히스토그램 기반의 과추정 방식을 이용한 잡음에 강인한 음성인식

Noise-Robust Speech Recognition Using Histogram-Based Over-estimation Technique

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

* 디지털코리아 (부경대기술사업단), ** 부산대학교 전자공학과
(접수일자: 1999년 8월 9일; 채택일자: 1999년 10월 7일)

잡음환경에서의 음성인식 성능향상을 위해서는 서로 다른 잡음환경으로 인한 mismatch를 줄이는 것이 중요하다. 이를 위해 계산이 간단하고 잡음환경에서 비교적 우수한 성능을 내고 있는 스펙트럼 차감법이 널리 사용되고 있다. 본 논문에서는 스펙트럼 차감법을 적용하기 위한 잡음 스펙트럼 추정방법으로 히스토그램 처리방법을 도입한다. 히스토그램 처리방법은 음성이 아닌 구간의 검출이 필요없으며 시간에 따라 변화하는 시변잡음에도 적용 가능한 장점이 있다. 그러나 히스토그램 처리방법으로 신뢰도 높은 잡음 스펙트럼의 평균값을 추정하더라도 스펙트럼 차감법을 적용했을 때의 잔여 잡음의 문제가 발생한다. 이를 해결하기 위하여 잡음추정 과정에 사용되었던 히스토그램의 분포특성을 고려한 새로운 over-estimation 적용방식을 제안한다. 제안된 방식은 측정된 잡음의 분포에 따라 적응적으로 over-estimation의 정도를 결정함으로써 SNR 변화에 따른 영향이 적은 장점이 있다. 자동차 소음 환경에서의 화자독립 고립단어 인식실험 결과, 기존의 over-estimation factor를 적용한 경우보다 제안된 방식의 인식성능이 개선되었다
핵심용어: 스펙트럼 차감법, 과추정, 잡음추정, 자동차 소음, 화자독립, 고립단어
투고분야: 음성처리 분야(2.5, 2.3)

In the speech recognition under the noisy environments, reducing the mismatch introduced between training and testing environments is an important issue. Spectral subtraction is widely used technique because of its simplicity and relatively good performance in noisy environments. In this paper, we introduce histogram method as a reliable noise estimation approach for spectral subtraction. This method has advantages over the conventional noise estimation methods in that it does not need to detect non-speech intervals and it can estimate the noise spectra even in time-varying noise environments. Even though spectral subtraction is performed using a reliable average noise spectrum by the histogram method, considerable amount of residual noise remains due to the variations of instantaneous noise spectrum about mean. To overcome this limitation, we propose 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 has advantages to be few the influnce of the SNR variation on the noise levels. According to speaker-independent isolated word recognition experiments in car noise environment under various SNR conditions, the proposed histogram-based over-estimation technique outperforms the conventional over-estimation technique.
Key words: Spectral subtraction, Over-estimation, Noise estimation, Car noise, Speaker independent, Isolated word

I. 서 론

잡음환경에서 음성인식의 성능향상을 위해 전처리과정을 통해 잡음을 제거하는 음질개선 방식들 중에서 현재 가장 널리 사용되는 대표적인 방법은 스펙트럼 차감법이다[1][2][3]. 이 방법은 입력된 음성에서 잡음 스펙트럼을 추정하여 이를 잡음이 섞인 음성 스펙트럼에서 빼주는 것으로, 계산이 간단하고 잡음환경에서 비교적 우수한 성능을 나타낸다. 스펙트럼 차감법의 성공적인 적용을 위해서는 신뢰도 높은 잡음 스펙트럼 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 자

체가 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 성립되지는 않으며 시변 잡음환경에 대처할 수 없다는 문제점이 있다.

스펙트럼 차감법이 지니는 두 번째 문제점은 잡음 스펙트럼의 평균을 빼주기 때문에 낮은 SNR에 대해서는 스펙트럼 차감법 이후에도 잔여(residual) 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 논문의 선행연구에서는 스펙트럼 차감법을 적용하기 위한 신뢰도 높은 잡음 스펙트럼 추정방법으로 히스토그램 처리방법을 도입한 바 있다. Hirsch에 의해 제안된 히스토그램 처리방법은 음성이 아닌 구간의 검출이 필요 없으며 SNR의 의존도가 적은 장점이 있다[4][5]. 특히 이 방법은 시간적으로 서서히 변화하는 잡음의 특성에 대해서도 적용이 가능하다[6]. 그러나 히스토그램 처리방법으로 신뢰도 높은 잡음 스펙트럼의 평균값을 추정해 내더라도 스펙트럼 차감법을 적용했을 때의 잔여잡음의 문제는 여전히 남아 있다. 이 문제의 해결을 위해 본 논문에서는 히스토그램 처리방법에 의한 잡음추정 과정에서 히스토그램의 분포특성을 고려한 히스토그램 기반의 over-estimation 적용방식을 제안한다. 이 방식은 측정된 잡음의 분포에 따라 적응적으로 over-estimation의 정도를 결정함으로써 기존의 스펙트럼 차감법에 비해서 잡음음성의 SNR에 대한 영향이 적은 장점이 있다. 실제로 자동차 소음 환경에서의 인식실험 결과 기존의 over-estimation factor를 적용한 경우보다 인식성능이 개선되었다.

II. 히스토그램 기반의 스펙트럼 차감법

2.1. 스펙트럼 차감법

실제 환경에서의 음성신호에는 다양한 종류의 잡음이 존재하며, 채널왜곡에 의한 영향을 무시한다면 대부분이 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. 일반적으로 $n(m)$은 정적이거나 $x(m)$에 비해 천천히 변화하는 잡음으로 $x(m)$과는 서로 독립적이라고 가정한다.

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

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

스펙트럼 차감법은 잡음이 섞인 신호에서 잡음을 억제하는 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. 스펙트럼 차감법은 일반적으로 다음 식과 같은 변형된 방법이 널리 사용된다[1][2][3].

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

여기서 α는 over-estimation factor이고 β는 flooring factor를 나타낸다. 그리고 $\hat{N}(f)$는 추정된 잡음 스펙트럼을 의미한다.

스펙트럼 차감법에서는 잡음 스펙트럼의 추정치를 어떻게 구하는가 하는 것이 중요한 관건이며, 일반적으로 비음성구간으로 확인된 잡음구간에서의 평균레벨을 추정된 잡음으로 간주한다. 이 경우의 스펙트럼 차감법은 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 식 (3)에서와 같이 α배만큼 올려주는 over-estimation 처리를 한다. 그리고 식 (3)에서와 같이 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[2][3]. 스펙트럼 차감법은 비교적 연산이 간단하면서도 잡음환경에서 상당한 효과가 있기 때문에 잡음환경에서의 음성인식에 많이 사용된다. 본 논문에서는 스펙트럼 차감법에서 먼저 신뢰성 있는 잡음을 추정하기 위하여 히스토그램 처리방법을 도입하였다.

2.2. 히스토그램 기반의 스펙트럼 차감법

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 높을

수록 잡음 음성신호의 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[4][5].

히스토그램 처리방법에 의해 잡음 스펙트럼을 추정하는 방법을 요약하면 다음과 같다[5]. 먼저 매 프레임 단위로 입력음성을 FFT하여 주파수 영역으로 변환한 다음, 청각기관의 특성을 고려한 멜-스케일 삼각 필터뱅크를 통과시켜 MFCC를 구한다. 본 논문에서는 8kHz의 샘플링 주파수를 가지는 입력음성을 대상으로 256-point FFT를 수행한 다음, 입력음성의 주파수 범위가 0에서 4kHz 사이에서 25개의 멜-스케일 필터뱅크 출력을 얻는다. 이들 각 대역에 대해 개별적으로 히스토그램 처리과정을 수행하여 그 주파수 대역에 해당하는 잡음레벨을 추정하게 되며, 그 결과를 종합하면 각 대역별 주파수 분석에 따른 잡음 스펙트럼 특성을 구할 수 있다. 이하에 각 대역별로 잡음레벨을 추정하며 앞서 설명한 히스토그램 분석구간(현재 프레임을 기준으로 그 전후의 일정 수의 프레임들, 본 논문에서는 50프레임)에 대해 각 프레임별로 구한 해당 주파수 대역의 스펙트럼 크기들을 이용하여 분포 밀도함수, 즉, 히스토그램을 구한다. 이 때, 히스토그램 분석의 분해능(resolution)이 M이라 하면, 0에서 스펙트럼 크기의 최대값 사이를 M등분하여 각각의 돗수값을 누적시켜 히스토그램을 구한다. 이렇게 구해진 히스토그램에서 돗수값이 가장 큰 스펙트럼 크기, 즉, 최빈값을 해당 주파수 band의 잡음레벨의 추정치로 정한다.

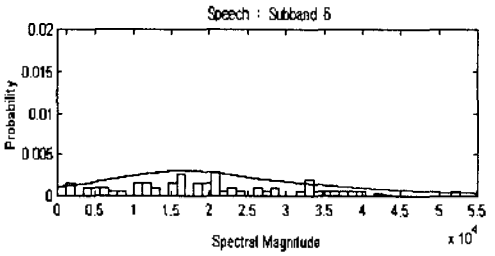
III. 히스토그램 기반의 over-estimation 방법

이미 언급한 바와 같이 히스토그램 처리방법을 통해 잡음의 평균 스펙트럼을 보다 신뢰성 있게 추정한다고 하더라도, 잡음 평균을 빼주는 스펙트럼 차감법을 적용할 경우 잔여잡음의 문제점이 해결되지 않는다. 따라서, 2.1절의 식 (3)에서와 같이 적절한 over-estimation factor α를 도입하여 잡음평균 스펙트럼보다 과도하게 빼 줌으로써 잔여 잡음의 문제를 완화시키는 방법이 일반적으로 사용된다. 이때 over-estimation factor를 선정함에 있어서, 낮은 SNR을 기준으로 하면 잡음 자체의 분산이 크기 때문에 over-estimation factor를 높여주어야 잔여 잡음이 줄어들게 된다. 그러나, 그 대신 음성구간에서는 오히려 음성의 스펙트럼을 왜곡시키게 된다. 반면에 높은 SNR을 기준으로 하여 over-estimation factor를 낮게 선정할 경우에는 낮은 SNR의 잡음구간에서 스펙트럼 차감법 이후에도 잡음이 처리되지 않고 남게 된다.

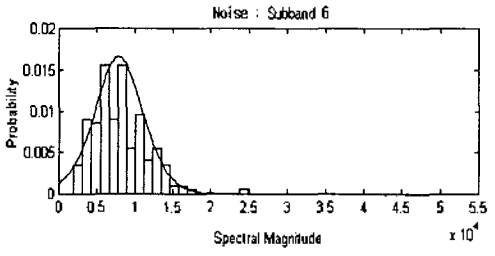
이러한 문제를 해결하기 위한 방법으로, 2.2절에 설명한 스펙트럼 크기의 통계적인 특성의 관찰결과에 근거하여 잡음이 부가된 잡음음성에서 그림 1과 같이 각각의 주파수 대역별 스펙트럼 크기에 대하여 두 개의 가우시안으로 모델화하고 잡음분포의 레벨을 결정하는 방식이 검토될 수 있다. 그림 1(a)는 음성에 대한 히스토그램과 가우시안 분포를 나타낸 것이며 그림 1(b)는 순수잡음에 대한 히스토그램과 가우시안 분포를 나타내었다. 그리고 그림 1(c)는 잡음음성에 대한 히스토그램 및 두 개의 가우시안 분포를 나타낸 것이다. 그림 1(c)에서 기존의 히스토그램 기반의 스펙트럼 차감법에서는 두 개의 분포 중에서 잡음영역 가우시안 분포의 최빈값 근처, 즉 히스토그램의 최빈값에 해당하는 스펙트럼 크기를 추정하고자 하는 평균적인 잡음레벨로 결정한다.

스펙트럼 차감법에서 over-estimation을 적용할 때 히스토그램 및 가우시안 분포에서 잡음과 음성의 분포특성의 경계(threshold)를 기반으로 over-estimation을 설정하게 되며, 경계의 높낮이에 따라 음성 및 잔여잡음의 특성은 식 (3)에서 over-estimation factor α를 적용한 경우와 같은 특성을 나타낸다. 따라서 음성의 왜곡을 최소화하는 범위에서 잡음을 최대한으로 제거하기 위해서는 두 개의 가우시안 분포가 만나는 경계에서 over-estimation을 결정하는 것이 바람직하다고 할 수 있겠다.

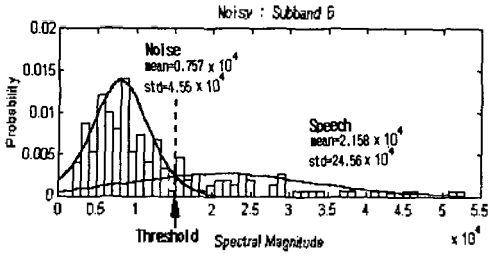
가우시안 분포특성을 고려한 over-estimation 방식은 높은 SNR의 잡음음성에서는 잡음과 음성의 두 가우시안의 분포가 뚜렷하게 분리되며, 이때는 잡음영역 가우시안 분포의 최소값 중에서 스펙트럼 크기가 큰 쪽의 값을 해당 주파수 대역에서의 잡음레벨이라 판정한다. 그리고 낮은 SNR의 잡음음성에서는 잡음과 음성의 두 가우시안의 분포가 겹쳐 있으며, 이 경우에는 잡음영역 가우시안 분포의 평균값(분포의 최대값) 또는 over-estimation을 교차되는 경계부분으로 설정하게 된다. 이러한 가우시안 분포는 잡음음성의 SNR에 따라 분산이 달라진다. 특히 잡음분포의 분산은 SNR에 따라 민감한 변화를 나타낸다.



(a)



(b)



(c)

그림 1. 가우시안 분포로 표현한 잡음음성의 스펙트럼
(a) 원 음성의 스펙트럼 크기에 대한 히스토그램 분포 및 가우시안 모델
(b) 5dB SNR 잡음의 스펙트럼 크기에 대한 히스토그램 분포 및 가우시안 모델
(c) 두 개의 가우시안 분포로 적응시킨 5dB SNR의 잡음음성에 대한 히스토그램

Fig. 1. Spectrum of noisy speech represented with Gaussian distribution,
(a) Histogram distribution and Gaussian models for spectral magnitude of speech,
(b) Histogram distribution and Gaussian models for spectral magnitude of noise with 5dB SNR,
(c) Histogram for noisy speech 5dB SNR fitted with two Gaussian distribution.

이와 같이 잡음과 음성의 두 가우시안 분포특성에서 경계를 이용한 over-estimation 결정방식은 입력음성의 SNR에 관계없이 잡음이 분포하는 정도를 항상 일정한 비율로 적응하는 잡음레벨을 결정할 수가 있다. 그러나 이 방법은 Expectation-Maximization(EM) 알고리즘에 의한 가우시안 분포의 모델링에 처리시간이 많이 소요되어 실시간 구현에 문제가 있다.

따라서 본 논문에서는 위에서 설명한 모델에 의한 방법을 사용하는 대신에 히스토그램 최대값의 γ 배인 $\gamma \cdot H_{max}$에 해당하는 스펙트럼 크기를 그 대역에서의 잡음레벨로 결정하는 히스토그램 기반의 over- estimation 방식을 제안한다. 이와 같이 기존의 히스토그램 기반의 스펙트럼 차감법에서 over-estimation의 정도를 적응적으로 결정하게 되면, 잡음음성의 SNR이 달라지더라도 잡음과 음성분포의 분산에 의한 경계를 비교적 잘 표현할 수 있다.

그림 2에서 중심주파수 f_0인 대역필터 출력 값들로부터 구한 히스토그램의 최대값 H_{max}에 해당하는 스펙트럼 크기의 최빈값 $\hat{N}(f_0)$가 기존의 히스토그램 처리방법에서 결정하는 잡음의 평균레벨을 나타내며, 기존의 스펙트럼 차감법에서는 식 (3)과 같이 $\hat{N}(f_0)$에 over-estimation factor α를 곱한 $\alpha \cdot \hat{N}(f_0)$를 사용한다. 이에 반하여 본 논문에서 제안한 히스토그램 기반의 over-estimation 방식은 그림 2의 히스토그램 최대값에서 γ배인 값 중에서 스펙트럼의 크기가 큰 쪽(그림 2의 두 교차점에서 우측부분)의 $\gamma \cdot H_{max}$에 해당하는 스펙트럼 크기 $\hat{N}_{OE}(f_0)$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 factor라고 부르기로 한다.

이러한 히스토그램 분포특성을 기반으로 하는 over-estimation에 의한 스펙트럼 차감법은 잡음이 거의 없는 높은 SNR의 음성뿐만 아니라 잡음이 많은 낮은 SNR의 잡음음성 등의 다양한 SNR에서도 추정한 잡음레벨을 잡음의 정도에 따라 별도로 조정할 필요가 없다. 즉, 잡음의 정도에 따라 히스토그램 분포의 분산이 달라지더라도 히스토그램의 최대값에 대한 비율로써 잡음의 레벨이 결정되기 때문에 SNR 변화에 따른 잡음의 분산이 달라지더라도 잡음의 분포를 항상 같은 비율로 반영시킬 수 있는 장점이 있다.

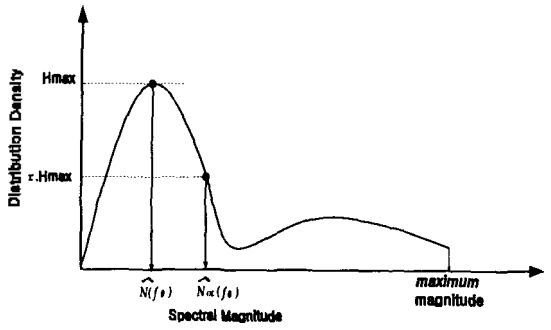


그림 2. 히스토그램 기반의 over-estimation 방식
Fig. 2. Histogram-based over-estimation method.

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

$$\hat{X}(f) = \begin{cases} Y(f) - \hat{N}_{OE}(f_0) & \text{if } Y(f) - \hat{N}_{OE}(f_0) > \beta \cdot \hat{N}_{OE}(f_0) \\ \beta \cdot \hat{N}_{OE}(f_0) & \text{otherwise} \end{cases} \quad (4)$$

여기서 $\hat{N}_{OE}(f_0)$는 히스토그램 기반의 over-estimation 방식에서 추정한 잡음레벨을 나타내며, 기존의 스펙트럼 차감법에서의 $\alpha \cdot \hat{N}(f_0)$에 대응되는 값이다. 그리고 β는 flooring factor를 나타내는 것으로 기존의 스펙트럼 차감법에서 사용한 것과 동일하다.

그림 3에 유색잡음이 부가된 단어음성 "재무관리실"에서 SNR에 따른 히스토그램 분포와 스펙트럼 차감법 적용을 위한 잡음레벨의 설정 예를 나타내었다. 그림 3(a)에서 $\alpha \hat{N}$으로 표현된 잡음레벨은 히스토그램 기반의 스펙트럼 차감법에서 적용한 기존의 over-estimation을 나타내는 것으로써, 잡음과 음성의 두 분포의 경계를 기준으로

히스토그램 기반의 over-estimation 방식에서의 잡음레벨 $\hat{N}_{OE}$보다도 낮게 설정되어 그 차이가 큼을 보여준다. 10dB SNR의 음성에 대한 그림 3(b)와 그리고 0dB SNR의 음성을 나타낸 그림 3(c)의 경우는 기존의 방법이 두 분포의 경계에서 스펙트럼의 크기가 높은 값으로 선정되어 음성 스펙트럼의 일부를 포함하고 있을 뿐만 아니라, 두 분포의 경계를 기준으로 히스토그램 기반의 over-estimation에 비해서 그 차이가 더 크게 나타남을 보여주고 있다. 반면에 히스토그램 기반의 over-estimation 방식에서 결정된 잡음레벨 $\hat{N}_{OE}$은 SNR이 달라지더라도 잡음과 음성의 두 분포의 경계에 보다 가깝게 설정되어 일정한 비율의 잡음 분포를 반영함을 알 수가 있다. 그림 3의 설명에서 히스토그램의 최빈값을 추정된 잡음레벨로 선정하여 잡음을 처리하는 과정에서 over-estimation factor α를 1.5로 하는 기존의 히스토그램 기반의 스펙트럼 차감법에서는 SNR에 따라 간혹 낮은 잡음레벨 또는 음성영역내의 분포에서 추정된 잡음레벨로 결정하는 경우가 있다. 반면에 히스토그램 기반의 over-estimation 방식에서 γ가 0.3일 때 대부분 잡음의 분포영역을 나타낼 뿐만 아니라 잡음과 음성의 두 가우시안 분포의 경계에 보다 근접하게 적응됨으로써 제거하고자 하는 잡음을 충분히 포함하면서 기존의 방식에 비해 음성의 훼손을 현저히 줄일 수가 있다.

그림 4는 히스토그램 기반의 over-estimation에 따른 잡음레벨 추정치와 기존의 잡음레벨 추정치의 비교를 나타내었다. 그림 4(a)는 유색잡음이 더해진 잡음음성에서 6번째 필터뱅크 출력에 대한 잡음의 레벨을 나타낸 것이며, 기존의 스펙트럼 차감법에서의 over-estimation factor α가 1.47인 경우와 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. 즉, 그림 4(a)에서 기존 방법의 184프레임에 대한 스펙트럼 크기의 평균이 78dB이며, 히스토그램 기반의 over-estimation 방식에서는 184프레임 평균이 78dB로 같은 평균값을 가진다. 그리고 그림 4(b)는 8번째 필터뱅크 출력에 대한 잡음의 레벨을 나타낸 것이며, 그림 4(a)와 같은 방법으로 over-estimation factor를 결정하였다. 기존 방법에 대한 평균이 29dB일 때 α가 1.33이며, 히스토그램 기반의 over-estimation 방식에서는 평균이 29dB일 때 γ 0.3으로 설정한 경우를 나타낸 것이다.

그림 4에서 기존의 히스토그램 처리방법에서 보다 히스토그램 기반의 over-estimation 방식이 잡음의 영역에서 대개 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.

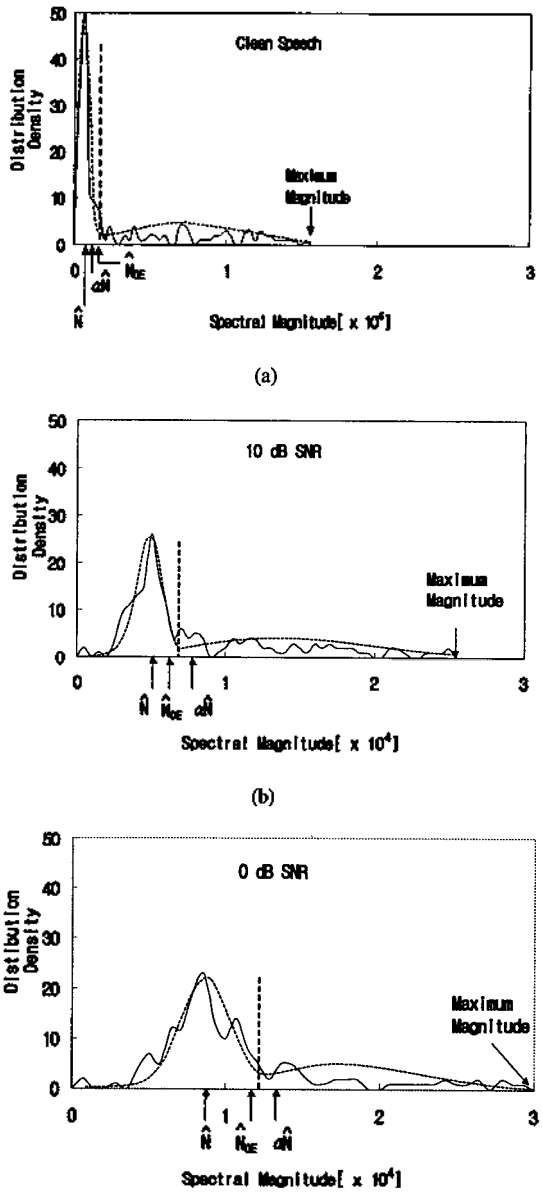


그림 3. SNR 변화에 따른 히스토그램 기반의 over-estimation 및 기존의 over-estimation의 비교(α =1.5, γ =0.3인 경우)
(a) 깨끗한 음성
(b) 10 dB SNR의 잡음음성
(c) 0 dB SNR의 잡음음성

Fig. 3. Comparison of histogram-based over-estimation with conventional over-estimation according to the change of SNR(in case of α =1.5 and γ =0.3),
(a) Original speech,
(b) Noisy speech with 10 dB SNR,
(c) Noisy speech with 0 dB SNR.

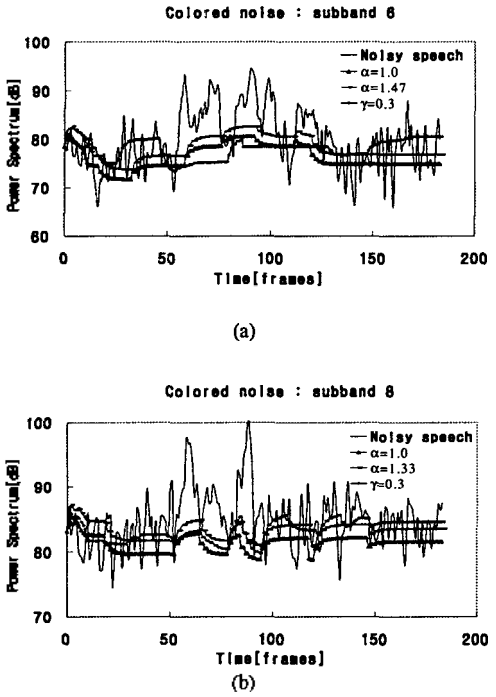


그림 4. 히스토그램 기반의 over-estimation에 따른 잡음레벨 추정치와 기존의 잡음레벨 추정치의 비교
(a) 6번째 필터뱅크의 출력
(b) 8번째 필터뱅크의 출력

Fig. 4. The comparison of noise levels using histogram-based over-estimation and conventional over-estimation,
(a) Output of 6th subband,
(b) Output of 8th subband.

IV. 인식실험 결과 및 고찰

본 논문에서 실험에 사용한 음성인식 시스템은 12차의 MFCC 및 12차의 델타 MFCC 계수들을 특징 파라미터로 삼았으며, 상용화된 HMM 인식도구인 HTK1.5를 이용하여 훈련 및 인식을 수행하였다[8]. 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[7]. 22개의 부서명을 50인 각 1회 발성한 것 중에서 임의로 선택한 35명의 잡음이 섞이지 않은 음성을 모델 형성을 위한 학습용으로 사용하였으며, 나머지 15명의 음성을 인식대상으로 사용하여 각각의 잡음레벨에 따라 인식실험을 하였다.

4.1. 히스토그램 처리방법에 의한 음성인식

스펙트럼 차감법에서 잡음의 추정방법으로는 기존의 프레임 평균한 방법과 본 논문에서 도입한 히스토그램 처리방법에 대해 각각 별도의 처리를 한 후 실제 자동차 소음환경에서 인식실험을 수행하였다. 기존의 프레임 평균에 의해 잡음 스펙트럼을 추정하는 방법은 입력된 잡음음성에서 미리 수작업에 의해 비음성 구간으로 확인된 처음 10프레임을 평균하여 추정하는 잡음레벨로 사용하였다.

표 1은 자동차 소음환경에 대하여 전처리 과정에서 잡음처리를 한 경우의 인식 결과를 나타낸 것이다. 표 1(a)는 정지 상태에서 서서히 속도를 증가시의 자동차 소음이 부가된 경우이며, 스펙트럼 차감법을 하지 않은 경우에 비해서 프레임 평균 및 히스토그램 처리방법을 이용한 스펙트럼 차감법을 수행한 경우가 clean을 제외한 모든 SNR에서 현격한 인식률 향상이 있음을 보여주고 있다. 특히 잡음의 영향이 심할 경우(5dB SNR 이하) 히스토그램 처리방법에서 현저한 성능개선을 보이고 있다. 표 1(b)는 복잡한 시내 도로를 주행시의 자동차 소음이 부가된 경우이며, 표 1(a)의 경우와 유사한 경향을 나타내고 있다.

표 1. 히스토그램 기반의 스펙트럼 차감법에 의한 자동차 소음환경에서의 인식 결과

Table 1. Recognition results in car noise environments using histogram-based spectral subtraction technique,
(a) 정지 상태에서 서서히 속도를 증가시의 소음
(a) In case of car noise at increased speed from idle,

Preprocessing Technique	Noise Estimation	Accuracy[%]						
		Clean	30 dB	20 dB	10 dB	5 dB	0 dB	Average
NO	NO	99.4	98.4	97.6	93.9	85.8	67.8	90.48
Spectral Subtraction	Frame Average	99.1	99.1	99.1	98.5	95.5	68.5	93.30
	Histogram	98.5	98.5	97.5	98.8	97.9	85.8	96.16

(b) 복잡한 시내 도로를 주행시의 소음
(b) In case of car noise at moving on a heavy traffic road.

Preprocessing Technique	Noise Estimation	Accuracy[%]						
		Clean	30 dB	20 dB	10 dB	5 dB	0 dB	Average
NO	NO	99.4	98.4	96.7	83.9	60.8	27.0	77.70
Spectral Subtraction	Frame Average	99.1	98.5	97.9	93.0	75.9	48.5	85.48
	Histogram	98.8	98.1	97.6	93.0	83.0	56.4	87.81

히스토그램 처리방법에 의해 추정한 잡음 스펙트럼을 이용하여 스펙트럼 차감법을 적용한 경우가 초기 프레임 평균방법에 의해 추정한 잡음 스펙트럼을 이용하여 스펙트럼 차감법을 적용한 경우보다 잡음환경에서의 인식 성능이 개선됨을 볼 수 있다. 한편, 계산량의 관점에서 볼 때, 초기 프레임의 평균에 의해 잡음 스펙트럼을 추정하는 방법의 경우 계산량의 증가가 거의 없는 반면에, 히스토그램 처리방법의 경우에는 히스토그램을 처리를 위해 일정한 크기의 입력이 요구됨으로써 약간의 계산량 추가가 필요하다.

그러나, 실제 인식실험에 따르면 히스토그램 처리방법에 의한 계산량 증가분은 전체 인식 소요 계산량의 약 3% 에 불과하므로 크게 문제되지 않는 것으로 판단된다[6].

4.2. 히스토그램 기반의 over-estimation 방식의 적용

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 2는 환경이 서로 다른 두가지 자동차 소음이 각각 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 2. 기존의 over-estimation 및 히스토그램 기반의 over-estimation 방식을 이용한 자동차 소음환경에서의 인식결과 비교

Table 2. Comparison of Recognition results using standard over-estimation and histogram-based over-estimation method in car noise environments,
(a) 정지 상태에서 서서히 속도를 증가시의 소음
(a) In case of car noise at increased speed from idle,

Over-estimation Technique	Over-estimation factor	Accuracy[%]						
		Clean	30 dB	20 dB	10 dB	5 dB	0 dB	Average
NO	NO	99.4	98.4	97.6	93.9	85.8	67.8	90.5
Standard Over-estimation	α = 1.0	98.5	98.2	97.6	95.8	92.4	83.3	94.3
	α = 1.2	98.5	98.2	98.5	95.8	91.2	81.2	93.90
	α = 1.5	98.5	98.5	97.5	98.8	97.9	85.8	96.2
	α = 2.0	98.2	97.8	97.0	87.6	72.4	52.4	84.2
Histogram-based Over-estimation	γ = 0.7	98.5	98.5	98.5	98.8	97.6	92.4	97.4
	γ = 0.5	98.5	98.2	98.5	98.2	97.6	91.2	97.0
	γ = 0.3	98.8	98.5	98.2	98.2	97.3	93.9	97.5
	γ = 0.1	98.5	98.2	97.9	97.0	93.0	80.6	94.2
	γ = 0.05	98.2	98.2	98.2	96.4	92.7	78.8	93.6

(b) 복잡한 시내 도로를 주행시의 소음
(b) In case of car noise at moving on a heavy traffic road.

Over-estimation Technique	Over-estimation factor	Accuracy[%]						
		Clean	30 dB	20 dB	10 dB	5 dB	0 dB	Average
NO	NO	99.4	98.4	96.7	83.9	60.8	27.0	77.70
Standard Over-estimation	α = 1.0	98.5	98.5	97.9	95.5	81.8	43.9	86.01
	α = 1.2	98.5	98.1	97.6	91.5	82.7	52.7	86.85
	α = 1.5	98.8	98.1	97.6	93.0	83.0	56.4	87.81
	α = 2.0	98.2	97.3	96.7	85.8	62.5	37.9	79.73
Histogram-based Over-estimation	γ = 0.7	98.5	98.5	98.2	94.9	85.2	50.8	87.68
	γ = 0.5	98.5	98.5	98.2	93.0	88.8	63.8	90.13
	γ = 0.3	98.8	98.5	97.6	95.1	89.1	64.1	90.53
	γ = 0.1	98.5	97.9	95.5	87.0	87.9	57.7	87.41
	γ = 0.05	98.2	97.9	94.6	81.2	77.9	48.6	83.06

4.3. 인식결과 검토

본 논문에서 이용한 잡음환경에서의 음성인식을 위한 전처리 과정에서 주로 사용하는 스펙트럼 차감법에서의 관건은 잡음 스펙트럼의 추정치를 어떻게 구할 것이며, 우수한 인식성능을 위해서 over-estimation factor를 어떻게 선정할 것인가이다. 일반적으로 입력음성 초기의 수 프레임을 비음성구간이라 가정하고 이들 구간에서 잡음을 추정하는 방법이 많이 사용되고 있으나, 이 가정이 항상 성립되지는 않으며, 또한 잡음의 특성이 변화할 때 이에 대한 대처가 곤란하다. 잡음의 특성이 시간에 따라 변하지 않는 정적인 특성에서는 기존의 초기프레임 평균에 의한 방식에 의해서도 어느 정도 잡음환경에서 음성인식 성능의 향상을 얻을 수가 있다. 그러나 잡음의 환경이 변화가 있을 때는 오히려 잡음처리를 하지 않은 경우보다도 인식률이 떨어지는 경향이 있다. 따라서 이를 해결하기 위한 방법으로 본 논문에서는 입력음성의 매 프레임에서 잡음의 레벨을 추정하는 히스토그램 처리방법 및 잡음의 레벨을 보다 신뢰성 있게 추정하는 히스토그램 기반의 over-estimation 방식을 이용하여 인식성능을 향상시킬 수가 있었다. 잡음의 영향이 심할수록 초기프레임 평균에 의한 방식에 비해서 히스토그램에 의한 스펙트럼 차감법에 의한 처리방식이 보다 향상된 성능을 얻을 수가 있었다.

그림 5는 주행중인 자동차 소음환경에서 clean에서 0dB까지의 다양한 SNR에 대하여 기존의 over-estimation factor를 사용할 때의 결과와 히스토그램 기반의 over-estimation 방식에 따른 평균 인식률의 결과를 나타내었다. 그림에서 Environment 1은 복잡한 시내도로를 주행시의 소음, Environment 2는 정지상태에서 서서히 속도를 증가시의 소음을 나타내었다. 그림 5에서 히스토그램 기반의 over-estimation 방식에서는 γ가 0.3일 때가 Clean 환경에서 0dB SNR까지의 평균 인식률이 94.0%로 가장 우수하며, 이는 기존 방식에서 가장 우수한 α가 1.5에서의 92.0%보다도 우수한 결과이다. 특히 히스토그램 기반의 over-estimation 방식에서는 γ=0.5 및 γ=0.7의 경우에

도 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 90.8%로, 최대값의 5%인 γ가 0.05인 경우는 평균 인식률이 88.3%로 떨어지는 결과를 보인다. 이는 상대적으로 과도한 over-estimation을 사용하기 때문에 잡음구간의 잡음은 대부분 제거되는 반면, 음성구간에서 음성의 왜곡이 크게 발생하기 때문이다.

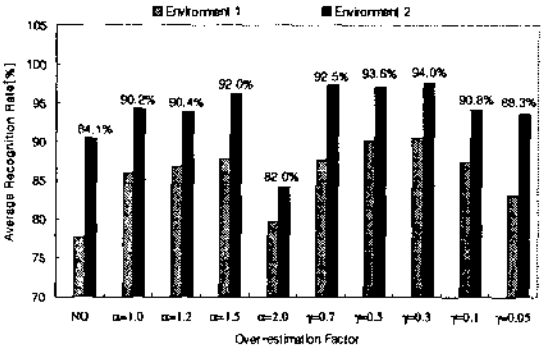


그림 5. 기존의 over-estimation 및 히스토그램 기반의 over-estimation 방식에 따른 인식결과 요약

Fig. 5. Summary of recognition results according to standard over-estimation and histogram-based over-estimation method.

표 2 및 그림 5의 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 분포특성을 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 92.0%였는데 반하여, 히스토그램 기반의 over-estimation 방식을 도입함으로써 평균 인식률이 94.0%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.

감사의 글

본 논문에서는 한국전자통신연구원이 구축한 부서명 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. J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1996.
3. S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, John Wiley & Sons Ltd., 1996.
4. H. G. Hirsch and C. Ehricher, "Noise estimation techniques for robust speech recognition," in Proc. IEEE ICASSP-95, pp. 153-156, May 1995.
5. 권영욱, 김형순, "히스토그램 처리방법에 의한 잡음 스펙트럼 추정을 이용한 잡음환경에서의 음성인식," 한국음향학회 논문집, 제16권 5호, pp.68-75, 1997년 7월.
6. 권영욱, 김형순, "히스토그램 처리방법을 이용한 시변 잡음환경에서의 음성인식," 한국음향학회 논문집, 제17권 3호, pp. 47-51, 1998년 4월.
7. 이영직 외, "ETRI의 음성 데이터베이스 구축 현황," 제12회 음성통신 및 신호처리 워크샵 논문집, pp. 265-267, 1995년 6월.
8. 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~1998년 8월 : 부산대학교 대학원 전자공학과 (공학박사)
1987년 10월~1999년 10월 : 부경대학교 전자공학과 조교
한국음향학회지 제16권 제5호 참조

▲ 김 형 순(Hyung Soon Kim)
1983년 2월 : 서울대학교 전자공학과(공학사)
1984년 2월 : 한국과학기술원 전기 및 전자공학과(석사과정 조기진학)
1989년 2월 : 한국과학기술원 전기 및 전자공학과(공학박사)
1987년 1월 ~ 1992년 6월 : (주)디지콤 부설 정보통신연구소 연구부장
1992년 7월 ~ 현재 : 부산대학교 전자공학과 조교수
부산대학교 정보통신연구소 연구원
1999년 8월 ~ 현재 : 카네기멜론대학 객원연구원
한국음향학회지 제16권 제5호 참조