한국음향학회지 제23권 제2호 pp. 180~186 (2004)

강인한 음성인식을 위한 통계적 특징벡터 추출방법의 개선

An Improvement of Stochastic Feature Extraction for Robust Speech Recognition

김 회 린*, 고 진 석*
(Hoi-Rin Kim*, Jin-Seok Ko*)

*한국정보통신대학교 공학부 음성인식기술연구실

(접수일자: 2003년 10월 24일; 수정일자: 2004년 1월 20일; 채택일자: 2004년 2월 18일)

음성 신호에 존재하는 잡음은 음성 인식기의 성능을 현저하게 감소시킨다. 이것은 잡음이 훈련 조건과 인식 조건 사이의 불일치를 가져오기 때문이다. 본 논문에서는 이러한 불일치를 최소화하기 위해서 통계적 특징벡터의 추출방법을 개선하기 위한 방법을 연구하였다. 밴드 SNR에 따라 잡음 스펙트럼의 차감 레벨을 조절하는 기존의 멀티 밴드 잡음 차감법 (MSS)을 개선하기 위하여 잡음 정규화 상수를 이용하여 잡음 스펙트럼의 차감 레벨을 보다 정확하게 조절하는 방법 (M-MSS)을 제시하였다. 다음으로, 기존의 통계적 특징벡터 추출방법 (SFE)에서 잡음 차감법을 파워 스펙트럼 영역에 적용함으로써 성능을 개선하였다 (M-SFE). 마지막으로, 위의 두 가지 방법의 장점을 결합하기 위해서 밴드 SNR에 근거한 통계적 특징벡터 추출방법 (MMSS-MSFE)을 제안하였다. 제안된 방법들은 다양한 잡음 환경 하에서 화자독립 고립 단어 인식으로 성능을 평가하였다. 기본적인 잡음 차감법 (SS)에 비하여 M-MSS, M-SFE와 MMSS-MSFE의 평균 에러율은 각각 18.6%, 15.1%와 33.9% 감소하였다. 위의 결과로부터 제안한 방법이 잡음에 강인한 음성인식을 위해 매우 효과적임을 입증하였다.

핵심용어: 음성인식, 음질 개선, 잡음 차감법, 통계적 특징벡터
투고분야: 음성처리 분야 (2,5)

The presence of noise in speech signals degrades the performance of recognition systems in which there are mismatches between the training and test environments. To make a speech recognizer robust, it is necessary to compensate these mismatches. In this paper, we studied about an improvement of stochastic feature extraction based on band-SNR for robust speech recognition. At first, we proposed a modified version of the multi-band spectral subtraction (MSS) method which adjusts the subtraction level of noise spectrum according to band-SNR. In the proposed method referred as M-MSS, a noise normalization factor was newly introduced to finely control the over-estimation factor depending on the band-SNR. Also, we modified the architecture of the stochastic feature extraction (SFE) method. We could get a better performance when the spectral subtraction was applied in the power spectrum domain than in the mel-scale domain. This method is denoted as M-SFE. Last, we applied the M-MSS method to the modified stochastic feature extraction structure, which is denoted as the MMSS-MSFE method. The proposed methods were evaluated on isolated word recognition under various noise environments. The average error rates of the M-MSS, M-SFE, and MMSS-MSFE methods over the ordinary spectral subtraction (SS) method were reduced by 18.6%, 15.1%, and 33.9%, respectively. From these results, we can conclude that the proposed methods provide good candidates for robust feature extraction in the noisy speech recognition.

Keywords: *Noise robust speech recognition, Speech enhancement, Spectral subtraction, Stochastic feature*
ASK subject classification: *Speech signal processing (2,5)*

I. 서 론

실제 환경에서 동작하는 음성 인식기는 잡음에 상당한 영향을 받는다. 무잡음 음성으로 훈련한 음성 인식기는 잡음 환경에서 얻은 음성을 적절하게 인식할 수 없다. 따라서 잡음으로 인해 음성 인식기의 성능은 저하된다. 이것은 잡음이 훈련 조건과 인식 조건 사이의 불일치를 가져오기 때문이다[1].

잡음에 강인한 음성 인식기를 만들기 위해 잡음으로

책임저자: 김회린 (hrkim@icu.ac.kr)
305-714 대전광역시 유성구 문지로 119
한국정보통신대학교
(전화: 042-866-6139; 팩스: 042-866-6245)

야기된 훈련 조건과 인식 조건 사이의 불일치를 보상할 필요가 있다. 이러한 불일치를 보상하기 위한 방법들은 대략 세 가지로 분류된다[2]. 첫째는 특징벡터 추출 과정에서 잡음이 섞인 음성으로부터 무잡음 음성 신호를 추정하고, 그것을 무잡음 음성 신호로 학습된 음향 모델로 인식하는 잡음 차감법을 이용한 음질 개선 방법이다. 둘째는 잡음에 둔감한 특징 파라미터를 얻는 데에 중점을 둔 잡음에 강인한 특징벡터 추출방법이다. 마지막은 무잡음 환경에서 학습된 음향 모델을 잡음 모델을 이용하여 인식 환경에 적합한 음향 모델로 변형하고, 잡음이 섞인 음성을 변형된 음향 모델을 이용하여 인식하는 모델 보상 방법이다.

본 논문에서는 훈련 조건과 인식 조건 사이의 불일치를 최소화하기 위해 잡음 차감법을 이용한 음질 개선 방법에 초점을 맞추었다. 잡음 차감법 (spectral subtraction)은 주변 잡음에 의해 손상된 음성 스펙트럼에서 추정된 잡음 스펙트럼의 크기 성분만을 제거하는 방법이다. 이 때 잡음 차감법이 음성 인식기의 성능을 향상시킨다 할지라도, 잡음 스펙트럼의 평균만 이용하기 때문에 잡음 스펙트럼의 분산을 적절하게 보상하지 못한다. 잡음의 분포를 특징벡터에 포함하여 잡음 스펙트럼의 분산을 보상하기 위해 통계적 특징벡터 추출 방법 (stochastic feature extraction)이 제안되었고, 잡음 차감법보다 더 좋은 인식 결과를 보여 주었다[3].

실제 환경의 잡음이 섞인 음성의 스펙트럼을 중첩되지 않은 몇 개의 밴드로 나누었을 때 SNR의 분포는 밴드에 따라 다르다[4]. 즉, 잡음은 음성 신호의 특정 주파수 밴드의 스펙트럼 성분만 심하게 영향을 주는 경향이 있다. 따라서, 각 밴드마다 잡음 스펙트럼의 차감 레벨을 적절하게 조절할 필요가 있다. 그러나, 기존의 통계적 특징벡터 추출 방법에서는 이러한 잡음의 영향을 정밀하게 고려하지 않는다. 그러므로, 잡음을 제거하는 동안 음성 신호의 왜곡을 줄이기 위해 멀티 밴드의 개념이 도입될 필요가 있다.

본 논문에서는 밴드 SNR에 근거한 통계적 특징벡터 추출 방법의 개선을 위한 방법을 연구하였다. 이것은 차감 레벨을 밴드 SNR에 따라 적절하게 조절함으로써 최적의 무잡음 음성을 추정하고, 잡음의 분포를 특징벡터에 포함함으로써 잡음 스펙트럼의 분산을 보상하기 위한 것이다. 제안된 방법은 다양한 잡음 환경 하에서 화자독립 고립 단어 인식으로 성능을 평가하였고, 잡음 차감법, 멀티 밴드 잡음 차감법, 통계적 특징벡터 추출 방법의 인식 성능 결과와 비교하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존의 특징벡터 보상 방법들을 간략히 소개하고, 3장에서는 밴드 SNR에 근거한 통계적 특징벡터 추출 방법의 개선 방법에 대하여 설명한다. 4장에서는 제안된 방법의 성능 평가를 위한 실험 및 결과를 정리하고, 마지막으로 본 논문의 결론을 맺도록 한다.

II. 기존의 특징벡터 보상 방법들

2.1. 잡음 차감법 (Spectral Subtraction, SS)

잡음이 섞인 음성 신호는 잡음 성분과 무잡음 음성 성분으로 구성되어 있다. 잡음과 무잡음 음성 신호 사이에 상관관계가 없고, 산술적으로 더해졌다고 가정하면, 잡음이 섞인 음성 신호 $y(n)$은 다음과 같이 표현할 수 있다[5,6].

$$y(n) = s(n) + d(n) \tag{1}$$

여기에서 $s(n)$과 $d(n)$은 각각 무잡음 음성 신호와 잡음을 나타낸다. 잡음이 섞인 음성 신호의 파워 스펙트럼 $|Y(k)|^2$은 다음과 같이 근사적으로 추정할 수 있다.

$$|Y(k)|^2 \approx |S(k)|^2 + |D(k)|^2 \tag{2}$$

여기에서 $|S(k)|^2$와 $|D(k)|^2$는 각각 무잡음 음성 신호와 잡음의 파워 스펙트럼을 나타낸다. 잡음 스펙트럼은 직접 얻을 수 없기 때문에 일반적으로 잡음이 섞인 음성 신호의 묵음 구간에서 잡음 스펙트럼의 추정치 $\hat{D}(k)$를 계산한다.

식 (2)를 기반으로 추정한 무잡음 음성의 스펙트럼은 다음과 같이 계산할 수 있다.

$$|\hat{S}(k)|^2 = |Y(k)|^2 - \alpha|\hat{D}(k)|^2 \tag{3}$$

$$|\hat{S}(k)|^2 = \begin{cases} |\hat{S}(k)|^2, & |\hat{S}(k)|^2 > \beta|\hat{D}(k)|^2 \\ \beta|\hat{D}(k)|^2, & otherwise \end{cases} \tag{4}$$

여기에서 α와 β는 각각 과추정 상수 (over-estimation factor)와 flooring 상수를 나타낸다. 과추정 상수는 프레임마다 계산되어 차감 레벨을 조절하는 역할을 담당하고, flooring 상수는 음성 신호의 스펙트럼이 기준치 아래로 떨어지는 것을 방지한다.

2.2. 멀티 밴드 잡음 차감법 (Multi-band Spectral Subtraction, MSS)

그림 1은 지하철 잡음에 의해 왜곡된 음성의 프레임별 SNR의 분포와 밴드별 SNR의 분포의 예를 보여준다. 이 예로부터, 단순히 프레임별 SNR에 따라 과추정 상수를 정하여 차감 레벨을 결정할 경우 더 큰 SNR 값을 나타내는 밴드에서는 실제보다 더 큰 차감 레벨이 적용되어 음성 왜곡을 초래하게 되고, 더 작은 SNR 값을 나타내는 밴드에서는 실제보다 더 작은 차감 레벨이 적용되어 잔류 잡음을 초래하게 됨을 확인할 수 있다. 따라서 각 주파수 밴드마다 잡음 스펙트럼의 차감 레벨을 적절하게 조절할 필요가 있다.

이러한 잡음의 특성을 고려하기 위하여 잡음 차감법에 멀티 밴드 접근법을 적용하였다. 먼저, 음성 스펙트럼을 중첩되지 않은 M개의 밴드로 나누고, 각 밴드마다 독립적으로 잡음 차감법을 적용한다.

따라서, m^{th}밴드에서 추정한 무잡음 음성 스펙트 럼은 다음과 같이 계산할 수 있다[4].

$$|\hat{S}_m(k)|^2 = |Y_m(k)|^2 - \alpha_m \delta_m |\hat{D}_m(k)|^2 b_m \le k \le e_m \quad (5)$$

여기에서 m은 밴드 인덱스 ($1 \le m \le M$)이고, b_m과 e_m은 각각 m^{th} 밴드의 시작과 끝 주파수 빈을 나타내고, α_m은 m^{th} 밴드의 과추정 상수 (over-estimation factor)를 나타내며, δ_m은 tweaking 상수를 나타낸다. m^{th} 밴드의 과추정 상수 α_m은 M개의 밴드마다 계산되는 밴드 SNR의 함수이다.

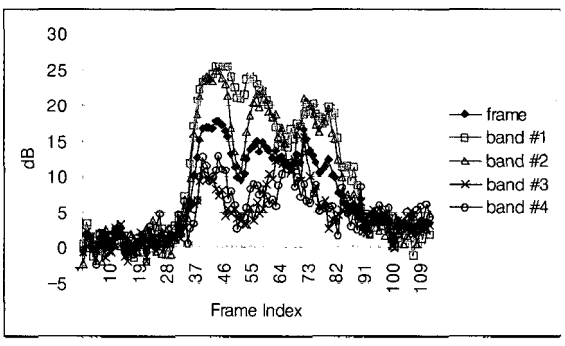

그림 1. 지하철 잡음에 의해 왜곡된 음성의 프레임별 SNR의 분포와 밴드별 SNR의 분포의 예

Fig. 1. The band-SNR examples of four frequency bands of speech corrupted by subway noise.

2.3. 통계적 특징벡터 추출 (Stochastic Feature Extraction, SFE)

식 (1)에서 잡음이 랜덤 변수이기 때문에 무잡음 음성 신호 또한 랜덤 변수로 간주할 수 있다. 따라서, 무잡음 음성 신호의 확률밀도함수 (probability density function) f는 잡음 성분의 확률밀도함수 g를 이용하여 다음과 같이 나타낼 수 있다[3,7].

$$f(s) = \begin{cases} J \cdot g(y-s), & s \in S \\ 0, & otherwise \end{cases} \quad (6)$$

여기에서 y와 s는 각각 잡음이 섞인 음성과 무잡음 음성 성분을 나타내고, J는 $\int_s f(s)ds = 1$를 만족하기 위한 정규화 상수를 나타낸다.

무잡음 음성 신호로 학습된 HMM (Hidden Markov Model)을 이용하여 잡음이 섞인 음성 신호의 관측 확률을 계산하는 것이 이 방법의 목적이다. 따라서, t^{th} 프레임에서 y_t가 관측되고, 잡음 성분의 확률밀도함수가 주어졌을 때, i^{th} state 출력 확률 $P_i(y_t, g)$은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} P_i(y_t, g) &= \int_s b_i(s) g(y_t - s) ds \\ &= \frac{1}{J_t} \cdot \int b_i(s) f_i(s) ds \end{aligned} \quad (7)$$

여기에서 b_i는 HMM에서 i^{th} state에 대한 출력 확률밀도 함수를 나타낸다. J_t는 인식 단계에서 인식 결과에 영향을 주지 않기 때문에 무시될 수 있다. 따라서, 추정한 무잡음 음성 신호에 대한 state 출력 확률 $B_i(f_t)$은 다음과 같이 나타낼 수 있다.

$$B_i(f_t) = \int_s b_i(s) f_t(s) ds \quad (8)$$

만일 f_t와 b_i가 모두 가우시안 (Gaussian) 분포를 따른다고 가정하면, 각각 $N(\xi, \Psi)$과 $N(\mu, \Sigma)$으로 표현할 수 있다. 여기에서 ξ와 Ψ는 각각 통계적 특징벡터의 평균과 분산이고, μ와 Σ는 각각 훈련 과정에서 학습된 무잡음 음성 신호의 평균과 분산이다. 결국, 추정한 무잡음 음성 신호의 i^{th} state 출력 확률은 다음과 같이 구할 수 있다.

$$\begin{aligned} B_i(N(\xi, \Psi)) &= \frac{1}{(2\pi)^{\frac{D}{2}} |\Psi + \Sigma|^{\frac{1}{2}}} \\ &\times \exp\left\{ -\frac{1}{2} (\xi - \mu)^T (\Psi + \Sigma)^{-1} (\xi - \mu) \right\} \end{aligned} \quad (9)$$

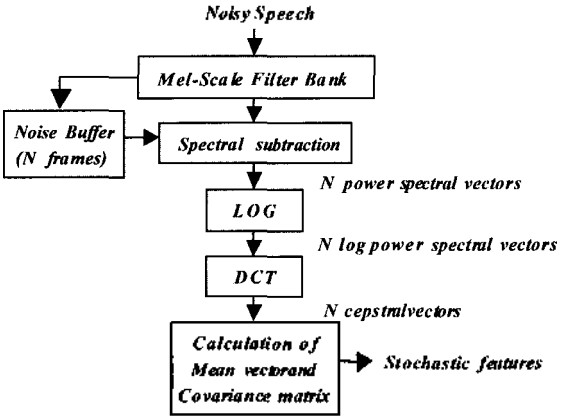

그림 2. MFCC 영역에서 통계적 특징벡터 추출방법의 구조
Fig. 2. Block diagram for the Gaussian stochastic feature extraction in the MFCC domain.

따라서, 이 방법은 잡음의 분포를 특징벡터에 적용함으로써 잡음 스펙트럼의 분산을 보상할 수 있다.

캡스트럴 영역에서 추정된 음성의 확률밀도함수에 대한 가우시안 표현을 얻기 위한 방법에는 여러 가지가 있다. 한가지 방법은 캡스트럴 영역에서 직접 평균과 분산을 계산하는 것이다. 그림 2는 MFCC 영역에서 가우시안 통계적 특징벡터 추출방법의 구조를 나타낸다.

III. 밴드 SNR에 근거한 통계적 특징벡터 추출방법의 개선

3.1 변형된 멀티 밴드 잡음 차감법 (M-MSS)

멀티 밴드 잡음 차감법의 개념을 이용하되 차감 레벨을 보다 정밀하게 조절하기 위하여 과추정 상수와 함께 잡음 정규화 상수를 사용하였다. 즉, 과추정 상수를 이용하여 밴드 SNR에 따른 차감 레벨을 결정하고, 여기에 잡음 정규화 상수를 적용하여 과추정 상수의 값을 조절함으로써 주파수에 따라 변화하는 잡음 스펙트럼의 영향을 고려한다. 이를 통해서 보다 정확하게 차감 레벨을 결정할 수 있다. 결국, m^{th} 밴드에서 추정한 무잡음 음성 스펙트럼은 다음과 같이 나타낸다.

$$|\hat{S}_m(k)|^2 = |Y_m(k)|^2 - \alpha_m \delta(k) |\hat{D}_m(k)|^2 \quad b_m \le k \le e_m \tag{10}$$

여기에서 b_m과 e_m은 각각 m^{th} 밴드의 시작과 끝 주파수 빈을 나타낸다. α_m은 m^{th} 밴드의 과추정 상수 (over-estimation factor)를 나타내고, $\delta(k)$는 주파수에 따른 잡음 정규화 상수를 나타낸다.

과추정 상수 α_m은 m^{th} 밴드의 SNR에 관한 함수이다. 과추정 상수를 계산하기 위해서 먼저 m^{th} 밴드의 SNR 값을 다음과 같이 계산한다.

$$SNR_m(dB) = 10\log_{10}\left(\frac{\sum_{k=b_m}^{e_m} |Y_m(k)|^2}{\sum_{k=b_m}^{e_m} |\hat{D}_m(k)|^2}\right) \tag{11}$$

식 (11)을 이용하여 과추정 상수는 다음과 같이 실험적으로 정의하였다.

$$\alpha_m = \begin{cases} 5 & , \quad SNR_m < -5 \\ 4.2 - \frac{4}{25}(SNR_m), & -5 \le SNR_m \le 20 \\ 1 & , \quad SNR_m > 20 \end{cases} \tag{12}$$

식 (12)에서 표현한 것처럼 과추정 상수는 각 주파수 밴드에서 잡음 스펙트럼의 차감 레벨을 조정하는 역할을 담당한다. 즉, m^{th} 밴드의 SNR 값이 크면 과추정 상수의 값을 작게 하여 잡음을 제거하는 동안 발생하는 음성 신호의 왜곡을 줄이는 데에 그 목적이 있다.

잡음 스펙트럼의 크기는 밴드 내에서 주파수에 따라 변화하기 때문에 잡음의 밴드별 변화는 물론 주파수별 변화도 고려하여 차감 레벨을 결정할 필요가 있다. 잡음의 주파수별 변화를 고려하기 위해서 잡음 정규화 상수 $\delta(k)$를 제안하였고, 다음과 같이 계산한다.

$$\delta(k) = \frac{\hat{D}(k)}{\max_k(\hat{D}(k))}, \quad 1 \le k \le \frac{FFTSIZE}{2} \tag{13}$$

잡음 정규화 상수는 0보다 크고, 1과 같거나 작은 값을 나타낸다. 만일 잡음 정규화 상수의 값이 작으면 그 주파수에서 잡음의 영향이 상대적으로 작다는 것을 의미한다. 이러한 경우, 음성 왜곡을 최소화하기 위해서 각 밴드에서 미리 정한 과추정 상수의 값을 줄일 필요가 있다. 따라서, 잡음 정규화 상수는 과추정 상수의 값을 주파수마다 조절함으로써 차감 레벨을 보다 정확하게 결정하는 데에 중요한 역할을 한다.

마지막 단계로, 추정한 무잡음 음성 신호의 스펙트럼이 기준치 (threshold) 아래로 떨어질 경우 뮤지컬 잡음이 발생한다. 이를 방지하기 위해서 flooring 상수 β를 사용하여 식 (14)처럼 무잡음 음성 신호의 스펙트럼 추정치를 얻게 된다.

$$|\hat{S}_m(k)|^2=\begin{cases}|\hat{S}_m(k)|^2, & |\hat{S}_m(k)|^2>\beta|\hat{D}_m(k)|^2\\ \beta|\hat{D}_m(k)|^2, & otherwise\end{cases} \quad (14)$$

여기에서 flooring 상수 β는 실험적으로 0.1로 정하였다.

3.2. 변형된 통계적 특징벡터 (M-SFE)

그림 2에서 보여진 것처럼 SFE 방법에서는 무잡음 음성의 스펙트럼을 추정하기 위해서 SS 방법이 mel-scale 영역에 적용되었다. 그림 3에서 보여지는 것처럼 SFE 방법의 구조를 변형하였다. 즉, SS 방법을 mel-scale 영역에 적용하지 않고 파워 스펙트럼 영역에 적용하여 무잡음 음성의 스펙트럼을 추정하였다. 이것은 무잡음 음성의 스펙트럼을 추정하기 위한 방법이 어떤 영역에서 적용될 때 더 좋은 인식 성능을 나타내는지 확인하기 위한 시도이다. 그림 3은 변형된 통계적 특징벡터 추출방법의 구조를 나타낸다.

3.3. M-MSS와 M-SFE 방법의 결합 (MMSS-MSFE)

M-MSS와 M-SFE 방법의 장점을 이용하기 위하여 밴드 SNR에 근거한 통계적 특징벡터 추출방법의 개선에 관하여 연구하였다. 그림 4는 제안된 방법의 구조를 나타낸다.

캡스트럴 static 계수를 계산하기 위한 MMSS-MSFE 방법의 절차는 다음과 같이 요약된다.

- 화자가 단어를 발성하면, 발성된 단어의 묵음 구간에서 N 프레임의 잡음 파워 스펙트럼을 각각 계산하고, 이것을 잡음 버퍼에 저장한다.
- 멀티 밴드 잡음 차감법을 이용하여 프레임마다 잡음이 섞인 음성 신호의 파워 스펙트럼으로부터 잡음 버퍼에 저장되어 있는 N개의 잡음 파워 스펙트럼을 각각 빼준다.
- 위의 과정에서 계산된 N개의 추정한 무잡음 음성의 파워 스펙트럼을 mel-scale 필터 뱅크에 적용하여 mel-scale 영역에서 N개의 파워 스펙트럼을 얻는다.
- N개의 파워 스펙트럼에 로그를 취하고 DCT (Discrete Cosine Transform)를 적용하여 N개의 cepstral vector로 변환한다.
- N개의 cepstral vector를 이용하여 각 프레임의 평균과 분산을 계산한다.
- 프레임마다 통계적 특징벡터를 얻게 된다.

IV. 실험 및 결과

제안된 방법들은 다양한 잡음 환경에서 화자독립 고립단어인식 실험으로 평가하였다. 실험에 사용한 음성 DB는 PBW (Phonetically Balanced Word) 452이고, 이것은 70명 (남자: 38명, 여자: 32명)의 화자가 각각 2회씩 발성한 단어로 구성되어 있다. 9:1의 비율로 63명의 화자가 발성한 56,952개의 단어를 훈련 데이터로 사용하였고, 나머지 7명의 화자가 발성한 6,328개의 단어를 테스트 데이터로 사용하였다. 음성 신호는 16 KHz로 샘플링 되었고, 16 bit로 양자화 되어있다. 하지만, 8 kHz로 샘플링 되어 있는 AURORA 잡음 DB를 음성 신호에 첨가하기 위하여 음성 신호를 8 kHz로 다운 샘플링하였다. 잡음 DB

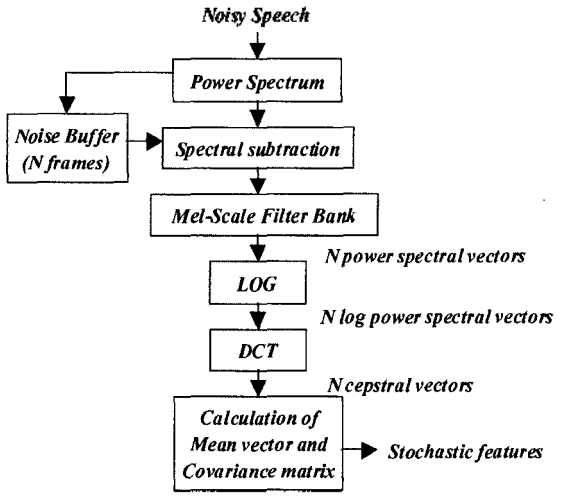

그림 3. 변형된 통계적 특징벡터 추출방법의 구조
Fig. 3. Block diagram for the modified stochastic feature extraction method (M-SFE).

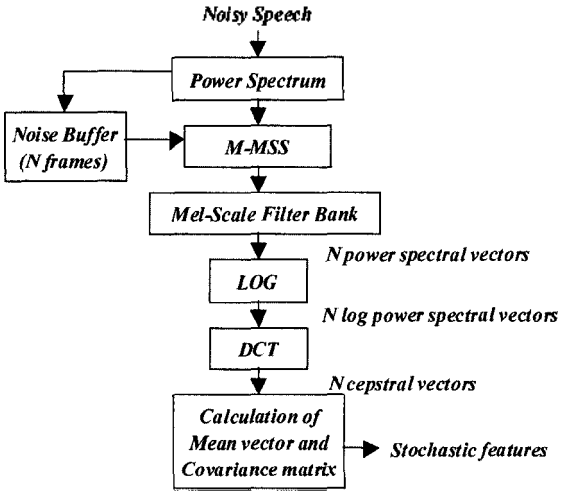

그림 4. 제안된 MMSS-MSFE 방법의 구조
Fig. 4. Block diagram for the combined method (MMSS-MSFE).

는 AURORA DB 중 babble, car, exhibition과 subway 잡음을 사용하였다.

음성 신호는 20 ms 단위의 프레임을 10 ms 단위씩 이동하여 특징벡터를 추출하였고, 각 프레임마다 총 39차 특징벡터로 표현하였다. 특징벡터는 12차 MFCC (Mel Frequency Cepstral Coefficients)와 로그에너지, delta 및 delta-delta로 구성되어 있다. 본 실험에서의 음향모델은 triphone을 사용하였고, 3 state의 left-to-right 구조의 연속밀도 HMM (Hidden Markov Model) 기반으로 하였다.

각각 녹음된 무잡음 음성 신호와 잡음을 이용하여 다양한 SNR을 가지는 잡음이 섞여 있는 음성 신호를 만들어 HTK로 실험하는데 사용하였다. 잡음 스펙트럼을 추정하기 위해 20 프레임의 묵음 구간을 이용하였다. 또한, 음성 스펙트럼을 4개의 중첩되지 않은 밴드로 나누었고, 각 밴드는 32개의 주파수 빈으로 구성된다.

표 1은 보상 방법을 적용하지 않은 base와 SS, MSS, M-MSS를 적용한 실험의 인식 성능 비교를 보여준다. 표 1로부터, 보상 방법을 적용하였을 때 더 좋은 인식 성능을 얻을 수 있음을 알 수 있다. 또한, M-MSS 방법이 20dB의 지하철과 전시회 잡음 환경을 제외한 모든 잡음 환경에 대해 SS와 MSS 방법보다 더 좋은 인식 성능을 나타냄을 알 수가 있다. M-MSS 방법에서, 과추정 상수는 밴드에 따른 잡음 스펙트럼의 변화를 반영하고 잡음 정규화 상수는 밴드 내에서 주파수에 따른 잡음 스펙트럼의 변화를 반영한다. 차감 레벨을 보다 적절하게 적용함으로써 최적의 음성 스펙트럼을 추정하기 때문에 다른 방법들보다 더 좋은 성능을 나타낸다. 특히, SS 방법을 기준으로 20dB와 10dB에 대한 M-MSS 방법의 평균 에러 감소율 (Error Reduction Rate)은 18.7%를 보여 주었다.

표 1. Baseline과 SS, MSS, M-MSS를 적용한 실험의 인식 성능 비교 (%)

Table 1. Recognition accuracies of baseline, SS, MSS, and M-MSS method (%).

Noise	SNR	Base	SS	MSS	M-MSS
Clean		99.12	99.12	99.12	99.12
Babble	20	97.90	98.12	98.12	98.34
	10	86.17	88.27	89.16	90.71
Car	20	98.12	98.12	98.23	98.23
	10	84.62	91.15	91.59	92.37
Exhibition	20	90.38	94.80	97.79	97.12
	10	59.40	77.77	79.65	80.31
Subway	20	94.80	96.35	97.35	97.12
	10	65.04	78.43	79.98	82.85

표 2는 SS와 SFE, M-SFE를 적용한 실험의 인식 성능 비교를 보여준다. 표 2로부터 M-SFE 방법이 모든 잡음 환경에 대해 SFE 방법보다 더 좋은 인식 성능을 나타냄을 알 수 있다. 이를 통해 제안된 방법의 구조에서 M-MSS 방법을 파워 스펙트럼 영역에 적용하기로 결정하였다. 또한 M-SFE 방법이 SS 방법보다 무잡음 환경과 10dB의 전시회 잡음 환경을 제외한 모든 잡음 환경에 대해 더 좋은 인식 성능을 나타냄을 볼 수 있다. 통계적 특징벡터의 평균과 분산을 모두 이용하여 잡음 스펙트럼의 분산을 보상하기 때문에 더 좋은 인식 성능을 나타낸다. 특히, SS 방법을 기준으로 20 dB와 10 dB에 대한 M-SFE 방법의 평균 에러 감소율 (Error Reduction Rate)은 15.1%를 보여 주었다.

표 3은 SS와 M-MSS, M-SFE, MMSS-MSFE를 적용한 실험의 인식 성능 비교를 보여준다. 표 3으로부터, 제안

표 2. SS와 SFE, M-SFE (통계적 특징벡터의 평균과 분산을 모두 이용)를 적용한 실험의 인식 성능 비교 (%)

Table 2. Recognition accuracies of SS, SFE, and M-SFE method using both the mean and variance of stochastic features (%).

Noise	SNR	SS	SFE	M-SFE
Clean		99.12	99.10	99.10
Babble	20	98.12	98.01	98.34
	10	88.27	88.94	89.82
Car	20	98.12	98.01	98.12
	10	91.15	93.14	93.25
Exhibition	20	94.80	97.12	97.12
	10	77.77	76.88	77.21
Subway	20	96.35	97.12	97.23
	10	78.43	78.98	79.76

표 3. SS와 M-MSS, M-SFE, MMSS-MSFE (통계적 특징벡터의 평균과 분산을 모두 이용)를 적용한 실험의 인식 성능 비교 (%)

Table 3. Recognition accuracies of SS, M-MSS, M-SFE, and MMSS-MSFE using both the mean and variance of stochastic features (%).

Noise	SNR	SS	M-MSS	M-SFE	MMSS-MSFE
Clean		99.12	99.12	99.10	99.10
Babble	20	98.12	98.34	98.34	98.67
	10	88.27	90.71	89.82	92.37
Car	20	98.12	98.23	98.12	98.23
	10	91.15	92.37	93.25	93.36
Exhibition	20	94.8	97.12	97.12	97.90
	10	77.77	80.31	77.21	86.06
Subway	20	96.35	97.12	97.23	97.79
	10	78.43	82.85	79.76	87.06

표 4. SS 방법을 기준으로 20 dB와 10 dB에 대한 제안된 방법의 평균 에러 감소율 (%)
Table 4. Average ERR of the proposed methods using both the mean and variance over SS method in 20 dB and 10 dB (%).

Noise	M-MSS	M-SFE	MMSS-MSFE
Babble	16.3	12.5	32.1
Car	9.8	11.9	15.4
Exhibition	28	21.1	48.5
Subway	20.8	15.1	39.7
Total	18.6	15.1	33.9

된 방법인 MMSS-MSFE 방법이 무잡음 환경을 제외한 모든 잡음 환경에서 SS나 M-MSS, M-SFE보다 더 좋은 인식 성능을 나타냄을 알 수 있다. 이것은 제안된 방법이 M-MSS 방법의 장점과 M-SFE 방법의 장점을 모두 반영하기 때문임을 알 수 있다. 특히, SS 방법을 기준으로 20dB와 10dB에 대한 MMSS-MSFE 방법의 평균 에러 감소율은 33.9%를 보여 주었다.

마지막으로, 표 4는 SS 방법을 기준으로 20dB와 10dB에 대한 제안된 방법의 평균 에러 감소율을 보여준다. 이러한 결과로부터, 제안한 방법이 잡음에 강인한 음성인식을 위한 좋은 방법을 제공함을 알 수 있다.

V. 결 론

본 논문에서는 잡음에 강인한 음성인식을 위해 밴드 SNR에 근거한 통계적 특징벡터 추출방법을 개선하기 위한 방법을 연구하였다. 제안된 방법은 과추정 상수와 잡음 정규화 상수를 이용하여 차감 레벨을 적절하게 조절함으로써 잡음을 제거하는 동안 음성 왜곡을 최소화하고, 통계적 특징벡터의 평균과 분산을 모두 이용하여 잡음 스펙트럼의 분산을 보상할 수 있었다. 실험 결과로부터 제안된 방법이 다른 보상 방법들보다 더 좋은 성능을 나타내었다. 특히, SS 방법을 기준으로 20 dB와 10 dB에 대한 평균 에러 감소율을 정리하면 다음과 같다.

- M-MSS의 에러 감소율: 18.6%
- M-SFE의 에러 감소율: 15.1%
- MMSS-MSFE의 에러 감소율: 33.9%

제안된 방법에서 통계적 특징벡터의 평균과 분산을 얻기 위해 요구되는 계산량은 상당히 크다. 따라서, 제안된 방법의 에러 감소율은 유지하면서 계산량을 줄이기 위한 연구가 필요하다. 좋은 성능을 얻기 위해 과추정 상수를 실험적으로 결정하였다. 만일 인식 환경이 본 논문에서 사용된 잡음 환경과 다를 경우 과추정 상수의 값을 바꾸어야 하는 경우가 발생할 수 있다. 따라서 차감 레벨을 조절하기 위한 상수들을 잡음 환경에 독립적으로 정하는 것이 중요하다. 마지막으로, 강력한 잡음 제거 방법인 AURORA front-end 방법과 비교 및 결합 방법에 대해서 연구할 것이다

참 고 문 헌

1. Gong Y., "Speech recognition in noisy environments: A survey," *Speech Communication*, 16, 261-292, 1995.
2. 오영환, "음성언어정보처리", 홍릉과학출판사, 1998.
3. N. Iwahashi, H. Pao, H. Honda, K. Minamono, and M. Omoto, "Stochastic features for noise robust speech recognition," *ICASSP*, 2, 633-636, 1998.
4. Sunil D. Kamath, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," Master thesis, University of Texas at Dallas, 2001.
5. Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, 27, 113-120, 1979.
6. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *ICASSP*, 208-211, 1979.
7. Nestor Becerra Yoma, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE Trans. Speech and Audio Processing*, **10** (3), Mar., 2002.
8. Nathalie Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Processing*, 7 (2), 1999.

저자 약력

● 김 회 린 (Hoi-Rin Kim)

1984년 2월: 한양대학교 전자공학과 (공학사)
1987년 2월: 한국과학기술원 전기및전자공학과 (공학석사)
1992년 2월: 한국과학기술원 전기및전자공학과 (공학박사)
1987년 10월~1999년 12월: 한국전자통신연구원 선임연구원
1994년 6월~1995년 5월: 일본 ATR-ITL 방문연구원
2000년~현재: ICU 공학부, 조교수
※ 주관심분야: 음성인식, 화자인식, 오디오 인덱싱, 음성처리 기술의 통신망 응용, 음성언어 번역

● 고 진 석 (Jin-Seok Ko)

2002년 2월: 광운대학교 전자공학부 (학사)
2002년 3월~2004년 2월: 한국정보통신대학교 (공학석사)
2004년 2월~현재: LG전자
※ 주관심분야: 잡음에 강인한 음성 인식, 음질 개선