한국음향학회지 제21권 제6호 pp. 552~558 (2002)

다중대역 음성인식을 위한 부대역 신뢰도의 추정 및 가중

Estimation and Weighting of Sub-band Reliability for Multi-band Speech Recognition

조 훈 영*, 지 상 문**, 오 영 환*
(Hoon-Young Cho*, Sang-Mun Chi**, Yung-Hwan Oh*)

*한국과학기술원 전자전산학과, **경성대학교 정보과학부

(접수일자: 2002년 4월 8일; 채택일자: 2002년 7월 25일)

최근에 Fletcher의 HSR (human speech recognition) 이론을 기초로 한 다중대역 (multi-band) 음성인식이 활발히 연구되고 있다. 다중대역 음성인식은 주파수 영역을 다수의 부대역으로 나누고 별도로 인식한 뒤 부대역들의 인식결과를 부대역 신뢰도로 가중 및 통합하여 최종 판단을 내리는 새로운 음성인식 방식으로서 잡음환경에 특히 강인하다고 알려졌다. 잡음이 정상적인 경우 무음구간의 잡음정보를 이용하여 부대역 신호대 잡음비(SNR)를 추정하고 이를 가중치로 사용하기도 하였으나, 비정상 잡음은 시간에 따라 특성이 변하여 부대역 신호대 잡음비를 추정하기가 쉽지 않다. 본 논문에서는 깨끗한 음성으로 학습한 은닉 마코프 모델과 잡음음성의 통계적 정합에 의해 각 부대역에서 모델과 잡음음성 사이의 거리를 추정하고, 이 거리의 역을 부대역 가중치로 사용하는 ISD (inverse sub-band distance) 가중을 제안한다. 1500 ~ 1800 Hz로 대역이 제한된 백색잡음 및 클래식 기타음에 대한 인식 실험 결과, 제안한 방법은 정상 및 비정상 대역제한 잡음에 대하여 부대역의 신뢰도를 효과적으로 표현하며 인식 성능을 향상시켰다.

핵심용어: 다중대역 음성인식, 부대역 신뢰도, 부대역 거리, 비정상 잡음, 대역제한 잡음
투고분야: 음성처리 분야 (2.5)

Recently, based on the human speech recognition (HSR) model of Fletcher, the multi-band speech recognition has been intensively studied by many researchers. As a new automatic speech recognition (ASR) technique, the multi-band speech recognition splits the frequency domain into several sub-bands and recognizes each sub-band independently. The likelihood scores of sub-bands are weighted according to reliabilities of sub-bands and re-combined to make a final decision. This approach is known to be robust under noisy environments. When the noise is stationary a sub-band SNR can be estimated using the noise information in non-speech interval. However, if the noise is non-stationary it is not feasible to obtain the sub-band SNR. This paper proposes the inverse sub-band distance (ISD) weighting, where a distance of each sub-band is calculated by a stochastic matching of input feature vectors and hidden Markov models. The inverse distance is used as a sub-band weight. Experiments on 1500 ~ 1800 Hz band-limited white noise and classical guitar sound revealed that the proposed method could represent the sub-band reliability effectively and improve the performance under both stationary and non-stationary band-limited noise environments.

Keywords: *Multi-band speech recognition, Sub-band reliability, Sub-band distance, Non-stationary noise, Band-limited noise*
ASK subject classification: *Speech signal processing (2.5)*

책임저자: 조훈영 (hycho@bulsai.kaist.ac.kr)
305-701 대전광역시 유성구 구성동 373-1
한국과학기술원 전자전산학과 전산학전공
(전화: 042-869-8720; 팩스: 042-869-3510)

I. 서 론

조용한 실험실 환경에서 높은 성능을 보이는 음성인식 (ASR; automatic speech recognition) 시스템은 다양한 잡음이 존재하는 실제 응용 환경에서 그 성능이 급격히 저하된다. 잡음 환경에서 인식기의 성능이 떨어지는 이유는 인식기의 학습 환경과 사용 환경간의 불일치 때문이며, 이러한 불일치를 보완하기 위해 잡음에 강한 특징추출, 모델 적응 및 보상, 잡음 제거 등의 분야에서 방대한 연구가 이루어져 왔다[1]. 이러한 연구결과에 의해 백색 잡음, 자동차 실내 소음 및 F16 조종실 소음과 같이 정상적이거나 시간에 따라 천천히 변하는 잡음에 대해서 음성인식 성능을 크게 향상시킬 수 있었다. 그러나 다양한 음성인식 태스크에 대해 인간과 기계의 음성인식 성능을 비교해 본 결과, 현존하는 최고 수준의 음성인식 시스템도 잡음 환경에서는 인간의 음성인식 (HSR; human speech recognition)에 비해 인식 오류율이 한 자리수 이상 크게 나타나 HSR 방식에 대한 보다 깊은 이해가 필요하게 되었다[2].

최근에 HSR에 대한 Fletcher의 연구 결과를 기초로 HSR의 원리를 음성인식 시스템에 적용하려는 연구들이 이루어지고 있다[3]. Fletcher의 연구 결과에 따르면 인간은 주파수 영역에서 각각의 부대역들에 대해 독립적으로 인식을 수행하고, 이들 인식 결과를 최적으로 통합하여 인식 결과를 얻는다. 즉 HSR의 경우 주파수 영역에서 부분적으로 발생한 오류가 다른 주파수 영역에서의 인식에 영향을 미치지 않으므로 정보 손실이 적게 발생한 부대역들에서의 인식 결과를 최대한 반영하여 높은 인식 성능을 획득할 수 있다. 반면에 기존의 ASR에서 특징 벡터는 전체 주파수 대역에 대한 하나의 대표값으로서 스펙트럼 상의 일부분만이 손상된 경우에도 특징 벡터의 요소 전체에 오류가 전파되어 인식률이 크게 저하된다.

다중대역 음성인식은 Fletcher의 연구 결과에 근거하여 음성 스펙트럼을 다수의 부대역으로 구분하고 부대역별로 독립적인 음성 인식을 수행한 후 각각의 인식 결과를 통합하는 새로운 음성인식 방식이다. 이 방식과 관련된 주요 논점으로는 부대역의 개수, 부대역의 범위, 부대역 특징 파라미터의 종류 및 음소, 음절, 단어 등 부대역 인식 결과의 통합을 위한 최적의 시간 단위, 그리고 부대역 인식 결과에 대한 가중 및 통합 방식 등을 들 수 있다. 이 중에서도 부대역 인식결과에 대한 가중 방식은 다중대역 음성인식의 핵심적인 문제로서 기존의 가중치 부여 방식을 크게 학습 자료에 대해 부대역 가중치를 최적화하는 고정 가중 (fixed weighting)과 매 입력 잡음에 대해 부대역 가중치를 최적화하는 적응 가중 (adaptive weighting)으로 구분해 볼 수 있다. 고정 가중 방식으로는 부대역 인식 결과에 동일한 가중 (equal weighting)을 주거나 부대역의 인식 정확도로 가중 (accuracy weighting)하는 방식 또는 MLP (multi-layer perceptron)에 의한 통합 등이 연구되었으며 [4,5], MCE (minimum classification error) 학습에 의해 각 부대역의 가중치가 인식 집단간의 변별력을 최대화하도록 가중치를 부여하는 방법이 제안되기도 하였다[6]. 이외에도 부대역 상호 정보 (mutual information), 또는 최대 우도 (ML; maximum likelihood)에 기반한 부대역 신뢰도 측정 방법이 제안되었다[7]. 적응 가중 방식으로는 부대역 신호대 잡음비에 의한 가중(SNR weighting) 방법, 부대역 상호 정보를 이용한 가중[8] 및 적응 최대 우도 가중[9] 등이 연구되었다. 이중에서 부대역 신호대 잡음비 가중은 상대적으로 적은 연산량으로도 인식성능을 크게 향상시킬 수 있는 방법이며, Fletcher가 제안한 인간의 음성인식 모형에서도 부대역 신호대 잡음비를 이용하여 부대역 인식 결과를 가중 및 통합한다[3]. 이때 잡음이 정상적인 경우에는 무음구간에서 획득한 잡음 정보 혹은 잡음의 히스토그램을 이용하여 부대역 신호대 잡음비를 추정할 수 있으나, 잡음이 비정상적이며 빠르게 변하는 실제 잡음 환경에서는 부대역 신호대 잡음비의 추정이 매우 어렵다.

본 논문에서는 이와 같이 잡음이 비정상적이어서 부대역 신호대 잡음비 추정이 어려운 경우, 학습된 HMM (hidden Markov model)을 이용하여 부대역의 가중치를 추정하는 모델기반의 부대역 신뢰도 추정방식을 제안한다. 제안한 방법은 부대역 HMM들을 PMC (parallel model combination)[10]에서 제안된 모델 변환을 통해 선형 스펙트럼 영역에서의 HMM으로 변환한다. 그리고 변환된 모델과 입력 필터뱅크 에너지 벡터의 거리비교를 통해 각 부대역 거리의 총합을 구하여 거리가 큰 부대역에 작은 가중치를 부여함으로써 정상적 및 비정상적인 잡음에 의해 크게 손상된 부대역의 영향을 줄인다.

본 논문의 2장에서는 Fletcher의 HSR 이론 및 다중대역 음성인식에 관하여 간략히 설명하고, 3장에서는 제안한 방법에 대해 자세히 기술한다. 4장에서 실험 및 결과를 기술하고 마지막으로 5장에서 결론을 맺는다.

II. 다중대역 음성인식

1900년도 초반에 Fletcher의 연구팀은 전화 음성의 명료도 및 선호도를 높이기 위한 정량적 실험을 위해 wif,

moush 등의 무의미한 CVC (consonant-vowel-consonant) 음절과 의미를 갖는 단어 및 문장에 대한 인간의 음성인식 성능을 연구하였다. 이때 무의미한 CVC 음절의 인식률을 명료도 (articulation), 의미를 갖는 단어의 인식률을 이해도 (intelligibility)라고 정의하였다. 이 실험에서 무의미한 CVC 음절을 사용하여 단어에 포함된 문맥 정보 (context information)를 제거함으로써 문맥 정보가 실험결과에 주는 변이를 배제하였으며, 분석 결과 무의미한 CVC 음절의 명료도는 각각의 음소의 명료도의 곱과 같다는 사실을 발견하였다. 따라서 인간은 음절을 인식함에 있어 독립적인 음소 단위로 인식한다고 결론지을 수 있었다. 더 나아가서 Fletcher는 인간이 음소를 어떤 방식으로 인식하는지를 알아보기 위해 고역통과 및 저역통과 필터에 통과시킨 음성의 명료도를 분석한 결과, 주파수 부대역에서의 명료도 오류 (articulation error)는 독립적이며 전대역 (full-band) 명료도 오류와 식 (1)과 같은 관계를 갖는다는 사실을 밝혔다[3].

$$e_F = e_L \cdot e_H \tag{1}$$

식 (1)에서 e_F, e_L 및 e_H는 각각 HSR의 전대역, 저대역 및 고대역 명료도 오류이며, 이 식은 고역통과 및 저역통과 필터를 결정하는 임의의 차단주파수 값에 대해서 성립한다. 이 식을 다시 B개의 부대역에 대해 확장하면 식 (2)와 같이 표현되며, 이를 Fletcher-Allen 법칙 혹은 오류적 (PoE; product-of-errors) 법칙이라고 한다[11].

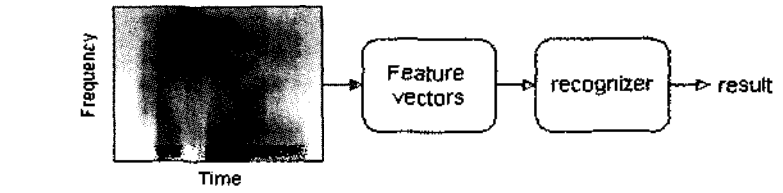
$$e_F = e_1 e_2 \cdots e_B = \prod_{b=1}^{B} e_b \tag{2}$$

이 식은 인간의 경우 언어정보가 주파수 부대역에서 독립적으로 해독되며 이들 결과를 최적으로 통합하여 인식을 수행한다는 사실을 의미한다.

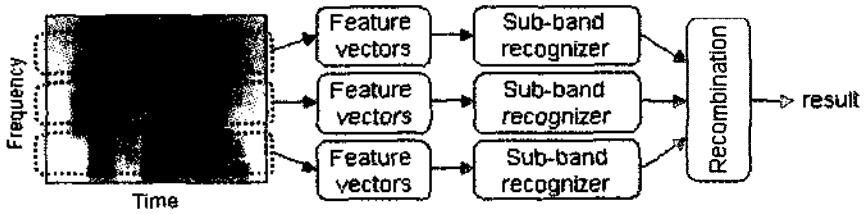
최근에 Fletcher의 연구결과에 근거하여 음성 스펙트럼을 다수의 부대역으로 구분하고, 부대역별로 독립적인 인식을 수행한 후에 인식결과를 통합하는 다중대역 음성인식이 활발히 연구되고 있다. 그림 1은 기존의 음성인식 방식과 다중대역 음성인식 방식의 차이를 나타낸다.

그림 1의 (b)에서 현재까지는 주로 2개에서 7개까지의 주파수 부대역에 대해 연구가 수행되었으며, 주파수 부대역별로 서로 다른 특징을 추출하거나 부대역들을 서로 다른 방법으로 인식할 수 있다. 또 서론에서 기술한 바와 같이 부대역 가중치를 부대역 신호대 잡음비, MCE, 상호 정보, 최대 우도 등 다양한 방식으로 부여할 수 있으며 부대역 인식결과의 통합 (recombination) 방식도 기하 가중 평균, MLP, 전조합 (full combination) 등을 적용할 수 있다. 이중에서 잡음 하에서 인식을 위해 자주 사용된 기하 가중 평균은 다음과 같이 정의된다. 전대역을 b개의 부대역으로 나누고 각 부대역에서 추출한 특징벡터를 $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_b$라 하면, 부대역 특징들이 서로 독립일 때 전대역 특징 $\mathbf{x}$는 $\mathbf{x}^t=(\mathbf{x}_1^t, \mathbf{x}_2^t, \cdots, \mathbf{x}_b^t)$이며 ($t$는 벡터의 전치),

$$\Pr(\mathbf{x}\mid\lambda) = \prod_{b=1}^{B} \Pr(\mathbf{x}_b \mid \lambda) = \prod_{b=1}^{B} \Pr(\mathbf{x}_b \mid \lambda_b) \tag{3}$$



(a) 기존의 음성인식 시스템
(a) conventional ASR system



(b) 다중대역 음성인식 시스템
(b) multi-band ASR system

그림 1. 기존의 음성인식 시스템과 다중대역 음성인식 시스템의 개념적 차이
Fig. 1. Conceptual difference between conventional ASR system and multi-band ASR system.

가 성립한다. 단, λ는 전대역 특징에 대한 모델이며 λ_b는 부대역 특징을 사용한 모델이다. 이 식의 양변에 로그 함수를 적용하면 다음과 같다.

$$\log \Pr(\mathbf{x} \mid \lambda) = \sum_{b=1}^{B} \log \Pr(\mathbf{x}_b \mid \lambda_b) \tag{4}$$

위 식은 잡음이 첨가되지 않은 경우이다. 잡음이 첨가되었을 때는 각 부대역의 신뢰도 w_b를 반영한 식 (5)를 사용한다.

$$\log \Pr(\mathbf{x} \mid \lambda) = \sum_{b=1}^{B} w_b \cdot \log \Pr(\mathbf{x}_b \mid \lambda_b) \tag{5}$$

III. 부대역 신뢰도의 추정 및 가중

다중대역 음성인식에서 부대역의 신뢰도를 의미하는 부대역 가중치는 인식 성능에 가장 큰 영향을 미치는 요인들 가운데 하나로서 본 논문에서는 매 입력마다 변하는 잡음에 대해 적응적인 가중치를 계산하는 방법에 중점을 두었다. Fletcher의 연구에 의하면 각 임계 대역 (critical band)에서 30 dB로 정규화된 신호대 잡음비가 해당 부대역의 인식 오류율을 결정한다[3]. 이 원리를 구현한 부대역 가중방식으로서 각 부대역의 신호대 잡음비를 사전 (a priori)에 계산하고, 신호대 잡음비가 10 dB 이하인 부대역을 우도 계산에서 제외하거나[4], 부대역의 에너지 히스토그램에서 구한 신호대 잡음비를 정규화하여 가중하였다[5]. 잡음이 정상적인 경우에는 무음구간의 잡음정보 혹은 히스토그램을 통해 추정한 잡음의 스펙트럼을 이용하여 부대역 신호대 잡음비를 구할 수 있었다. 그러나 잡음이 비정상적인 경우에는 잡음 스펙트럼이 시간에 따라 변하므로 기존의 방식으로는 부대역 신호대 잡음비를 정확히 추정하기가 어렵다. 본 논문에서는 연속 HMM과 입력 잡음음성의 통계적 정합을 통해 부대역의 거리를 추정하는 ISD (inverse sub-band distance)를 제안한다. 이 방법은 먼저 입력 음성에서 추출한 부대역 MFCC (mel frequency cepstral coefficient)에 대해 모델 공간에서 최적 상태열을 추출한다. 다음으로 최적 상태열의 모델 파라미터들을 선형 스펙트럼 영역으로 변환하고 입력 MFCC도 역변환에 의해 선형 스펙트럼 영역으로 변환한 후 둘 사이의 거리를 계산한다. 마지막으로 부대역별로 구한 거리의 역을 부대역의 가중치로 사용한다. 다음의 절들에서 각각의 단계에 대해 보다 자세히 설명하기로 한다.

3.1. 입력 음성과 모델의 통계적 정합

켑스트럼 파라미터는 로그 필터뱅크 에너지를 코사인 변환하여 얻는 벡터값으로서 필터뱅크 에너지에 비해 벡터 차수를 줄일 수 있고, 벡터 요소들 사이의 독립성이 향상된다. 또한 켑스트럼 파라미터는 HMM의 공분산 행렬로 대각행렬을 사용할 수 있어 모델 파라미터 수를 크게 줄일 수 있으며 동시에 모델링 정확도를 높이고 연산량을 줄일 수 있다. 따라서 기존의 대다수 음성인식기는 MFCC와 같은 켑스트럼 벡터를 사용하였으며, 본 논문에서도 부대역의 특징 벡터로 MFCC를 사용한다. 입력 잡음음성의 부대역 MFCC 특징을 $\mathbf{Y}^c=\{\mathbf{y}_1^c\ \mathbf{y}_2^c\ \cdots\ \mathbf{y}_T^c\}$, 잡음이 가산되지 않은 원음의 MFCC 특징을 $\mathbf{X}^c=\{\mathbf{x}_1^c\ \mathbf{x}_2^c\ \cdots\ \mathbf{x}_T^c\}$라 하고, 각각에 해당하는 필터뱅크 에너지 특징을 $\mathbf{Y}^e$ 및 $\mathbf{X}^e$, 로그 필터뱅크 에너자를 $\mathbf{Y}^l$ 및 $\mathbf{X}^l$이라 하면, 이들 사이에 식 (6), (7)과 같은 관계가 성립한다. 식 (6)의 $\mathbf{b}_t^c$는 잡음에 해당하는 벡터이고, 식 (7)의 $\mathbf{C}$는 DCT (discrete cosine transform) 행렬이다.

$$\mathbf{x}_t^e = \mathbf{y}_t^e - \mathbf{b}_t^e \tag{6}$$

$$\mathbf{y}_t^c = \mathbf{C} \cdot \mathbf{y}_t^l \tag{7}$$

제안한 방법은 식 (6)에서 벡터 $\mathbf{b}_t^e$의 크기, 즉, $\mathbf{y}_t^e$와 $\mathbf{x}_t^e$ 사이의 평균 거리를 부대역별로 계산하여 부대역 가중치로 사용한다. 이 때, $\mathbf{x}_t^e$는 직접 구할 수 없으므로 HMM 상태의 평균벡터에서 획득한다. 따라서 제안한 방법에서 HMM 모델과 입력과의 거리 D는 식 (8)에서 나타낸 바와 같이 Mahalanobis 거리 $|r_k|$의 총합을 프레임 개수 T로 나눈 평균 Mahalanobis 거리를 사용한다. 식 (8)에서 $x_t^e(k)$는 식 (6)의 $\mathbf{x}_t^e$에서 k번째 벡터요소이며, K는 현재 부대역의 멜 필터 개수이다.

$$D = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left|\frac{y_t^e(k)-x_t^e(k)}{\sigma_t^e(k)}\right| \approx \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left|\frac{y_t^e(k)-\mu_t^e(k)}{\sigma_t^e(k)}\right| = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}|r_k| \tag{8}$$

켑스트럼으로 학습된 모델 파라미터 공간을 Λ^c라고 하자. 제안한 방법은 우선 식 (9)과 같이 Λ^c에 대한 입력 벡터열 $\mathbf{Y}^c=\{\mathbf{y}_1^c\ \mathbf{y}_2^c\ \cdots\ \mathbf{y}_T^c\}$의 ML 상태열 $S^c=\{s_1^c\ s_2^c\ \cdots\ s_T^c\}$을 비터비 (Viterbi) 알고리즘에 의해 구한다.

$$S^c = \arg\max_{S} \Pr(\mathbf{Y}^c, S \mid \Lambda^c) \tag{9}$$

이 때, s_i^c에 포함된 평균벡터 μ_i^c는 켑스트럼 벡터이므로 식 (8)에 직접 적용할 수 없다. 따라서 다음 절에서 기술하는 바와 같이 모델 파라미터 공간의 변환이 필요하다.

3.2. HMM 변환 및 부대역 신뢰도 추정

켑스트럼 모델 파라미터 공간에서 임의의 HMM 상태 s^c에 포함된 평균벡터와 공분산 행렬을 각각 μ^c 및 $\sum^c$라 하면 DCT 역변환 $\mathbf{C}^{-1}$에 의해 이 모델 파라미터들을 켑스트럼 영역에서 로그 스펙트럼 영역으로 식 (10)과 (11)처럼 변환할 수 있다[10].

$$\mu^l = \mathbf{C}^{-1}\mu^c \tag{10}$$

$$\sum^l = \mathbf{C}^{-1}\sum^c(\mathbf{C}^{-1})^t \tag{11}$$

이와 같이 변환된 평균 벡터의 k번째 요소를 $\mu^l(k)$, 공분산 행렬의 (k,k)번째 요소를 $\sum^l(k)$라고 하면 로그 스펙트럼 영역에서 선형 스펙트럼 영역으로의 변환은 다음 식과 같이 근사할 수 있다.

$$\mu^e(k) = \exp(\mu^l(k) + \sum^l(k)/2) \tag{12}$$

$$\sum^e(k) = (\mu^l(k))^2[\exp(\sum^l(k)) - 1] \tag{13}$$

식 (9)에서 구한 상태열에 포함된 모델 파라미터들을 식 (10)에서 식 (13)까지의 절차를 통해 변환한 후의 상태열을 $S^e = \{s_1^e s_2^e \cdots s_N^e\}$라고 하면, 상태 s_i^e에 포함된 평균 벡터 μ_i^e와 공분산 벡터 $\sum_i^e$를 식 (8)의 $\mu_i^e(k)$와 $\sigma_i^e(k)$에 각각 적용하여 부대역 거리 D를 얻을 수 있다. b번째 부대역에서 구한 부대역 거리를 D_b라고 할 때 부대역 가중치 $w_b = 1/D_b$을 이용하여 식 (14)과 같이 최종 우도를 구한다.

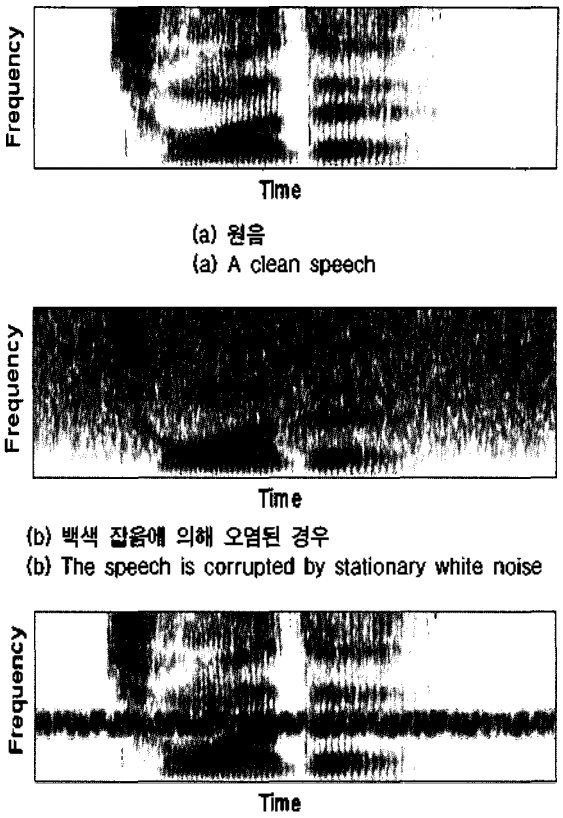
$$\log \Pr(\mathbf{y} \mid \lambda) = \sum_{b=1}^{B} w_b \cdot \log \Pr(\mathbf{Y}_b \mid \lambda_b) \tag{14}$$

위 식에서 부대역 거리가 클수록 해당 부대역의 신뢰도는 낮으며, 반대로 거리가 작을수록 부대역은 높은 신뢰도를 가지고 인식 결과에 크게 반영된다.

IV. 실험 및 결과

4.1. 실험 환경

제안한 ISD의 성능을 검증하기 위해 100 단어 규모의 고립단어 인식실험을 수행하였다. 실험에 사용한 음성 데이터베이스는 국어공학센터의 PBW (phoneme balanced word)-452이며[12], 이 중에서 100개의 단어를 임의로 선정하였다. 학습 자료로는 각 단어별로 남녀 화자 50명에 대한 2회의 발성을 사용하였으며, 평가 자료로는 남녀 화자 15명이 발성한 3000개의 발성을 사용하였다. 평가 자료에는 정상 대역제한 잡음 (stationary band-limited noise), 비정상 대역제한 잡음 (non-stationary band- limited noise) 및 백색 잡음을 신호대 잡음비 15, 10, 5, 0 dB로 가산하였다. 정상 대역제한 잡음은 백색 잡음을 1500~1800 Hz의 대역통과필터에 통과시켰고, 비정상 대역제한 잡음은 클래식 기타 연주곡을 1500~1800 Hz의 대역통과필터에 통과시켜 얻었다. 그림 2는 깨끗한 음성, 신호대 잡음비 0 dB의 백색 잡음이 가산된 경우, 신호대 잡음비 0 dB의 대역제한 잡음이 가산된 경우의 음성 스펙트럼을 나타낸다.



(a) 원음
(a) A clean speech

(b) 백색 잡음에 의해 오염된 경우
(b) The speech is corrupted by stationary white noise

(c) 백색 대역제한 잡음에 의해 일부 주파수 부대역이 손상된 경우
(c) A sub-band of the speech is corrupted by white band-limited noise

그림 2. 다양한 가산 잡음에 의해 오염된 음성(/청외대/)의 스펙트로그램 비교
Fig. 2. Comparison of speech spectrograms corrupted by various additive noises.

실험에 사용한 다중대역 ASR 시스템의 부대역 주파수 범위는 0~1155 Hz, 1050~2996 Hz, 2723~8000 Hz이며, 부대역별로 8차의 멜 필터뱅크 에너지를 구하고 이로부터 켑스트럼 0차를 포함한 4차의 MFCC를 추출하였다.

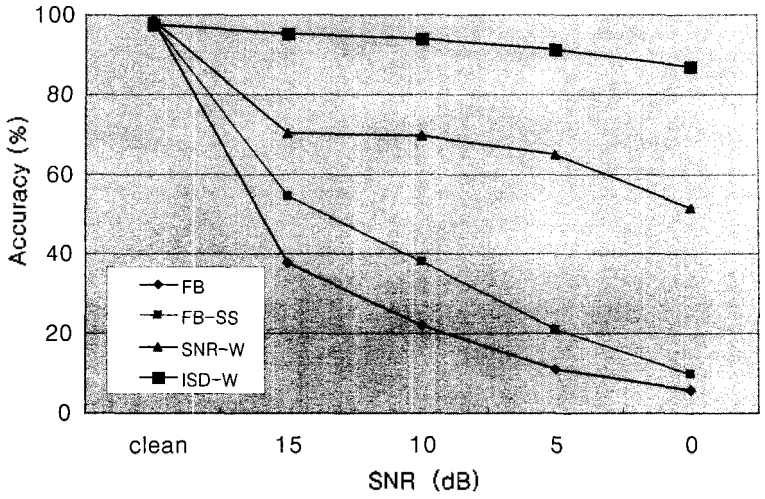


그림 3. 1500~1800 Hz 대역제한 클래식 기타음에 대한 단어 인식률(%) 비교
Fig. 3. Comparison of word accuracies (%) for 1500~1800 Hz band-limited classical guitar sound.

비교를 위한 기존의 전대역 (full-band) 음성인식기는 24차의 멜 필터뱅크로부터 캡스트럼 0차를 포함한 12차의 MFCC를 추출하였다.

4.2. 실험 결과

표 1은 1500~1800 Hz 백색 대역제한 잡음에 대한 기존의 전대역 음성인식 및 다중대역 음성인식의 성능을 나타낸다.

표에서 SB1, SB2, SB3는 부대역 각각의 인식결과이며, 잡음에 의해 심하게 손상된 두번째 대역의 인식결과가 매우 낮았고, 세번째 대역은 부대역 자체에 포함된 음성정보가 적어 첫번째 부대역보다 낮은 인식률을 보였다. FB는 기존의 음성인식 방법으로서 두번째 대역에서 발생한 오류가 전체 대역에서 추출한 특징벡터에 전반적으로 영향을 미쳐 인식률이 저하되었다. 다중대역 음성인식에서 기존의 신호대 잡음비 가중 (SNR-W)과 제안한 ISD 가중 (ISD-W)에 의해 부대역 정보를 통합한 결과, 기존의 FB 방식보다 월등한 성능을 나타내었다. 또한, 이 방식들은 기존의 잡음처리 방법으로서 잘 알려진 스펙트럼 차감법 (spectral subtraction)[13]을 적용한 결과 (FB-SS)보다도 인식률이 높았다. 정상 대역제한 잡음의 경우, 제안한 방법이 잡음이 그다지 크지 않을 때는 SNR-W보다 높은 성능을 보였으나 잡음의 크기가 커짐에 따라 SNR-W가 더 높은 성능을 보였다. 마지막으로 PoE는 식 (2)에 정의된 인간의 부대역 통합방식이다. 표에서 SB1, SB2, SB3가 인간에 의해 수행되었다고 가정하고 이 식에 대입하여 인식률을 계산하였으며 가장 좋은 성능을 나타내었다.

그림 3은 비정상 대역제한 잡음에 대한 기존의 FB 및 FB-SS, 그리고 SNR-W와 ISD-W의 성능을 비교한 것이다. 그림에서 다중대역 인식방식이 전대역 인식방식에 비해 높은 인식률을 나타내었다. 또, 잡음이 비정상적인 특성을 가지는 경우 제안한 ISD-W가 다른 방식들에 비해 월등한 성능을 나타내어 제안한 방법이 비정상 대역제한 잡음에 효과적임을 알 수 있었다.

마지막으로 표 2는 전체 주파수 대역에 걸쳐 음성을

표 1. 1500~1800 Hz 대역제한 백색잡음에 대한 단어 인식률 (%) 비교
Table 1. Comparison of word accuracies (%) under 1500~1800 Hz band-limited white noise environments.

	SNR				
	clean	15 dB	10 dB	5 dB	0 dB
SB1	92.6	92.0	91.3	89.5	86.2
SB2	87.8	3.1	2.2	1.7	1.3
SB3	61.7	61.4	60.8	58.4	55.1
FB	98.8	34.8	21.2	11.9	6.7
SNR-W	98.5	85.0	92.1	93.7	90.5
ISD-W	97.6	95.0	93.2	87.9	80.2
FB-SS	98.4	66.3	50.7	32.5	17.1
PoE	99.7	97.0	96.7	95.7	93.9

표 2. 전대역 백색 잡음에 대한 단어 인식률(%) 비교
Table 2. Comparison of word accuracies (%) under full-band white noise environments.

	SNR				
	clean	15 dB	10 dB	5 dB	0 dB
FB	98.8	22.2	6.0	2.8	1.2
SNR-W	98.5	38.4	22.6	11.3	5.1
ISD-W	97.6	46.3	23.2	8.0	2.2

오염시키는 백색 잡음에 대한 SNR-W 및 ISD-W의 성능을 비교한 것으로서 이 두 방식 사이에 백색 대역제한 잡음의 경우와 유사한 성향을 보였다. 또한, 주파수 전대역이 오염된 경우는 주파수 대역 일부가 손상된 경우에 비해 다중대역 음성인식 방식과 기존의 인식방식 사이에 성능차이가 크지 않았다.

V. 결 론

본 논문에서는 비정상 및 정상 잡음에 대해 다중대역 음성인식의 성능을 향상시키기 위한 부대역 신뢰도 추정 방법을 제안하였다. 제안한 ISD 방법은 잡음음성의 특징벡터와 HMM의 통계적 정합에 의해 둘 사이의 평균 Mahalanobis 거리를 각 부대역 별로 구하고, 이 거리의 역을 부대역 가중치로 사용하였다. 이 방법은 잡음에 심하게 오염되어 신뢰도가 낮아질수록 부대역의 거리가 커져 해당 부대역이 최종 인식결과에 적게 영향을 미치게 된다. 실험결과 제안한 방법은 비정상 대역제한 잡음에 대해 기존의 부대역 신호대 잡음비 가중보다 월등한 성능을 보였다. 또한, 대역제한 잡음에 대해 다중대역 인식은 기존의 전대역 음성인식에 스펙트럼 차감법을 적용한 경우보다도 더욱 효과적이었다. 향후에는 다중대역 음성인식에서 부대역의 정의, 고정 및 적응 가중의 혼합, 기존의 잡음제거 방법의 적용 등에 관한 연구가 필요하다.

참 고 문 헌

1. Y. Gong, "Speech recognition in noise environments: A survey," *Speech Communication*, 16, 261–291, 1995.
2. R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, 22, 1–15, 1997.
3. J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. On Speech and Audio Processing*, 2 (4), 567–577, 1994.
4. H. Hermansky, S. Tibrewala and M. Pavel, "Towards ASR on partially corrupted speech," *Proc. Int. Conf. on Spoken Language Processing*, 1, 462–465, 1996.
5. H. Bourlard and S. Dupont, "ASR based on independent processing and recombination of partial frequency bands," *Proc. Int. Conf. on Spoken Language Processing*, 1, 422–425, 1996.
6. C. Christophe, H. J. Paul and F. Dominique, "Towards a global optimization scheme for multi-band speech recognition," *Proc. EUROSPEECH*, 2, 587–590, 1999.
7. Y. C. Tam and B. Mak, "Optimization of sub-band weights using simulated noisy speech in multi-band speech recognition," *Proc. Int. Conf. on Spoken Language Processing*, 2000.
8. S. Okawa, T. Nakajima and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," *Proc. EUROSPEECH*, 2, 603–606, 1999.
9. A. Hagen, H. Bourlard and A. Morris, "Adaptive ML-weighting in multi-band recombination of Gaussian mixture ASR," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1, 257–260, 2001.
10. M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. On Speech and Audio Processing*, 4 (5), 352–359, 1996.
11. A. Morris, A. Hagen, H. Glotin and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, 34, 25–40, 2001.
12. 이용주, "음성데이터베이스의 현황 및 과제," 제 13회 음성통신 및 신호처리 워크샵, 13 (1), 279–287, 1996.
13. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. On Speech and Audio Processing*, 27 (2), 113–120, 1979.

저자 약력

● 조 훈 영 (Hoon-Young Cho)

1995년 8월: 한국과학기술원 전자전산학과 (학사)
1998년 2월: 한국과학기술원 전자전산학과 (석사)
1998년 3월~현재: 한국과학기술원 전자전산학과 전산학 전공 박사과정
※ 주관심분야: 잡음에 강한 음성인식, 패턴인식

● 지 상 문 (Sang-Mun Chi)

1991년: 서울대학교 수학교육과 (학사)
1993년: 한국과학기술원 수학과 (석사)
1998년: 한국과학기술원 전자전산학과 (박사)
1993년~2000년: 삼성전자 정보통신 선임연구원
2000년~2001년: L&H 연구개발본부 책임연구원
2001년~현재: 경성대학교 정보과학부 전임강사
※ 주관심분야: 패턴인식

● 오 영 환 (Yung-Hwan Oh)

1972년: 서울대학교 공과대학 (학사)
1974년: 서울대학교 교육대학원 (석사)
1980년: Tokyo Institute of Technology 정보공학 전공 (박사)
1981년~1985년: 충북대학교 컴퓨터 공학과 조교수
1983년~1984년: University of California (Davis) 연구교수
1995년~1996년: Carnegie-Mellon University 연구교수
1985년~현재: 한국과학기술원 전자전산학과 전산학 전공 교수
※ 주관심분야: 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가시스템