부분 손상된 음성의 인식성능 향상을 위한 가중 필터뱅크 분석 및 모델 적응

조훈영(KAIST), 오영환(KAIST)

<차 례>

<Abstract>

Weighted filter bank analysis and model adaptation for improving the recognition performance of partially corrupted speech

Hoon-Young Cho, Yung-Hwan Oh

We propose a weighted filter bank analysis and model adaptation (WFBA-MA) scheme to improve the utilization of uncorrupted or less severely corrupted frequency regions for robust speech recognition. A weighted mel frequency cepstral coefficient is obtained by weighting log filter bank energies with reliability coefficients and hidden Markov models are also modified to reflect the local reliabilities. Experimental results on TIDIGITS database corrupted by band-limited noises and car noise indicated that the proposed WFBA-MA scheme utilizes the uncorrupted speech information well, significantly improving recognition performance in comparison to multi-band speech recognition systems.

* 주제어: 다중대역, 음성인식, 손실데이터 기법

1. 서 론

실제 응용환경에서 동작하는 음성인식 시스템은 다양한 특성의 잡음에 의해 성능이 저하되어 실용화에 어려움을 겪고 있다. 잡음에 강한 음성인식 분야에서 현재까지 이루어진 방대한 연구결과로 정상적 잡음(stationary noise) 등의 제한적인 잡음 환경에서는 상당 수준의 인식성능 개선이 가능하게 되었다. 그러나, 실세계에는 버스트(burst) 잡음, 기관총, 문닫힘 소리(door slam)와 같이 시간축 상에서 음성의 일부 구간을 손상시키는 시간선택(time-selective) 잡음과 전화기의 신호음, 벨소리, 자동차 경적소리처럼 음성의 주파수 영역 일부를 손상시키는 주파수선택(frequency-selective) 혹은 대역제한(band-limited) 잡음도 다수 존재하며, 이러한 종류의 잡음에 대해서 음성인식의 성능을 향상시키는 방법에 관한 연구가 필요하다[1]-[12].

본 연구는 잡음에 의해 음성의 주파수 영역 일부가 손상된 경우에 중점을 두었다. 이 경우, 손실 데이터 기법(missing data technique) 또는 다중대역 음성인식(multi-band speech recognition)을 고려해 볼 수 있다. 손실 데이터 기법은 음성의 스펙트로그램에서 손상된 영역을 찾고, 이 영역을 출력확률 계산 단계에서 제외하는 방식이다[1]. 이 방법은 잡음에 대한 가정이 불필요하다는 장점이 있으나, 직교화된 특징벡터를 사용하기 어려우므로 잡음이 없는 음성에 대해 성능이 떨어지는 단점이 있다[2]. 다중대역 음성인식은 Fletcher의 다중독립 채널 모형(multi-independent channel model)에 기반하여 음성의 전체 주파수 대역을 다수의 부대역으로 나누고, 각 부대역에 대해 독립적으로 음성인식을 수행한 다음 부대역 인식결과를 통합하여 최종적인 인식 결과를 얻는 방식으로서 주파수 영역의 일부가 상대적으로 심하게 손상된 경우에 매우 효과적이라고 알려졌다[4,6]. 그러나, 부대역의 경계선이 학습 단계에서 결정되고 고정적이어서 임의의 주파수 영역에서 발생한 대역제한 잡음에 효과적으로 대처할 수 없으며, 각각의 부대역에서 독립적인 특징을 추출하므로 부대역 간의 상관 정보(correlation information)가 손실된다. 또한, 기존의 전대역(full-band) 음성인식 시스템에 이 방법을 적용할 경우 시스템 구조의 차이로 인해 전체 시스템을 재구축해야 하므로 비용부담이 크다. 따라서, 본 연구에서는 기존의 전대역 음성인식에 다중대역 인식과 동일한 기능을 추가하기 위한 가중 필터뱅크 분석 및 모델 적응(weighted filter-bank analysis and model adaptation; WFBA-MA) 방법을 제안하였다.

본 논문의 2장에서는 Fletcher의 다중독립 채널 모형과 다중대역 음성인식에 관하여 기술하고, 본 논문에서 검토한 문제점을 언급한다. 제 3장에서는 제안한 가중 필터뱅크 분석 및 모델 적응 방식을 설명하며, 제 4장과 5장에서 실험 및 결과를 기술하고 결론을 맺도록 한다.

2. 다중대역 음성인식

본 장에서는 음성신호에서 부분적으로 덜 손상된 정보를 가중하여 인식하는 인간의 음성인식과 이를 응용한 다중대역 음성인식에 대해서 설명하고, 본 연구의 동기가 된 다중대역 인식방법의 개선점을 기술한다.

2.1. 다중독립 채널 모형

1918년부터 1950년 사이에 Fletcher의 연구그룹은 wif, moush 등의 무의미한 CVC (consonant-vowel-consonant) 음절과 의미를 갖는 단어 및 문장에 대한 방대한 청취실험과 통계적 분석을 통해 인간의 음성인식(human speech recognition; HSR) 원리를 연구한 결과 다음과 같은 사실들을 발견하였다[3].

(a) HSR에서는 주파수 국부적인 특징이 독립적인 채널들에서 추출된다.
(b) 특정 주파수 영역에서의 부분 인식 오류(partial recognition error)는 다른 주파수 영역에서의 부분 인식 오류에 영향을 미치지 않는다.
(c) 독립적인 주파수 채널들에서 추출된 특징에 대한 인식결과는 음소와 같은 기본적 소리단위로 통합되고, 다시 음절 및 단어 등의 큰 단위로 인식된다.

이 연구에서 무의미한 CVC 음절의 인식률을 명료도(articulation), 의미를 갖는 단어의 인식률을 이해도(intelligibility)라고 정의하였다. 실험에서 무의미한 CVC 음절을 사용하여 단어에 포함된 문맥정보를 제거함으로써 문맥정보가 실험결과에 주는 변이를 배제하였으며, 통계적으로 분석한 결과 무의미한 CVC 음절의 명료도는 각 음소들의 명료도의 곱과 같다는 사실을 발견하였다. 따라서, 인간은 음절을 인식함에 있어 독립적인 음소 단위로 인식한다고 결론지을 수 있었다. 더 나아가서 Fletcher는 인간의 음소인식 방식을 알아보기 위해 공통의 차단주파수를 갖는 고역통과 및 저역통과 필터에 통과시킨 음성의 명료도를 분석하였다. 그 결과 고주파수 및 저주파수 부대역의 명료도 오류(articulation error)는 서로 독립적이며, 전대역(full-band) 명료도 오류와 다음과 같은 관계를 갖는다는 사실을 밝혔다.

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

식 1에서 e_F, e_L 및 e_H는 각각 전대역, 저주파수 부대역 및 고주파수 부대역의 명료도 오류이며, 두 개의 대역은 동일한 차단 주파수 f_c에 의해 구분되고 임의의 f_c에 대해 성립한다. 이를 B개의 부대역에 대해 확장하면 식 2와 같이 표현

되며, 이를 Fletcher의 다중독립 채널(multi-independent channel; MIC) 모형 또는 오류적(product-of-error; PoE) 법칙이라고 한다[3,6].

(2) $$e_F = e_1 e_2 \cdots e_B$$

이 식에 의하면 인간은 인식 오류가 0인 주파수 부대역이 존재하기만 하면 다른 주파수 대역이 손상되었다 해도 전체 대역의 인식 오류가 0이 되어 정확한 인식이 가능하다. 따라서 인간은 주파수선택 잡음에 의해 음성이 손상된 경우에도 매우 높은 인식률을 나타낸다. 반면에 현재 널리 사용되는 HMM 음성인식 시스템은 주파수 영역의 일부가 상대적으로 심하게 손상된 경우에 대한 처리방법이 고려되지 않아 이 경우 성능저하가 크다. 따라서 비교적 손상정도가 적은 주파수 영역을 강조하여 인식하는 능력을 기존의 음성인식기에 부여할 필요가 있다.

2.2. 다중대역 음성인식

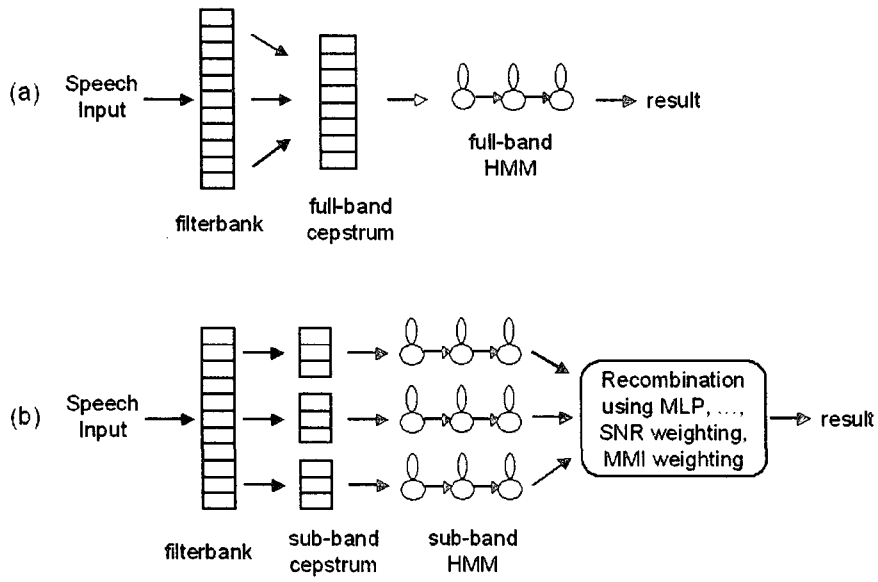
1996년부터 Fletcher의 MIC 모형에 기반하여 음성의 전체 주파수 대역을 다수의 부대역으로 나누고, 부대역별로 독립적인 인식을 수행한 후에 인식결과를 통합하는 다중대역 음성인식 관련 연구결과가 활발히 발표되었다[4,5]. 그림 1은 현재 일반적으로 널리 사용되고 있는 전대역 특징기반의 음성인식 시스템과 다중대역 음성인식의 차이를 간략히 나타낸다. 다중대역 음성인식에서 $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\cdots$, $\boldsymbol{X}_B$를 B개의 부대역에서 추출한 특징벡터라고 하고, M_j^b를 클래스 j의 b번째 부대역 HMM이라고 하면, 이 클래스에 대한 로그 우도는 다음과 같이 가중 통합될 수 있다.

(3) $$\log \Pr(X|M_j) = \sum_{b=1}^{B} w_b \cdot \log \Pr(X_b|M_j^b)$$

위 식에서 w_b는 부대역의 신뢰도 혹은 정보량을 나타내는 항으로서 이 값이 0이면 해당 부대역이 인식결과에 아무런 영향을 미치지 않게 된다.

다중대역 인식에서는 주파수 부대역의 개수 및 범위, 부대역 특징벡터의 종류 등과 같은 부대역 정의 방법, 부대역 인식결과의 신뢰도 추정 및 통합방법, 부대역 인식결과의 통합 시점 등에 관한 연구가 진행되어 왔다.

주파수 부대역의 개수로는 주로 2개에서 7개 사이를 주로 사용해 왔으며, 부대역의 범위는 멜 단위로 변환된 신호의 전체 주파수 영역을 부대역 개수로 균등하게 분할하는 경우가 많았다. 부대역 인식결과의 신뢰도 혹은 가중치로는 부대역의



<그림 1> 전대역(full-band) 음성인식과 다중대역(multi-band) 음성인식의 개념적 차이 (a) 전대역 음성인식 (b) 다중대역 음성인식

SNR, 부대역 상호 정보(mutual information) 및 적응 최대우도를 정규화하여 가중하는 방식들이 제안되었다[8,9,12]. 부대역 인식 결과의 통합방법으로는 식 1과 같은 선형 가중 통합, MLP 등의 비선형 통합 외에도 부대역들의 모든 가능한 조합에 대해 최적의 조합을 선택하는 전조합(full combination) 및 부대역의 신뢰도를 별도로 추정할 필요가 없는 PUM (probabilistic union model) 방식 등이 연구되었다[2,7]. 현재 많은 인식 시스템들이 채택하고 있는 전대역(full-band) 음성인식 방식은 전체 주파수 대역의 전반적인 스펙트럼 형태를 표현하는 p차의 직교화된 특징벡터를 사용하므로 스펙트럼의 일부 영역에 오류가 발생할 경우에도 전체 p차의 벡터요소에 오류가 전파된다. 반면에 다중대역 시스템은 다수의 부대역에 대해 독립적인 특징을 추출하여 독립적으로 인식하므로 손상되지 않은 주파수 영역의 음성정보를 최대한 활용할 수 있다[4].

비록 다중대역 인식방식이 부분 손상된 음성에 매우 효과적인 것으로 알려져 왔으나, 현재의 방식에는 몇 가지 한계점이 있다. 첫째, 주파수 부대역의 경계선이 학습 단계에 결정되고 고정적이므로 임의로 발생하는 주파수선택 잡음의 특성을 반영함에 있어 한계가 있다. 예를 들어 동일한 대역폭을 갖는 주파수선택 잡음이라고 하더라도 부대역 경계선에 걸쳐 발생하는 경우는 두 개의 부대역을 동시에 손상시켜 잡음이 경계선에 걸치지 않은 경우에 비해 인식률이 떨어진다. 또한, 잡음의 대역폭이 좁을 경우, 해당 부대역 내부에 여전히 음성인식에 유효한 정보가 존재할 수 있음에도 불구하고 이를 효과적으로 활용할 수 없다. 둘째, 광대역 잡

음의 경우 부대역 음성정보간의 상관(correlation) 정보가 인식에 중요하게 사용된다. 그러나, 각 부대역에서 독립적인 특징을 추출하는 다중대역 인식의 경우 부대역 간 상관정보를 잘 활용하지 못하므로 광대역 잡음에 대해서는 전대역 방식에 비해 그다지 효과적이지 않다고 알려졌다[8]. 마지막으로 전대역 인식기와의 큰 구조적 차이로 인해 전체 음성인식 시스템을 재구축해야 하므로 비용의 부담이 크다. 따라서, 기존의 전대역 인식방식의 범주 내에서 동일한 부분정보 가중효과를 얻을 수 있는 방법에 대한 연구가 필요하다.

3. 가중 필터뱅크 분석 및 모델 적응

본 장에서는 전대역 음성인식 방식을 유지하면서도 앞 절에서 언급한 다중대역 음성인식 방식의 문제점을 해소할 수 있는 가중 필터뱅크 분석 및 모델 적응(weighted filter bank analysis and model adaptation; WFBA-MA)을 가중 필터뱅크 분석부와 모델 적응부로 구분하여 소개한다.

3.1. 가중 필터뱅크 분석

멜 필터뱅크 분석에서 음성 프레임의 파워스펙트럼을 $|X(k)|^2$라고 할 때, i번째 채널의 필터뱅크 에너지 x_i 및 로그 필터뱅크 에너지 x_i^l 은 다음과 같이 계산된다.

$$x_i = \sum_{k=1}^{K} |X(k)|^2 \cdot \psi_i(k) \tag{4}$$

$$x_i^l = \log(x_i), \quad 1 \le i \le Q \tag{5}$$

위 식에서 k는 FFT 인덱스, $\psi_i(k)$는 i번째 멜 대역통과 필터를 나타낸다. 위첨자 l은 이 변수가 로그 스펙트럼 영역의 값임을 의미한다. 위 식에서 구한 Q차의 로그 필터뱅크 에너지 벡터를 $\boldsymbol{x}^l = (x_1^l, x_2^l, \cdots, x_Q^l)^t$라고 할 때, 이 벡터의 요소들은 채널의 신뢰도에 따라 음성인식에 유효한 정보를 다른 정도로 포함하게 된다. 가중 필터뱅크 분석은 잡음에 의해 크게 손상된 채널의 벡터 요소가 표현하는 값은 적은 정보량을 가지므로 인식에 중요하게 사용하지 않도록 하고, 덜 손상된 채널에서 추출한 벡터 요소는 신뢰도가 높은 정보를 포함하므로 인식에 크게

기여하도록 한다. 필터뱅크 채널의 신뢰도를 0에서 1사이의 값으로 표현한 값을 $(w_1, w_2, \cdots, w_Q)$라 할 때, 이 값들로부터 대각 가중 행렬 $W = diag(w_1, w_2, \cdots, w_Q)$를 정의할 수 있다. 이 행렬을 이용하여 가중 로그 필터뱅크 에너지 벡터 $\widehat{x^l}$을 다음과 같이 구한다.

$$\widehat{x^l} = W \cdot x^l = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_Q \end{bmatrix} \begin{bmatrix} x_1^l \\ x_2^l \\ \vdots \\ x_Q^l \end{bmatrix} \tag{6}$$

본 연구에서는 각 채널의 신호대 잡음비(signal-to-noise ratio; SNR)를 계산하고, 다음과 같은 시그모이드 함수를 이용하여 SNR 0dB에서 30dB 사이를 0에서 1사이의 값이 되도록 정규화하였다.

$$w_i = \frac{1}{1 + \exp(-\alpha(\rho_i - 15))} \tag{7}$$

위 식에서 α는 함수의 0에서 1사이 값의 기울기를 조절하는 변수로서 본 연구에서는 0.3을 실험적으로 결정하여 사용하였고, ρ_i 는 채널의 SNR이다.

켑스트럼 추출의 마지막 단계로 이산 코사인 변환(discrete cosine transform; DCT) 행렬을 $C = \{c_{ij}\}$라고 할 때, c_{ij}는 다음과 같이 정의할 수 있으며,

$$c_{ij} = \sqrt{\frac{2}{Q}} \cos\left(\frac{\pi(i-1)(j-0.5)}{Q}\right), \quad 1 \le i \le D, \quad 1 \le j \le Q \tag{8}$$

앞서 구한 가중 로그 필터뱅크 에너지 벡터 $\widehat{x^l}$에 대해 다음과 같이 DCT 변환을 적용하여 D차의 가중 멜켑스트럼 특징벡터를 얻는다.

$$\widehat{x^c} = C \cdot \widehat{x^l} \tag{9}$$

3.2. HMM 파라미터 변환

HMM 상태 s의 가우시안 혼합밀도 함수의 평균 벡터를 μ, 대각 공분산 행렬을 Σ라고 하면 이 상태에 대한 멜켑스트럼 벡터 x의 로그 우도는 다음과 같이

계산된다.

$$\log \Pr(\boldsymbol{x}|s) = -0.5((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \log|\boldsymbol{\Sigma}|) + C \tag{10}$$

위 식에서 C는 상수항이다. 여기서 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$는 인식기의 학습 단계에서 결정되는 모델 파라미터로서 앞 절에서 기술한 가중치가 모두 1인 경우이다. 만약 입력 음성으로부터 구한 가중치 $\boldsymbol{W}$를 적용하여 가중 멜켑스트럼 벡터 $\hat{\boldsymbol{x}}$을 구하였다면, 위의 로그 우도 계산식에서 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에 대해서도 동일한 가중치를 적용하여 $\hat{\boldsymbol{\mu}}$ 및 $\hat{\boldsymbol{\Sigma}}$를 얻어야 한다. 켑스트럼 영역에서의 평균 벡터 및 공분산 행렬을 각각 $\boldsymbol{\mu}^c$, $\boldsymbol{\Sigma}^c$라고 할 때, 제안한 모델 파라미터의 변환 방식은 다음과 같다.

단계 1. 모델 파라미터의 DCT 역변환

$$\begin{aligned} \boldsymbol{\mu}^l &= C^{-1}\,\boldsymbol{\mu}^c \\ \boldsymbol{\Sigma}^l &= C^{-1}\,\boldsymbol{\Sigma}^c\,(C^{-1})^t \end{aligned} \tag{11}$$

단계 2. 필터뱅크 가중

$$\begin{aligned} \hat{\boldsymbol{\mu}}^l &= \boldsymbol{W}\,\boldsymbol{\mu}^l \\ \hat{\boldsymbol{\Sigma}}^l &= \boldsymbol{W}\,\boldsymbol{\Sigma}^l\,\boldsymbol{W}^t \end{aligned} \tag{12}$$

단계 3. 모델 파라미터의 DCT 변환

$$\begin{aligned} \hat{\boldsymbol{\mu}}^c &= C\,\hat{\boldsymbol{\mu}}^l \\ \hat{\boldsymbol{\Sigma}}^c &= C\,\hat{\boldsymbol{\Sigma}}^l\,C^t \end{aligned} \tag{13}$$

단계 4. determinant $|\hat{\boldsymbol{\Sigma}}^c|$ 계산

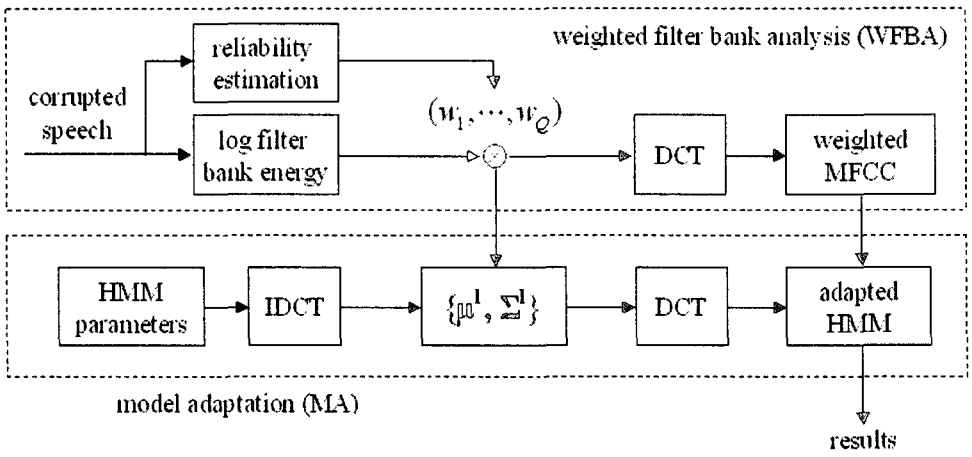
첫 번째 단계에서는 코사인 역변환에 의해 켑스트럼 영역의 모델 파라미터들을 로그 스펙트럼 영역으로 변환한다. 그 다음으로 로그 필터뱅크 에너지 벡터에 입력 신호에서 획득한 가중치를 적용하며, 마지막으로 코사인 변환에 의해 다시 가중 멜켑스트럼 모델 파라미터를 얻는다.

위 식 11에서 식 13까지를 하나의 변환 행렬 V로 표현하면 식 14와 같으며,

이 행렬을 이용하여 모델 파라미터를 켑스트럼 영역에서 직접 변환함으로써 동일한 가중 효과를 얻을 수 있다.

$$\hat{\mu}^c = C W C^{-1} \mu^c = V \mu^c \qquad (14)$$
$$\hat{\Sigma}^c = V \Sigma^c V^t$$

그림 2는 지금까지 기술한 WFBA-MA 기법의 블럭도를 나타낸다.



<그림 2> 제안한 WFBA-MA 방식의 블럭도

4. 실험 및 결과

제안한 방법을 평가하기 위해 TIDIGITS 데이터베이스를 사용하였다. 이 데이터베이스는 "zero"부터 "nine" 및 "oh"의 11개 영어 숫자음 발성을 포함하고 있으며, 학습 자료는 8623개의 연속 숫자음으로 구성되어 있고, 평가 자료는 8700개의 연속 숫자음으로 이루어져 있다. 본 연구에서는 전체 학습자료를 사용하여 각 숫자별로 연속 확률밀도 HMM 단어모델을 학습하였으며, 각 HMM은 상태별로 16개의 Gaussian 혼합밀도 함수를 갖는 10개의 상태로 구성하였다. 인식 성능의 평가를 위해서는 평가 자료 중 2486개의 단일 숫자음 자료를 사용하였고, 이 자료에 대해 기존의 인식방식인 전대역(full-band) 인식기와 3개 및 4개의 부대역으로 구성된 다중대역 인식기, 그리고 제안한 WFBA-MA 방식을 적용한 인식기를 비교 평가하였다. 다중대역 인식기의 대역분할은 표 1에 나타낸 바와 같으며, 부대역 인식결과의 통합을 위해서는 SNR 가중방식을 적용하였다. 전대역 인식기 및 제안한 시스템은 24차의 멜 필터뱅크 에너지로부터 12차의 MFCC를 추출하였고, 3-band 및 4-band 시스템은 각 부대역별로 4차 및 3차의 MFCC를 추출하여 모든 비교 시스

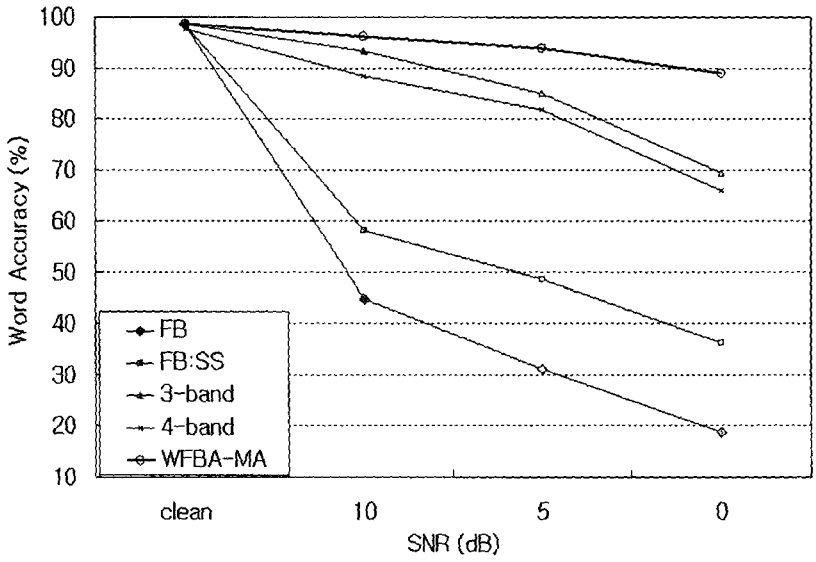
템이 동일한 차수의 MFCC를 사용하도록 하였다.

<표 1> 3대역 및 4대역 multi-band ASR 시스템의 대역분할

	부대역 주파수 범위 (Hz)
3-band	0-1155, 1050-2996, 2723-8000
4-band	0-950, 850-1860, 1691-3625, 3295-8000

첫 번째 실험에서는 1500-1800 Hz 대역제한 잡음을 사용하여 평가자료를 SNR 0, 5, 10 dB로 손상시켰다. 이 잡음은 3-band 시스템의 두 번째 부대역을 15% 손상시키며, 4-band 시스템의 두 번째 부대역의 36% 및 세 번째 부대역의 6%를 손상시킨다. 그림 3은 비교한 시스템들의 단어 인식률을 보인다. 이 결과에서 음성의 일부 주파수 대역이 손상된 경우에는 기존의 전대역 인식(FB) 및 전대역 인식에 기존의 대표적 잡음처리 방법인 스펙트럼 차감법을 적용한 경우(FB:SS)에 비해 다중대역 방식 및 제안한 WFBA-MA 방식은 덜 손상된 음성 대역을 보다 효과적으로 사용함으로써 월등한 인식성능을 나타냄을 알 수 있다. 또한, 3-band 시스템은 잡음이 두 번째 부대역 내부에 발생하였으므로 해당 부대역만 손상된 반면, 4-band 시스템은 발생한 잡음이 두 번째 및 세 번째 부대역 사이의 경계선 상에서 발생하여 두 대역이 모두 손상되었으므로 3-band 시스템에 비해 성능이 저하됨을 알 수 있다. 따라서, 다중대역 인식시스템은 잡음의 발생위치에 따라 성능에 영향을 받음을 확인할 수 있었다. 제안한 WFBA-MA 방식은 잡음의 발생 위치에 상관없이 손상되지 않은 주파수 대역의 음성 정보를 최대한 활용하므로 다중대역 시스템들에 비해 보다 높은 인식 성능을 나타내었다.

두 번째 실험에서는 500-800 Hz 및 2500-2800 Hz의 음성대역을 손상시키는 대역제한 잡음을 사용하여 동일한 실험을 수행하였다. 표 2에서 나타낸 바와 같이 3-band 시스템은 세 개의 부대역들이 모두 잡음에 손상되어 이 경우 전대역 시스템과 비슷한 성능을 나타냄을 알 수 있다. 반면에 4-band 시스템은 두 번째 및 네 번째 부대역에서 손상되지 않은 음성정보를 이용하여 높은 인식률을 얻을 수 있었다. 표 2의 마지막 두 줄은 제안한 WFBA-MA 방식에서 HMM 상태의 평균 벡터 및 공분산 행렬을 모두 가중한 경우(mean:cov)와 평균 벡터만을 가중한 경우로서 이 방식의 성능 향상이 주로 평균 벡터의 변환에 의해 달성됨을 알 수 있었다.



<그림 3> 대역제한 잡음에 의해 1500-1800 Hz의 주파수 대역이 손상된 경우에 대한 전대역 인식(FB), 스펙트럼 차감법(FB:SS), 3-band 및 4-band 인식 및 제안한 방식(WFBA-MA)의 인식성능 비교

<표 2> 평가 음성의 500-800 Hz 및 2500-2800 Hz가 손상된 경우에 대한 인식 시스템들의 단어 인식률(%) 비교

	clean	10 dB	5 dB	0 dB
FB	98.5	10.4	9.0	7.2
FB:SS	98.4	17.3	17.5	16.7
3-band	98.7	9.5	9.2	9.3
4-band	97.8	61.7	59.1	48.8
WFBA-MA (mean:cov)	98.6	83.5	89.2	85.8
WFBA-MA (mean)	98.7	88.1	91.2	77.2

세 번째 실험은 전체 평가 자료에 Volvo 자동차 소음을 추가하고, 전대역 인식 시스템, 3-band 및 WFBA-MA 시스템에 스펙트럼 차감법을 동일하게 적용하여 성능을 비교하였다. 표 3에 정리한 실험 결과에서 3-band 시스템은 음성의 전대역에 잡음이 존재하는 Volvo 잡음의 경우, 비교한 시스템들에 비해 다소 낮은 인식률을 나타내는 반면에 제안한 WFBA-MA는 이 경우에도 보다 월등한 인식 성능을 나타내었다.

<표 5> Volvo 자동차 소음에 대한 각 시스템들의 단어 인식률(%) 비교

	clean	10 dB	5 dB	0 dB
FB:SS	98.4	92.0	88.1	84.2
3-band:SS	98.6	90.9	84.8	78.8
WFBA-MA:SS	98.6	95.5	91.6	85.2

5. 결 론

본 연구에서 제안한 WFBA-MA 기법은 기존의 전대역 음성인식에서 필터뱅크 분석 단계에 가중항을 도입하여 주파수 영역에서 부분 손상된 음성의 인식 성능을 향상시킬 수 있었다. 제안한 방법은 다중대역 음성인식에서처럼 인위적인 부대역의 구분을 필요로 하지 않으면서도 손상되지 않은 주파수 영역의 음성정보를 더욱 효과적으로 활용할 수 있었다. 향후에는 제안한 방법을 동적 MFCC 특징벡터를 포함하도록 확장하고, 필터뱅크 채널의 신뢰도를 보다 효과적으로 표현하는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] Cook, M., P. Green, L. Josifovski and A. Vizinho (2001), Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Communication 34*, pp.267~285.

[2] Morris, A., A. Hagen and H. Bourlard (1999), The Full Combination Sub-bands Approach to Noise Robust HMM/ANN based ASR, in *Proc. of European Conference on Speech Communication and Technology*, pp.599~602.

[3] J. B. Allen (1994), How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing 2(4)*, pp.567~577.

[4] Hermansky, H., S. Tibrewala, and M. Pavel (1996), Towards ASR on Partially Corrupted Speech, in *Proc. of International Conference on Spoken Language Processing 1*, pp.462~465.

[5] 조훈영, 지상문, 오영환(2002), 다중대역 음성인식을 위한 부대역 신뢰도의 추정 및 가중, 「한국음향학회」 21(6), pp.552~558.

[6] Morris, A., A. Hagen, H. Glotin, and H. Bourlard (2001), Multi-stream adaptive evidence combination for noise robust ASR, *Speech Communication 34*, pp.25~40.

[7] Ming, J., P. Jancovic, and F. J. Smith (2002), Robust Speech Recognition Using Probabilistic Union Models, *IEEE Trans. on Speech and Audio Processing 10(6)*, pp.403~414.

[8] Okawa, S., E. Bocchieri, and A. Potamianos (1998), Multi-band speech recognition in noisy environments, in *Proc. of International Conference on Acoustics, Speech and Signal*

Processing, pp.641~644.

[9] Okawa, S., T. Nakajima, and K. Shirai (1999), A Recombination Strategy for Multi-band Speech Recognition based on Mutual Information Criterion, in *Proc. of European Conference on Speech Communication and Technology*, pp.603~606.

[10] H. Y. Cho, L. Y. Kim, and Y. H. Oh (2002), Segmental reliability weighting for robust recognition of partly corrupted speech, *Electronics Letters 38(12)*, pp.611~612.

[11] 조훈영(2003), 부분 정보 기법에 기반한 강인한 음성인식, 박사학위논문, 한국과학기술원 전자전산학과.

[12] Hagen, A., H. Bourlard, and A. Morris (2001), Adaptive ML-weighting in multi-band recombination of Gaussian misture ASR, in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp.257~260.

접수일자: 2002년 10월 22일
게재결정: 2002년 12월 12일

▶ 조훈영 (Hoon-Young Cho)
주소: 대전시 유성구 구성동 373-1
소속: 한국과학기술원 전자전산학과 전산학전공
전화: 042) 869-3556
Fax: 042) 869-3510
E-mail: hycho@bulsai.kaist.ac.kr

▶ 오영환 (Yung-Hwan Oh)
주소: 대전시 유성구 구성동 373-1
소속: 한국과학기술원 전자전산학과 전산학전공
전화: 042) 869-3516
Fax: 042) 869-3510
E-mail: yhoh@bulsai.kaist.ac.kr