k) = \max[\hat{Y}(w_k) - \hat{N}(w_k), 0] \tag{9}$$

이러한 반파정류의 결과로 musical tone 형태의 잡음이 발생하는 문제가 있는데, 이를 보완하기 위해서 다음과 같은 변형 방법이 널리 사용된다.

$$\hat{X}(w_k) = \begin{cases} \hat{Y}(w_k) - \alpha \cdot \hat{N}(w_k) & \text{if } \hat{Y}(w_k) - \alpha \cdot \hat{N}(w_k) > \beta \cdot \hat{N}(w_k) \\ \beta \cdot \hat{N}(w_k) & \textit{otherwise} \end{cases} \tag{10}$$

여기서 w_k는 k번째 band의 주파수, α는 over-estimation factor 그리고 β는 flooring factor를 나타낸다. 이와 같은 스펙트럼 차감법은 비교적 연산이 간단하면서도 잡음환경에서 상당한 효과가 있기 때문에 잡음환경에서의 음성인식에 많이 사용된다. 스펙트럼 차감법에서의 관건은 잡음 스펙트럼의 추정치를 어떻게 구하는가 하는 것이다. 대부분 잡음의 추정시 음성/비음성 구간을 먼저 구분하여 비음성 구간의 프레임을 평균하여 부가잡음의 추정치로 하게 되는데, 이러한 경우에는 음성/비음성의 판단의 정확도에 크게 의존하게 되며, 특히 음성 스펙트럼의 왜곡이 심한 낮은 SNR에 대해서 잡음을 정확히 추정하는데는 어려움이 있다. 일반적으로 입력음성 초기의 몇 프레임을 비음성구간이라 가정하고 이들 구간에서 잡음을 추정하는 방법이 사용되고 있으나, 이 가정이 항상 성립되지는 않으며, 또한 잡음의 특성이 시간적으로 변화할 때 이에 대한 대체가 불가능하다. 이러한 문제의 해결을 위해 음성/비음성 구간을 검출하는 대신에 과거의 복수개의 프레임으로부터 구한 단구간 스펙트럼의 이동평균(moving average)을 추정된 잡음레벨로 하는 연속 스펙트럼 차감법(Continuous Spectral Subtraction)이 제안되었다.[4] 그러나, 이 방법의 경우에도 추정된 잡음이 앞선 음성 프레임의 스펙트럼의 영향을 받게 되어 결과적으로 음성스펙트럼의 왜곡이 발생하는 문제가 있다.

Ⅲ. 히스토그램 처리에 의한 잡음 스펙트럼 추정

3.1 개 요

Hirsch에 의해 제안된 히스토그램 처리방법은 특정 주파수 대역(subband)에서의 잡음에 대한 진폭 스펙트럼의 통계적인 특성, 즉 분포밀도 함수(Distribute Density Function)를 이용하여 잡음의 스펙트럼을 추정하는 방법으로, 다음과 같은 관찰결과에 근거를 두고 있다.[6-7]

(1) 잡음이 포함된 잡음음성 신호의 SNR이 낮을수록 진폭 스펙트럼 밀도의 분포가 진폭 스펙트럼의 값이 큰 값으로 분포하게 되며, 반면에 SNR이 높을수록 잡음음성 신호의 진폭 스펙트럼 밀도분포는 진폭 스펙트럼의 값이 작은 값으로 분포하게 된다.

(2) 잡음음성 신호의 SNR이 낮을수록 진폭 스펙트럼 밀도의 분포에서 진폭 스펙트럼의 분산도 큰 값으로 분포(broad distribution)한다.

히스토그램 처리방법에서는 이러한 사실들에 근거하여 각각의 주파수 대역에 대해서 진폭 스펙트럼의 분포밀도 함수의 값이 최대로 되는 진폭 스펙트럼을 해당 주파수 대역에서의 잡음레벨이라 판정한다. 따라서, 이 방법의 경우 음성이 아닌 구간의 검출이 필요 없으며 또한 SNR의 의존도가 적은 장점이 있다. 특히 이 방법은 시간적으로 변하는 잡음의 특성에 대해서도 적용이 가능하다.

실제의 단어음성 "재무관리실"에 대해서 각 잡음의 레벨에 따른 스펙트럼 분포 및 분포밀도 함수를 그림 2에 나타내었다. 그림 2는 입력음성에 대해 critical-band 대역필터군에 의해 주파수 분석을 하고, 그 중에서 6번째 필터(1.6kHz 부근)에 대한 결과이다. 그림 2(a)는 깨끗한 음

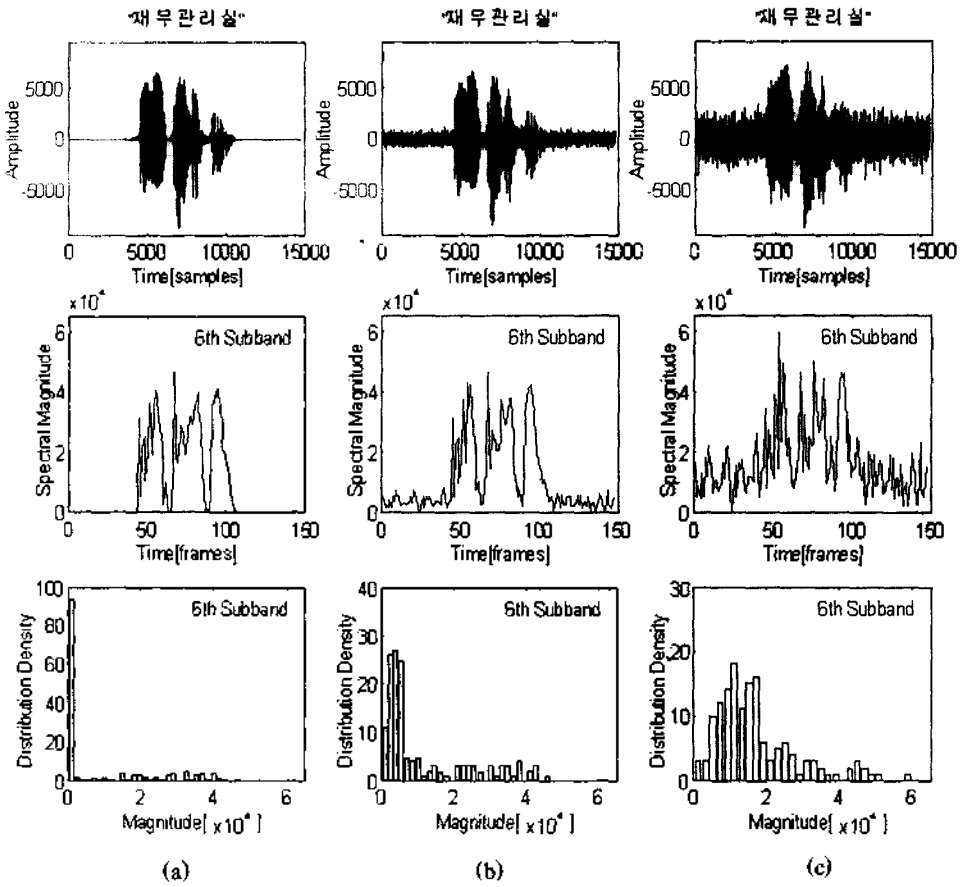


그림 2. SNR 변화에 따른 음성신호, 진폭 스펙트럼 및 분포밀도 함수
(a) 원래의 음성
(b) 10 dB SNR의 잡음음성
(c) 0 dB SNR의 잡음음성

Fig. 2. Speech signal, magnitude spectrum and distribution density function according to the change of SNR,
(a) Original speech,
(b) Noisy speech of the 10 dB SNR,
(c) Noisy speech of the 0 dB SNR.

성(clean speech), 그림 2(b)는 10dB SNR에서의 잡음음성 스펙트럼을 나타내며, 그리고 그림 2(c)는 0dB SNR에서의 잡음음성 스펙트럼을 나타낸다. 이들 세 경우에 대한 분포밀도 함수를 비교해 볼 때, 깨끗한 음성에서의 진폭 스펙트럼의 분포는 0 부근에 대부분이 모여 있으며 10dB 잡음음성의 진폭 스펙트럼 분포의 값은 깨끗한 음성에서보다 큰 값으로 분포하고 있으며, 또한 0dB 잡음음성의 진폭 스펙트럼 분포는 10dB 잡음음성에서보다 더 큰 값으로 분포하는 것을 쉽게 알 수 있다. 즉 낮은 SNR일수록 진폭 스펙트럼의 값이 큰 쪽으로 이동하며 분포의 피크 위치가 그 주파수 대역에서의 평균적인 잡음레벨을 표현하게 된다. 그림 2(a)-(c)로부터 SNR이 낮아짐에 따라 분포밀도 함수의 분산도 커짐을 알 수 있다.

3.2 히스토그램 처리방법의 구현

앞서 설명한 히스토그램 처리방법에서는 입력 음성구간 전체에 대해 각 band별로 진폭 스펙트럼의 분포밀도 함수, 즉 히스토그램을 구해서 잡음 스펙트럼을 추정하는 것을 가정하였다. 그러나, 히스토그램 처리방법의 실시간 구현을 위해서는 입력음성이 다 들어올 때까지 기다릴 수 없으며 프레임 단위로 처리하는 것이 필수적으로 요청된다. 따라서 현재 프레임을 기준으로 하여 그 이전의 일정 구간에 대한 분포밀도 함수를 구하는 방법이 사용된다. 히스토그램 분석기법의 특성상 분석구간이 긴 것이 바람직 하나, 분석구간을 길게 하면 처리시간이 많이 소요되고 시간적으로 변화하는 잡음특성을 표현할 수 없는 문제점이 있으므로, 적절한 분석구간 크기가 요구된다.

히스토그램 처리방법에 의해 잡음 스펙트럼을 추정하는 방법을 요약하면 다음과 같다. 먼저 매 프레임 단위로 입력음성을 FFT하여 주파수 영역으로 변환한 다음, 청각기관의 특성을 고려한 critical-band scale로 warping 한

다. 본 논문에서는 8 kHz의 샘플링 주파수를 가지는 입력 음성을 대상으로 256-point FFT를 수행하였다. 입력음성의 주파수 범위가 0에서 4kHz 사이이므로, 식 (1)에 의해 15개의 critical-band 출력을 얻게 된다. 이들 각 band에 대해 개별적으로 히스토그램 처리과정을 수행하여 그 주파수 band에 해당하는 잡음레벨을 추정하게 되며, 그 결과를 종합하면 15개의 주파수 band 분석에 따른 잡음 스펙트럼 특성을 구할 수 있다.

이하에 각 band 별로 잡음레벨을 추정하는 과정을 설명한다. 앞서 설명한 히스토그램 분석구간(현재 프레임을 기준으로 그 이전의 일정 수의 프레임들)에 대해 각 프레임별로 구한 해당 주파수 band의 진폭 스펙트럼들을 이용하여 그림 2의 맨 아래 그림과 같은 분포밀도 함수, 즉, 히스토그램을 구한다. 이 때, 히스토그램 분석의 분해능(resolution)이 M이라 하면, 0에서 진폭 스펙트럼의 최대값 사이를 M등분 하여 각각의 돗수값을 누적시켜 히스토그램을 구한다. 이렇게 구해진 히스토그램에서 돗수값이 가장 큰 진폭 스펙트럼, 즉, 최빈값을 해당 주파수 band의 잡음레벨의 추정치로 정한다. 히스토그램 분석과정의 분해능 크기는 추정된 잡음 스펙트럼의 정밀도와 밀접한 관계가 있으며, 히스토그램 분석구간의 길이(프레임 수)를 고려하여 선정할 필요가 있다. 본 논문에서는 이들 분석구간 길이 및 분해능 크기 값을 인식 성능결과에 따라 실험적으로 결정하였다.

대부분의 음성 프레임들에 대해 히스토그램 분석 방법이 잡음 레벨을 매우 신뢰도 높게 추정하지만, 일부 프레임들에 대해서는 엉뚱한 추정값을 구하게 되는 경우가 있다. 이러한 문제점을 해결하기 위해 다음 두 가지 보상방법이 검토되었다. 첫 번째로 잡음의 특성이 프레임 단위로 급격히 변화하지 않는다는 가정하에 히스토그램 처리방법에 의해 구해진 잡음레벨을 그 이전 프레임들에 대해 구해진 잡음레벨들을 이용하여 smoothing 하도록 하였다. 두 번째로 잡음신호의 히스토그램이 일반적으로 Rayleigh 분포에 가까운 형태를 가지는데 반하여 특정한 음성파형이 지속될 경우 높은 레벨에서 잘못된 피크(peak)가 형성될 수 있는 문제를 해결하기 위해, 히스토그램 상의 최빈값이 평균값보다 클 경우 잡음레벨의 추정치를 평균값으로 대치하는 방법을 적용하였다.

3.3 히스토그램 처리방법을 이용한 부가잡음의 추정실험

본 논문에서의 부가잡음 추정은 음성특징 분석방법으로 2.1절에서 설명한 PLP 방법을 사용함을 전제로 하였다. 입력 음성의 샘플링 주파수는 8kHz이고, 20msec의 분석구간으로 하여 10msec씩 이동하면서 특징을 추출하였다. 그리고 잡음에 대한 모의 실험을 위해 컴퓨터에서 발생시킨 잡음을 각 SNR에 따라 입력 음성에 더하고, 이를 잡음음성에서 잡음을 추정하는 시스템의 입력으로 사용하였다. 본 실험에서 1차적으로 사용한 잡음은 백색 가우시안 잡음을 900Hz에서 1.647kHz를 통과대역으로 하는 대역필터를 거치게 하여 만든 유색잡음 형태이다. 다음 장에서 다룰 인식실험에서는 유색 가우시안 잡음 이외에도 주행중인 자동차환경 잡음 및 백색 가우시안 잡음 자체에 대한 잡음음성에 대해서도 잡음 추정기법에 의한 스펙트럼 차감법을 수행하고 인식실험을 하였다.

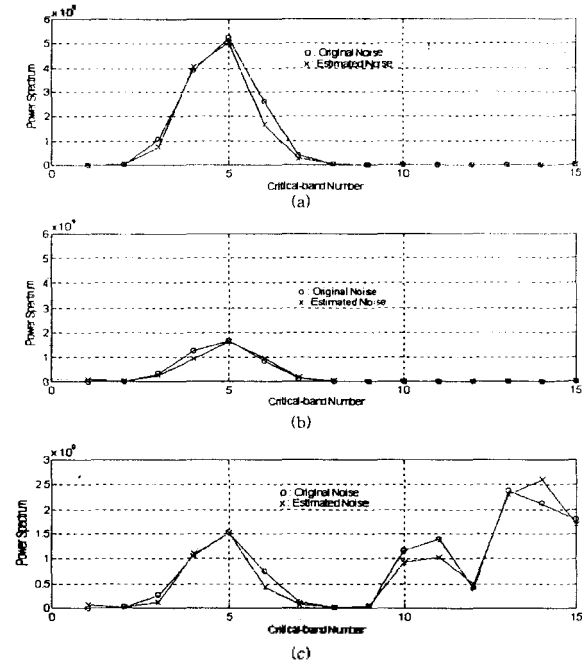
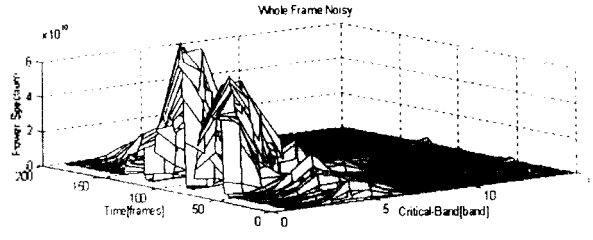


그림 3. 히스토그램 처리 방법을 이용하여 잡음음성으로부터 추정한 평균잡음
(a) 5dB SNR의 유색 가우시안 잡음
(b) 10dB SNR의 유색 가우시안 잡음
(c) 5dB SNR의 보다 복잡한 유색 가우시안 잡음

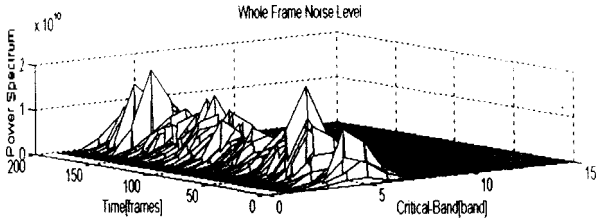
Fig. 3. Average noise spectrum estimated from the noisy speech using the histogram processing method,
(a) Colored Gaussian noise with 5dB SNR,
(b) Colored Gaussian noise with 10dB SNR,
(c) More complicated colored Gaussian noise with 5dB SNR.

그림 3에서는 단어 "재무관리실"에 유색 가우시안 잡음을 더한 잡음음성에 대해서 히스토그램 처리방법을 이용하여 잡음을 추정한 결과를 나타내고 있다. 그림 3(a)는 5dB SNR, 그리고 그림 3(b)에서는 10dB의 SNR에 대해 실험용 잡음의 스펙트럼 레벨과 추정된 잡음의 전체 프레임에 대한 평균을 각각 나타내고 있다. 각각에 대해 입력음성에 더하기 이전의 잡음의 스펙트럼 분포와 거의 일치함을 알 수 있다. 그림 3(c)는 의도적으로 보다 복잡한 형태로 구성한 유색잡음에 대해 히스토그램 방법을 이용하여 잡음 스펙트럼을 추정한 결과를 나타내고 있으며, 역시 잡음음성의 스펙트럼에서 추정한 잡음이 순수 잡음신호와 매우 유사한 결과를 보여주고 있다.

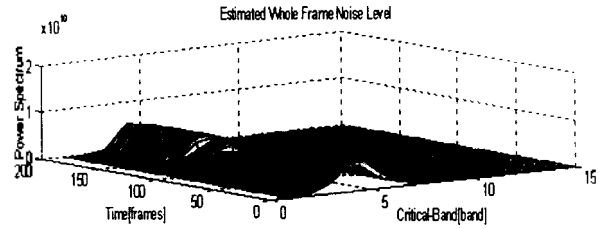
그림 4는 단어 "재무관리실"에 대한 잡음음성[그림 4(a)] 및 모의환경에서 발생시킨 잡음의 스펙트럼[그림 4(b)],



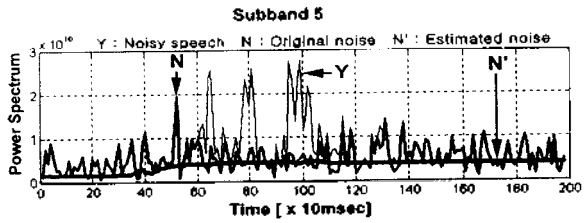
(a)



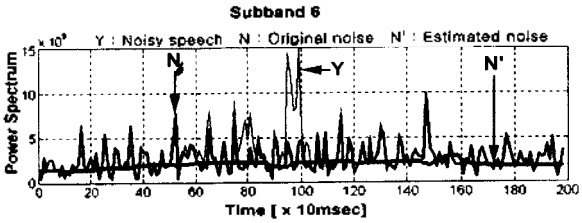
(b)



(c)



(d)



(e)

그림 4. 히스토그램 처리방법에 의한 잡음추정의 예
(a) 5dB SNR로 잡음이 더해진 잡음음성의 스펙트럼
(b) 순수잡음의 스펙트럼
(c) 잡음음성으로부터 추정한 잡음레벨
(d) 5번째 밴드의 잡음음성, 잡음 및 추정된 잡음레벨
(e) 6번째 밴드의 잡음음성, 잡음 및 추정된 잡음레벨

Fig. 4. Examples of noise estimation by the histogram processing method,
(a) Spectra of noisy speech with additive noise of 5dB SNR,
(b) Original noise spectra,
(c) Estimated noise spectra from the noisy speech.
(d) Noisy speech, original noise and estimated noise for 5th subband,
(e) Noisy speech, original noise and estimated noise for 6th subband.

그리고 히스토그램 분석기법에 의하여 추정한 잡음[그림 4(c)]의 스펙트럼 레벨을 나타내었으며 또한 band 5와 band 6에 대해서 잡음음성 및 모의 발생된 잡음 그리고 추정한 잡음레벨을 그림 4(d)-(e)에 각각 나타내었다. 이들 그림에서 스펙트럼의 진폭에 Y로 표기한 것이 잡음이 섞인 음성의 스펙트럼을 나타내고 N으로 표기한 것이 잡음의 스펙트럼 레벨을 나타내며, 그리고 비교적 일정한(stationary) 값을 유지하는 것(N'로 표기)이 추정된 잡음의 레벨을 나타내고 있다. 그림 4(c)-(e)에서 추정된 잡음은 초기 10 프레임에 대해서는 스펙트럼 평균을 취하여 잡음레벨로 추정한 것이며 그 이후에는 초기 10프레임을 포함하여 히스토그램 분석 구간별로 스펙트럼의 분포밀도 함수에 의해 추정된 잡음의 레벨을 나타내고 있다. 추정된 잡음의 레벨은 모의 발생된 잡음의 스펙트럼 레벨 평균과 거의 일치하고 있음을 잘 보여주고 있다. 그러나, 그림 4(b), (d) 및 (e)에서 보는 바와 같이 원래의 잡음 자체는 프레임간의 편차가 매우 큰 특성을 가지므로, 잡음레벨로 추정된 잡음을 스펙트럼 차감법에 의해 빼주는 것 만으로는 잡음의 영향이 충분히 제거되지 않으며, 또한 상당수의 프레임에서 음의 값이 나오게 된다. 실제로는 2.2절에서 설명한 바와 같이 반파정류 및 overestimation factor 도입 등의 방법으로 이러한 문제를 극복하게 된다.

Ⅳ. 인식실험 결과 및 고찰

인식실험은 22개의 부서명을 대상으로 한 한국전자통신연구소의 부서명 음성 데이터베이스[9] 중에서 고립단어 형태의 음성 데이터 만을 이용하여 화자독립으로 수행하였다. 음성자료는 표 1에 나타낸 것과 같이 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 인식 실험을 하였다.

본 논문에서의 음성인식 시스템은 12차의 PLP 켑스트럼 계수를 특징 파라메터로 사용하였으며, 상용화된 HMM 인식도구인 HTK1.5를 이용하여 훈련 및 인식을 수행하였다.[10] 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. 훈련과정의 iteration수는 초기모

델 작성시에 15회로 하고 reestimation 알고리즘 수행시에는 20회로 하였다. 전처리 과정에서 잡음의 추정방법으로는 미리 비음성구간으로 확인된 초기 10 프레임의 잡음평균을 이용하는 방법과 히스토그램 처리방법의 두 가지 방법에 대한 실험을 별도로 수행하였다. 스펙트럼 차감법은 식(10)에 의해 처리하였으며, 이때 β는 10^{-6}으로 하고 α는 1.0 및 1.5 각각에 대하여 인식실험을 하였다.

표 1. 인식대상 어휘목록
Table 1. Word list for recognition.

1	총무부	7	재무관리실	13	설비과	19	인력개발부
2	운영관리실	8	회계과	14	영선과	20	연력계획과
3	총무과	9	내사과	15	근로복지실	21	인사과
4	자산관리과	10	외자과	16	근로과	22	연수실
5	인전관리과	11	건설관리실	17	복지과		
6	예비군대대본부	12	건설과	18	서울사무소		

먼저 3장에서 설명한 히스토그램 처리방법을 이용한 잡음 스펙트럼 추정방법을 스펙트럼 차감법에 적용하기 위해서는 먼저 히스토그램 처리과정에서의 여러 가지 파라메터들을 적절하게 선정할 필요가 있다. 이때 검토대상이 되는 파라메터들로는 히스토그램 처리를 위한 분석구간 길이, 분석구간의 이동 간격, 히스토그램의 분해도(resolution), 그리고 잡음추정 이전에 critical-band 출력에서 과거 프레임과의 smoothing 여부 등이다. 이들 파라메터의 선정을 위해 다양한 SNR(5~30dB)에서의 유색 가우시안 잡음이 부가된 상황에서 스펙트럼 차감법을 적용한 인식실험을 수행하고, 인식성능이 가장 우수한 경우의 파라메터들을 선정하였다. 그 결과 히스토그램 처리를 위한 분석구간 길이는 50 프레임(500msec), 분석구간의 이동 간격은 2 프레임(20msec), 히스토그램의 분해도는 50, 그리고 잡음 추정 이전에 과거 프레임과의 smoothing을 하는 것으로 결정하고, 이들 파라메터들을 이후의 실험에 적용하였다. 또한, 스펙트럼 차감법에서의 과추정 상수 α에 대해서는 1.0과 1.5의 값을 검토하였으며, $\alpha=1.5$인 경우가 SNR이 낮을 때에는 보다 우수한 결과를 나타내었지만 높은 SNR에서 저조한 결과를 나타내어 최종적으로 1.0의 값을 사용하였다. SNR을 추정하여 그 결과에 따라 α 값을 적응시키는 방법은 본 논문에서는 시도하지 않았다.[4]

표 2는 본 실험에서의 인식 결과를 나타낸 것이다. 즉 기존의 분석방법(LPC, PLP), 음성의 초기 프레임을 평균한 방법 그리고 히스토그램 분석기법에 대한 인식결과를 나타내었다. 표 2에서 사용된 특징벡터는 모두 12차로 하였으며 초기 프레임의 평균에 의해 잡음 스펙트럼을 추정하는 방법에서 초기 프레임은 미리 수작업에 의해 비음성 구간으로 확인된 처음 10프레임으로 하였다.

표 2(a)는 유색 가우시안 잡음이 부가된 경우이며 실험에 사용한 유색잡음은 3.3절에서 언급된 바와 같이 모의

표 2. 잡음환경에서의 인식 결과
Table 2. Recognition results in noise environments.
(a) 유색 가우시안 잡음의 경우
(a) In case of colored gaussian noise.

Feature	Noise Estimation	Accuracy[%]				
		Clean	30 dB	20 dB	10 dB	5 dB
PLP	Histogram	96.1	96.4	95.2	84.2	74.6
PLP	Frame Average	94.2	93.6	86.7	64.2	47.0
PLP	NO	96.7	87.3	56.7	18.2	4.2
LPC	NO	97.3	69.7	26.1	9.4	6.1

(b) 자동차 소음의 경우
(b) In case of car noise.

Feature	Noise Estimation	Accuracy[%]				
		Clean	30 dB	20 dB	10 dB	5 dB
PLP	Histogram	96.1	95.8	88.2	51.8	33.6
PLP	Frame Average	94.2	90.0	74.9	28.5	12.1
PLP	NO	96.7	79.1	67.9	31.5	17.0
LPC	NO	97.3	76.7	49.4	19.7	11.8

발생시킨 백색 가우시안 잡음을 900Hz에서 1.6kHz를 통과대역으로 하는 대역필터를 통과시킨 것을 사용하였다. 실험 결과, 잡음제거 처리를 하지않은 LPC 분석방법의 경우 clean 환경일 때 97.3%로 상대적으로 가장 우수한 성능을 나타내었으나 SNR이 낮아짐에 따라 급격한 성능저하를 보여주고 있다. PLP 분석방법의 경우에도 스펙트럼 차감법을 적용하지 않았을 때에는 SNR이 낮아짐에 따라 현저한 성능저하를 나타내지만, 기존의 LPC 분석방법보다는 잡음에 대해 상대적으로 강인한 특성을 가짐을 확인할 수 있다. 히스토그램 처리방식 또는 초기 프레임 평균 방법에 의해 추정된 잡음 스펙트럼을 이용하여 스펙트럼 차감법을 적용할 경우 잡음환경에서의 인식 성능이 개선되었다. 그 중에서도 특히 히스토그램 처리방식이 가장 우수한 결과를 나타내었다.

실제로 표 2(a)에서 히스토그램 처리방법을 이용한 인식성능은 20dB SNR에 대해 95.2%로서 초기 프레임을 평균한 방법에서 clean환경의 인식률(94.2%)보다도 우수하다. 또한 10dB SNR에서의 인식률(84.2%)은 초기 프레임을 평균한 방법에서 20dB SNR의 인식률과 유사하며, 잡음처리를 하지 않은 LPC 분석방법에서 30dB SNR의 결과보다도 훨씬 우수한 결과를 보여주고 있다.

표 2(b)는 100km/hour로 주행중인 자동차 소음이 더해진 경우이며 유색 가우시안 잡음의 경우와 동일한 경향, 즉, 히스토그램 처리에 의해 성능이 크게 향상됨을 보여주고 있다. 예를 들어 히스토그램 처리방식의 경우 20dB SNR에서 인식률이 88.2%로서 인식률은 30dB SNR에서 초기 프레임을 평균한 방법의 인식률(90.0%)과 유사한 결과를 보여주며, 잡음처리를 하지 않은 경우의 30dB SNR에서 PLP나 LPC 분석방법의 인식률(79.1% 및 76.7)

보다도 현저한 성능향상을 나타내고 있다. 표 2(a)와 (b)를 비교해 볼 때 유색 가우시안 잡음의 경우 자동차 주행소음의 경우에 비해 동일한 SNR에서 인식성능이 우수한 결과를 보여주고 있다. 이것은 모의 발생된 유색 가우시안 잡음이 실제 자동차 주행소음에 비해 보다 정적(stationary) 특성을 가지기 때문으로 해석된다.

계산량의 관점에서 볼 때, 초기 프레임의 평균에 의해 잡음 스펙트럼을 추정하는 방법의 경우 계산량의 증가가 거의 없는 반면에, 히스토그램 처리방법의 경우에는 약간의 계산량 추가가 필요하다. 그러나, 본 논문에서 수행한 실험에 따르면 히스토그램 처리방법에 의한 계산량 증가분은 전체 인식 소요 계산량의 약 3%에 불과하므로 크게 문제되지 않는 것으로 판단된다.

V. 결 론

본 논문에서는 부가잡음에 강인한 음성인식 시스템을 개발하기 위하여 현재까지 제안된 전처리 단계에서 많이 이용되고 있는 PLP 처리 및 음질개선을 위한 스펙트럼 차감법에 대해 검토하고, 잡음 스펙트럼의 추정방법으로서 히스토그램 처리방식을 도입하여 기존의 방식과 잡음환경에서의 인식성능을 비교하였다. 히스토그램 처리방법은 음성/비음성 구간의 검출과정이 없이도 신뢰도 높은 잡음 스펙트럼 추정을 가능케 하며, 시간에 따라 서서히 변화하는 잡음도 추정할 수 있는 장점을 가진다.

모의 발생시킨 유색 가우시안 잡음과 실제 자동차 주행소음을 이용한 잡음환경에서의 화자독립 고립단어 인식실험 결과, 히스토그램 처리 방법에 의해 잡음 스펙트럼을 추정하고 스펙트럼 차감법을 적용한 경우가 초기 프레임의 평균으로 잡음 스펙트럼을 추정한 후 스펙트럼 차감법을 적용한 경우보다 모든 SNR에 대해 우수한 성능을 나타내었으며, 특히 SNR이 낮아질수록 성능차이가 두드러짐을 확인하였다. 앞으로 히스토그램 처리 방법의 보완과 더불어 비선형 스펙트럼 차감법 및 Wiener 필터링 방법 등에 대한 적용이 계속 진행될 예정이다. 또한, 시간에 따라 변화하는 잡음에 대한 인식실험도 수행할 계획이다.

※본 논문에서 사용한 단어 데이터베이스는 한국전자통신연구원이 구축한 부서명 음성 데이터베이스를 사용하였습니다.

참 고 문 헌

1. B. H. Juang, "Speech recognition in adverse environments," Computer Speech and Language, vol. 5, pp. 275-294, 1991.
2. J. C. Junqua and J. P. Haton, Robustness *in Automatic Speech Recognition*, Kluwer Academic Publishers, USA, 1996.
3. 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.
4. P. Lockwood and J. Boudy, "Experiments with a nonlinear Spectral Subtraction(NSS) and hidden markov models and the projection, for robust speech recognition in cars," Speech Communication, 11(2-3):215-228, 1992.
5. A. A. Nolazco Flores and S. J. Young, "Continuous speech recognition in using spectral subtraction and adaptation," in Proc. IEEE ICASSP-94, pp. 409-412, 1994.
6. H. G. Hirsch, "Estimation of noise spectrum and its application to SNR estimation and speech enhancement," Technical Report TR-93-012, International Computer Science Institute, Berkeley, USA, 1993.
7. H. G. Hirsch and C. Ehricher, "Noise estimation techiques for robust speech recognition," in Proc. IEEE ICASSP-95, pp. 153-156, May 1995.
8. H. Hermansky, "Perceptual linear predictive analysis for speech," J. Acoust. Soc. Am., pp. 1738-1752, 1990.
9. 이영직 외, "ETRI의 음성 데이터베이스 구축 현황," 제12회 음성통신 및 신호처리 워크샵 논문집, pp. 265-267, 1995. 6.
10. S. J. Young et al., *HTK: Hidden Markov Model Toolkit V1.5*, Entropic Research Laboratory, Inc., 1993.

▲권 영 욱(Young Uk Kwon)

1986년 2월:부경대학교 전자공학과(공학사)
1991년 8월:영남대학교 전자공학과(공학석사)
1993년 3월~현재:부산대학교 대학원 전자공학과(박사과정)
1987년 10월~현재:부경대학교 공과대학 전자공학과 조교
※주관심분야:음성신호처리 및 그 응용

▲김 형 순(Hyung Soon Kim)

1983년 2월:서울대학교 전자공학과(공학사)
1984년 2월:한국과학기술원 전기 및 전자공학과(박사과정 조기진학)
1989년 2월:한국과학기술원 전기 및 전자공학과(공학박사)
1987년 1월~1992년 6월:디지콤 정보통신연구소 선임연구원, 연구부장
1992년 7월~현재:부산대학교 전자공학과 조교수 부산대학교 정보통신연구소 연구원