히스토그램 처리방법을 이용한 시변 잡음환경에서의 음성인식

Speech Recognition in Time-varying Noisy Environments using the Histogram Technique

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

요 약

잡음 환경에서의 음성인식을 위해서는 일반적으로 전처리 과정에서 잡음의 스펙트럼을 잘 추정할 필요가 있다. 본 논문에서는 시변잡음 환경에서 히스토그램 처리방법에 의해 잡음의 스펙트럼을 추정하고 이를 제거하는 방법으로 스펙트럼 차감법을 사용하였다. 히스토그램 처리방법은 음성/비음성 구간의 구분을 할 필요가 없으며 서서히 변화하는 잡음의 스펙트럼도 추정할 수 있다는 점에서 기존 방식에 비해 장점을 지닌다. 다양한 SNR 조건하에서 시간에 따라 에너지, 그리고 주파수가 변화하는 유색 가우시안 잡음을 부가시킨 음성에 대해, 화자독립 고립단어 인식실험을 수행하였다. 실험결과, 히스토그램 처리방법에 기반을 둔 스펙트럼 차감법을 적용할 경우가 기존의 잡음 스펙트럼 추정방법에 비해 인식성능이 우수하였다.

ABSTRACT

Preprocessing techniques for speech recognition in additive noise environments usually require a good estimate of the noise power spectrum. In this paper, we choose spectral subtraction as preprocessing method, and employ the histogram technique for the estimation of noise spectra in time-varying noisy environments. This technique has advantages over other noise estimation methods in that it does not requires speech/non-speech detection. It can also estimate even time-varying noise spectra. We perform speaker-independent isolated word recognition in time-varying colored Gaussian noise environments under various SNR conditions. Experimental results show that histogram-technique-based spectral subtraction method yields superier performence to that with conventional noise estimation method.

I. 서 론

잡음환경에서의 음성인식을 위한 처리방식으로는 전처리과정을 통해 잡음을 제거하는 음질개선 방식과 청각기관의 모델 등에 근거한 잡음에 강인한 음성특징 추출방식, 잡음에 강인한 거리측정법, 그리고 이미 만들어진 모델을 토대로 환경에 따라 파라미터를 보상하는 모델적응화 기법 등이 있다[1][2]. 이들 방식 중에서 음질개선방식은 음성인식 시스템의 전처리과정에서 잡음을 제거함으로 기존의 인식시스템의 구조를 변화시키지 않고 처리가능하다는 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 차감법과 Wiener 필터링 방법이 대표적인 방법으로 알려져 있다[2][3][4]. 이들 두 방법 모두 잡음 스펙트럼의 추정치를 이용하며, 일반적으로 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. 이러한 경우, 시간에 따라 서서히 변화하는 시변 잡음에 대한 대처가 곤란하다. Hirsch는 음성/비음성 구간의 구분없이 잡음 스펙트럼을 효과적으로 추정하는 히스토그램 처리방법을 제안하였다[5]. 이 방법은 또한 시간에 따라 서서히

*부산대학교 전자공학과
접수일자 : 1998년 1월 22일

변화하는 잡음의 스펙트럼도 추정할 수 있다는 점에서 여타의 잡음 추정방법에 비해 장점을 지닌다. 본 논문의 선행연구로서 시불변 잡음환경에서 히스토그램 처리방법으로 잡음을 추정하여 화자독립 음성인식의 성능을 향상시키는 연구결과가 발표된 바 있다[6]. 본 논문에서는 히스토그램 처리방법이 시변 잡음환경에서 효과적인지를 검토하기 위하여 시간에 따라 특성이 변화하는 유색 가우시안 잡음환경에서의 인식실험을 수행하였다. 음성특징 파라미터로는 선행연구에서 사용한 인지선형예측 계수(Perceptual Linear Prediction(PLP)) 대신에 잡음환경에 보다 우수한 인식 성능을 나타내는 Mel 주파수 켑스트럼 계수(Mel-scaled Frequency Cepstral Coefficient(MFCC))를 사용하였다. 실험결과, 히스토그램 처리방법에 기반을 둔 스펙트럼 차감법이 기존의 비음성구간의 스펙트럼 평균치에 의한 방법보다 시변 잡음환경에서 우수한 성능을 나타냄을 확인하였다.

Ⅱ. 잡음환경에서의 전처리과정

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

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

여기서, $x(m)$은 잡음이 섞이지 않은 원래의 음성신호이고 $n(m)$은 부가잡음이다. 일반적으로 $x(m)$과 $n(m)$은 상관성이 없다고 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 각각의 주파수 대역 f_k에 대해 다음과 같은 관계식이 성립한다.

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

여기서 f_k는 k번째의 subband를 나타낸다.

따라서, 식 (2)에서 잡음 성분인 $N(f_k)$를 먼저 추정하고 이를 잡음이 섞인 $Y(f_k)$에서 빼줌으로써, 잡음이 없는 원음성 $X(f_k)$를 추정할 수 있다. 이 작업은 특징벡터를 추출하는 전처리 과정에서 수행하게 되며, 결과적으로 인식을 위한 특징벡터는 잡음이 없는 깨끗한 환경에서 얻어진 특징벡터와 유사하게 된다. 이때 사용되는 특징벡터는 PLP 계수 또는 MFCC 등이 주로 사용되고 있으며, MFCC가 보다 우수한 인식성능을 나타내는 것으로 알려져 있다[2]. MFCC는 기존의 켑스트럼 계수의 변형된 형태로서, 심리음향학적 지식에 기반을 둔 청각 스펙트럼 특성을 비선형 주파수 눈금인 mel-scale로 근사화 한 것이다. 그리고 잡음을 제거하기 위해서 비교적 연산이 간단하면서도 잡음환경에서 상당한 효과가 있는 스펙트럼 차감법이 주로 사용된다[3][6]. 따라서 본 논문에서는 히스토그램 처리방법으로 잡음을 먼저 추정하고, 이를 제거하기 위하여 스펙트럼 차감법을 전처리 과정에서 사용하였으며, 인식을 위한 특징벡터로서 잡음환경에서 우수한 성능을 내는 MFCC를 사용하였다.

Ⅲ. 잡음 스펙트럼의 추정

Hirsch에 의해 제안된 히스토그램 처리방법은 특정 주파수 대역(subband)에서의 잡음에 대한 진폭 스펙트럼의 통계적인 특성을 이용하여 잡음의 스펙트럼을 추정하는 방법으로, 다음과 같은 관찰결과에 근거를 두고 있다[5].

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

(2) 잡음 음성신호의 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 분포 밀도함수의 값이 최대로 되는 진폭 스펙트럼을 해당 주파수 대역에서의 잡음레벨이라 판정한다[5][6].

본 논문의 잡음추정 실험에서는 세 가지 형태의 유색 가우시안 잡음을 모의 발생시켜 사용하였다. 그 중 첫 번째는 시불변 잡음으로서 백색 가우시안 잡음을 입력으로 1.25KHz의 중심주파수 및 700Hz의 대역폭을 갖는 대역필터를 통과시켜서 만들었다. 두 번째는 필터특성은 동일하나 이득이 시간에 따라 변화하는 시변잡음으로서, 각 단어음성의 시작 부분에서부터 끝부분까지 단구간 잡음 에너지가 두배가 되도록 선형적으로 증가시킨 것이다. 그림 1(b)에 시간에 따라 에너지가 서서히 변화하는 유색잡음의 스펙트럼 분포를 나타내었다.

그리고 세 번째는 역시 시변 잡음으로서, 백색 가우시안 잡음을 대역폭이 700Hz이고 중심주파수가 프레임당 약 27Hz의 속도로 0Hz에서 4kHz까지 선형적으로 증가하는 특성을 가지는 시변 대역필터에 통과시켜 만든 유색 가우시안 잡음이다. 이때 0~4KHz까지의 가변 주파수 범위를 적용할 때 시간적인 변화구간은 0~3.75sec(8KHz 샘플링에서 30,000 샘플)의 범위가 된다. 이 중에서 실제 원음성에 부가될 때는 음성의 시작과 끝에 따라 생성된 잡음의 시작에서 원음성의 시간 길이만큼 잡음의 각 샘플을 부가하여 잡음음성을 구성하였다. 그림 2(b)에 시간에 따라 중심주파수가 선형적으로 변화하는 유색잡음을 Mel-

scaled 대역 필터를 거친 후의 결과를 나타낸 것이다.

본 논문에서의 부가잡음 추정은 시불변 잡음 및 그림 1(b) 및 그림 2(b)의 두가지 시변 잡음환경에 대하여 별도의 실험을 수행하였다. 입력 음성의 샘플링 주파수는 8kHz이고, 분석구간을 20msec로하여 10msec씩 이동하면서 특징벡터를 추출하였다. 그리고 그림 1 및 그림 2의 두가지 잡음에 대한 모의 실험을 위해 컴퓨터에서 발생시킨 잡음을 각 SNR에 따라 입력 음성에 더하고, 이를 잡음음성에서 잡음을 추정하는 시스템의 입력으로 사용하였다.

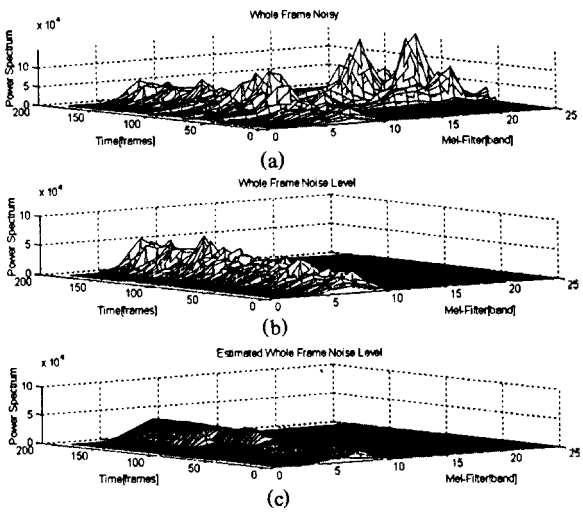

그림 1. 시간에 따라 에너지가 변화하는 유색잡음에 대한 히스토그램 처리방법을 이용한 잡음추정의 예
(a) 5dB SNR로 잡음이 더해진 잡음음성의 스펙트럼
(b) 잡음의 스펙트럼
(c) 잡음음성으로부터 추정한 잡음레벨

Fig. 1 Examples of noise estimation by histogram processing for the colored noise with time-varying energy;
(a) Spectra of noisy speech with additive noise of 5dB SNR,
(b) Spectra of original noise,
(c) Estimated noise spectra from the noisy speech.

그림 1 및 그림 2는 단어 "재무관리실"에 대하여 시간에 따라 에너지가 변화하는 경우와 시간에 따라 주파수가 변화하는 경우에 대해서 잡음음성 및 모의 발생된 시변잡음 그리고 추정한 잡음레벨을 나타내었다. 그림 1은 시간에 따라 에너지가 서서히 변화하는 유색 가우시안 잡음[그림 1(b)]이 포함된 잡음음성[그림 1(a)]에서의 추정된 잡음[그림 1(c)]의 예를 보인 것이다. 추정된 잡음의 레벨은 모의 발생된 잡음의 스펙트럼 레벨 평균과 거의 일치하고 있음을 잘 보여주고 있다. 그러나, 잡음신호는 프레임간의 편차가 매우 큰 특성을 가지므로, 잡음레벨로 추정된 잡음을 스펙트럼 차감법에 의해 빼주는 것만으로는 잡음의 영향이 충분히 제거되지 않으며, 또한 상당수의 프레임에서 음의 값이 나오게 된다. 실제로는 overestimation factor와 flooring factor 등을 도입하여 이러한 문제를 극복하게 되며, 본 논문에서도 이를 이용하였다. 그림 2에서는 시간에 따라 주파수가 변화하는 유색 가우시안 잡음[그림 2(b)]이 포함된 잡음음성[그림 2(a)]에서의 추정된 잡음[그림 2(c)]의 예를 보인 것이다. 잡음음성에서 시간에 따라 주파수가 변화하는 경우에는 특정 밴드에서 잡음 및 음성의 에너지 분포가 유사하여 구분이 어려운 구간이 간혹 발생하는 경우가 있으나, 비교적 잡음의 스펙트럼을 잘 추정하고 있음을 그림에서 보여주고 있다.

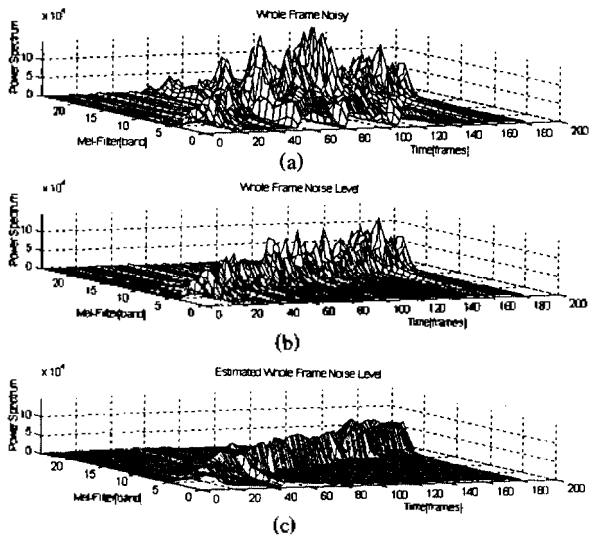

그림 2. 시간에 따라 중심주파수가 변화하는 유색잡음에 대한 히스토그램 처리방법을 이용한 잡음추정의 예
(a) 5dB SNR로 잡음이 더해진 잡음음성의 스펙트럼
(b) 잡음의 스펙트럼
(c) 잡음음성으로부터 추정한 잡음레벨

Fig. 2 Examples of noise estimation by histogram processing for the colored noise with time-varying center-frequency;
(a) Spectra of noisy speech with additive noise of 5dB SNR,
(b) Spectra of original noise,
(c) Estimated noise spectra from the noisy speech.

Ⅳ. 인식실험 결과 및 고찰

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.

본 논문에서의 음성인식 시스템은 12차의 MFCC 및 12차의 델타 MFCC 계수들을 특징 파라미터로 사용하였으며, 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. 잡음환경에서의 음성인식 전처리 방법으로는 스펙트럼 차감법에 대해 별도로 실험을 수행하였다. 전처리 과정에서 잡음의 추정방법으로는 미리 비음성구간으로 확인된 초기 10 프레임의 잡음평균을 이용하는 방법과 히스토그램 처리방법을 사용하였다.

표 1. 시불변 잡음환경에서의 인식 결과
Table 1. Recognition results in stationary noise environments.

Preprocessing Technique	Noise Estimation	Accuracy[%]				
		Clean	30 dB	20 dB	10 dB	5 dB
NO	NO	99.4	97.0	95.2	88.5	71.8
Spectral Subtraction	Frame Average	98.8	98.5	98.2	97.3	96.1
	Histogram	98.8	98.8	98.5	96.7	94.2

표 1은 시불변 유색잡음 환경에 대한 인식 결과를 나타낸 것이며, 잡음생성 방법은 3장에 기술된 바와 같다. 초기 프레임의 평균에 의해 잡음 스펙트럼을 추정하는 방법에서 초기 프레임은 미리 수작업에 의해 비음성 구간으로 확인된 처음 10프레임을 사용하였다. 표 1의 시불변 유색잡음 환경에서는 잡음의 정적인 특성으로 인해 초기 프레임 평균에 의한 잡음처리 방법이 낮은 SNR에서도 우수한 인식성능을 보여주고 있다.

표 2. 시변 잡음환경에서의 인식 결과
Table 2. Recognition results in time-varying noise environments:
(a) 시간에 따라 에너지가 증가하는 유색잡음
(a) Colored noise with increasing energy according to time,

Preprocessing Technique	Noise Estimation	Accuracy[%]				
		Clean	30 dB	20 dB	10 dB	5 dB
NO	NO	99.4	96.7	95.8	84.6	58.2
Spectral Subtraction	Frame Average	98.8	98.5	98.5	96.1	87.6
	Histogram	98.8	98.8	98.8	96.1	91.5

(b) 시간에 따라 주파수가 증가하는 유색잡음
(b) Colored noise with increasing frequency according to time.

Preprocessing Technique	Noise Estimation	Accuracy[%]				
		Clean	30 dB	20 dB	10 dB	5 dB
NO	NO	99.4	99.1	97.6	88.5	72.7
Spectral Subtraction	Frame Average	98.8	98.2	95.2	87.3	63.9
	Histogram	98.8	98.8	98.5	94.6	88.5

표 2는 시간에 따라 서서히 변화하는 유색잡음 환경에서의 인식 결과를 나타낸 것이다. 표 2(a)는 시간에 따라 에너지가 서서히 변화하는 유색 가우시안 잡음이 부가된 경우의 실험 결과이며 잡음생성 방법은 3장에 기술된 바와 같다. 표에서 잡음의 영향이 심할 경우(5dB SNR) 히스토그램 처리방법에서 현저한 성능개선을 보이고 있다. 그리고 표 2(b)는 시간에 따라 주파수가 서서히 증가하는 유색잡음 환경에서의 인식결과를 나타낸 것이다. 역시 잡음생성 방법은 3장에 기술된 바와 같다. 표 2(a)에서와 유사한 경향을 보이고 있으나 기존의 초기 프레임 평균 방법에 의한 잡음처리의 경우는 오히려 잡음처리를 하지 않은 경우보다도 나빠지는 경향을 보이고 있다. 이는 시간에 따라 스펙트럼 형태가 변화되는 잡음에 대해 일부 구간에 국한된 잡음추정을 함에 기인한 당연한 결과이다. 이에 비해 표 2(a)에서는 잡음 스펙트럼의 크기만 변할 뿐 형태 자체에는 변화가 없는 경우이므로 기존의 초기 프레임 평균방법에 의해서도 성능개선이 이루어졌다. 이상에서 히스토그램 처리방식에 의해 추정한 잡음 스펙트럼을 이용하여 스펙트럼 차감법을 적용한 경우가 초기 프레임 평균방법에 의해 추정한 잡음 스펙트럼을 이용하여 스펙트럼 차감법을 적용한 경우보다 시변 잡음환경에서의 인식 성능이 개선됨을 볼 수 있다.

V. 결 론

본 논문에서는 부가잡음에 강인한 음성인식 시스템을 개발하기 위하여 현재 전처리 단계에서 많이 이용되고 있는 스펙트럼 차감법을 사용하고, 시변잡음 스펙트럼의 추정방법을 검토하였다. 기존의 방식은 정적인 잡음의 경우에 있어서는 비교적 우수한 성능을 나타내고 있으나 시변 잡음환경에서는 성능저하를 초래한다. 이에 대한 대책으로 동적인 잡음에도 적용 가능한 히스토그램 처리방식으로 잡음 스펙트럼을 추정함으로써 인식성능의 개선을 도모하였다. 시간에 따라 에너지 또는 주파수가 변화하는 유색 가우시안 잡음을 이용한 잡음환경에서의 화자독립 고립단어 인식실험 결과, 히스토그램 처리방법에 의해 잡음 스펙트럼을 추정한 경우가 초기 프레임의 평균으로 잡음 스펙트럼을 추정한 경우보다 우수한 성능을 나타내었다.

※본 논문에서는 한국전자통신연구원이 구축한 부서명 음성 데이터베이스의 일부를 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, 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. S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, John Wiley & Sons Ltd., 1996.

5. H. G. Hirsch and C. Ehricher, "Noise estimation techniques for robust speech recognition," in Proc. IEEE ICASSP-95, pp. 153-156, May 1995.
6. 권영욱, 김형순, "히스토그램 처리방법에 의한 잡음 스펙트럼 추정을 이용한 잡음환경에서의 음성인식," 한국음향학회 논문집, 제16권 5호, pp. 68-75, 1997년 7월.
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~현재 : 부산대학교 대학원 전자공학과(박사과정)
1987년 10월~현재 : 부경대학교 공과대학 전자공학과 조교

▲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 : (주)디지콤 부설 정보통신연구소 연구부장
1992년 7월~현재 : 부산대학교 전자공학과 조교수