한국음향학회지 제22권 제4호 pp. 315~322 (2003)

부분 손상된 음성의 인식 향상을 위한 채널집중 MFCC 기법

Channel-attentive MFCC for Improved Recognition of Partially Corrupted Speech

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

*한국과학기술원 전자전산학과 전산학전공, **경성대학교 컴퓨터과학과

(접수일자: 2003년 3월 25일; 채택일자: 2003년 4월 24일)

본 논문에서는 주파수 영역의 일부가 상대적으로 심하게 손상된 음성에 대한 음성 인식기의 성능을 향상시키기 위해 채널집중 멜 켑스트럼 특징추출법을 제안한다. 이 방법은 기존 멜 켑스트럼 특징추출의 필터뱅크분석 단계에서 각 채널의 신뢰도를 구하고, 신뢰도가 높은 주파수 영역이 음성인식에 보다 중요하게 사용되도록 멜 켑스트럼 추출 및 HMM의 출력확률 계산식에 채널 가중을 도입한다. TIDIGITS 데이터베이스에 음성의 일부 주파수를 손상시키는 다양한 주파수선택 잡음을 가산하여 인식 실험을 수행한 결과, 제안한 방법은 덜 손상된 주파수영역의 음성 정보를 효과적으로 활용하며, 주파수선택 잡음에 대해 우수하다고 알려진 다중대역 음성인식에 비해 평균 11.2% 더 높은 성능을 얻었다.

핵심용어: 부분 손상된 음성, 주파수선택 잡음, 다중대역 음성인식, 채널집중 MFCC
투고분야: 음성처리 분야 (2,5)

We propose a channel-attentive Mel frequency cepstral coefficient (CAMFCC) extraction method to improve the recognition performance of speech that is partially corrupted in the frequency domain. This method introduces weighting terms both at the filter bank analysis step and at the output probability calculation of decoding step. The weights are obtained for each frequency channel of filter bank such that the more reliable channel is emphasized by a higher weight value. Experimental results on TIDIGITS database corrupted by various frequency-selective noises indicated that the proposed CAMFCC method utilizes the uncorrupted speech information well, improving the recognition performance by 11.2% on average in comparison to a multi-band speech recognition system.

Keywords: *Partially corrupted speech, Frequency-selective noise, Multi-band speech recognition, Channel-attentive MFCC*

ASK subject classification: *Speech signal processing (2,5)*

I. 서 론

본 논문에서는 사이렌, 자동차 경적소리, 전화기의 DTMF 신호음 및 벨소리와 같이 음성의 주파수 영역 일부를 손상시키는 잡음에 대해 음성인식 시스템의 성능을 향상시키는 방법에 중점을 둔다. 이러한 주파수선택 (frequency-selective) 잡음은 실세계에서 빈번히 발생되는 신호로서 주파수 영역의 임의 위치에서 발생하여 음성의 일부분을 상대적으로 심하게 손상시킨다. 음성인식기의 성능 저하를 막기 위하여 스펙트럼 차감법 (spectral subtraction)과 같이 무음구간에서 잡음을 추정하여 제거하는 방식 및 병렬 모델 조합 (parallel model combination; PMC)처럼 잡음 신호에 대해 인식 모델을 적응하는 방법을 고려해 볼 수 있다[1,2]. 이때, 전자의 경우는 잡음 음성으로부터 평균 잡음 스펙트럼을 차감하는 과정에서 잡음에 의해 손상되지 않은 주파수 영역의 정보가 손실될 가능성이 있으며, 후자의 경우는 잡음의

책임저자: 조훈영 (hycho@bulsai.kaist.ac.kr)
305-701 대전광역시 유성구 구성동 373-1
한국과학기술원 전자전산학과
(전화: 042-869-3556; 팩스: 042-869-3510)

신호를 미리 예측하여 적응 자료를 수집해야 한다는 어려움이 있다.

·손실 데이터 기법 (missing data technique) 또는 다중대역 음성인식 (multi-band speech recognition)은 음성의 시간주파수 영역에서 잡음에 의해 손상되지 않은 부분정보 (partial information)를 극대화하여 인식하는 새로운 음성인식 패러다임으로서 비교적 최근에 활발히 연구되어 왔다. 손실 데이터 기법은 음성의 스펙트로그램에서 손상된 영역을 찾고, 이 영역을 출력확률 계산 단계에서 제외하거나, 손상되기 이전의 값을 추정 및 대치하여 인식하는 방식이다[3]. 이 방법은 잡음에 대한 가정이 불필요하다는 장점이 있으나, 직교화된 특징벡터를 사용하기 어려우므로 잡음이 없는 음성에 대해서 상대적으로 낮은 성능을 보인다는 단점이 있다[4]. 다중대역 음성인식은 Fletcher의 다중독립 채널 모형 (multi-independent channel model)에 기반하여 음성의 전체 주파수 대역을 다수의 부대역 (sub-band)으로 나누고, 각 부대역에 대해 독립적으로 음성인식을 수행한 후 부대역 인식결과를 통합하여 최종적인 인식 결과를 얻는 방식으로서 주파수 영역의 일부가 상대적으로 심하게 손상된 경우에 매우 효과적이라고 알려졌다[5,6,13]. 그러나 부대역의 경계선이 학습 단계에서 결정되고 고정적이어서 인식기 사용시에 임의의 주파수 영역에서 발생한 주파수선택 (또는 대역제한) 잡음을 보다 효과적으로 국부화 (localize)할 수 없다[7,8]. 또한 각각의 부대역에서 독립적인 특징을 추출 및 인식하므로 부대역 간의 상관 정보 (correlation information)가 손실되는데, 이 정보는 광대역 잡음환경에서 음성을 인식함에 있어 중요한 정보라고 알려졌다[9].

본 연구에서는 다중대역 음성인식의 고정된 부대역 경계에 의해 발생하는 국부화 능력의 한계를 해소함과 동시에 기존의 전대역 (full-band) 음성인식방식에 부분정보의 강조기능을 추가하기 위해 채널집중 멜 켑스트럼 특징 추출 및 HMM (hidden Markov model)의 변형된 출력확률 계산식을 제안한다. 본 논문의 2장에서는 다중대역 음성인식에 관하여 보다 자세히 기술하고, 본 논문에서 검토한 문제점을 언급한다. 제 3장에서는 제안한 채널집중 멜 켑스트럼 특징추출방법을 설명하며, 제 4장과 5장에서 실험 및 결과를 기술하고 결론을 맺도록 한다.

II. 다중대역 음성인식

인간의 음소인식 원리에 대한 Fletcher의 다중독립 채널 모형은 식 (1)과 같이 기술될 수 있으며, 이를 Fletcher의 다중독립 채널 (multi-independent channel; MIC) 모형 또는 오류적 (product-of-error; PoE) 법칙이라고도 한다[5,10].

$$e_F = e_1 e_2 \cdots e_B = \prod_{b=1}^{B} e_b \qquad (1)$$

위 식에서 e_F와 e_b는 전대역 및 b번째 부대역의 인식 오류율이며, 특정 주파수 영역에서의 부분 인식 오류 (partial recognition error)는 다른 주파수 영역에서의 부분 인식 오류에 영향을 미치지 않는다. 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[10]. 따라서 인간은 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 않고 있으며, 비교적 손상정도가 적은 주파수 영역을 강조하여 인식하는 능력을 기존의 음성인식기에 부여할 필요가 있다.

Fletcher의 MIC 모형에 기반하여 음성의 전체 주파수 대역을 다수의 부대역으로 나누고, 부대역별로 독립적인 인식을 수행한 후에 인식결과를 통합하는 다중대역 음성인식 관련 연구결과가 활발히 발표되었다. 그림 1은 현재 일반적으로 널리 사용되고 있는 전대역 특징기반의 음성인식 시스템과 다중대역 음성인식의 차이를 나타낸다.

다중대역 음성인식에서 $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_B$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.

$$\log \Pr(\mathbf{X} \mid \lambda) = \sum_{b=1}^{B} w_b \cdot \log \Pr(\mathbf{X}_b \mid M_j^b) \qquad (2)$$

다중대역 음성인식에서 주파수 부대역의 개수로는 주로 2개에서 7개 사이를 주로 사용해 왔으며, 부대역의 범위는 멜 단위 주파수 범위를 부대역 개수로 균등하게 분할하여 결정한 경우가 많았다. 부대역 인식결과의 신뢰도 혹은 가중치로는 부대역의 SNR, 부대역 상호 정보 (mutual information) 및 최대 우도를 정규화하여 가중하는 방식들이 제안되었다[9,11,12]. 부대역 인식 결과의 통합방법으로는 식 (2)와 같은 선형 가중 통합, MLP (multi-layer perceptron) 등의 비선형 통합 외에도 부대역들의 모든 가능한 조합에 대해 최적의 조합을 선택하는 전조합

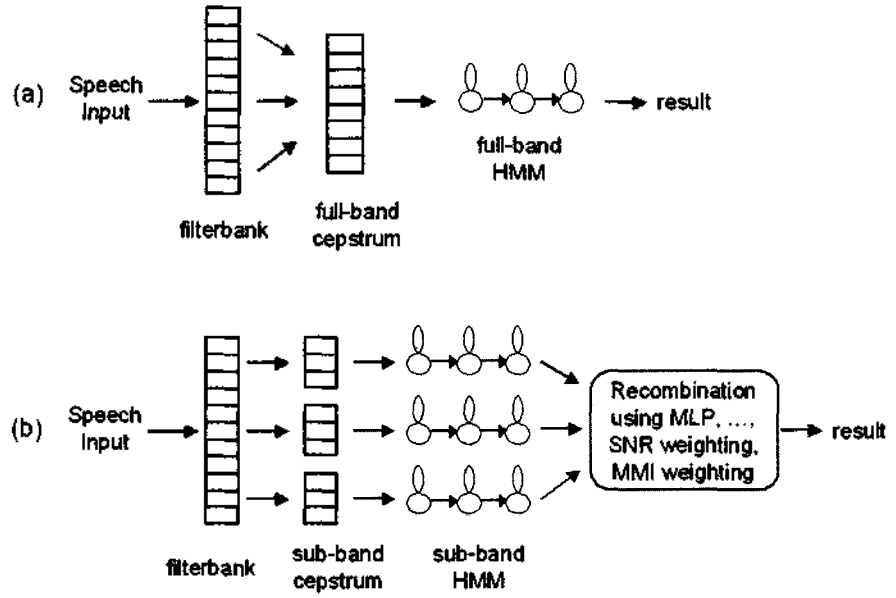

그림 1. 전대역 음성인식과 다중대역 음성인식 방식의 비교 (a) 전대역 음성인식 (b) 다중대역 음성인식
Fig. 1. Diagrams of (a) full-band speech recognition (b) multi-band speech recognition.

(full combination) 및 부대역의 신뢰도를 별도로 추정할 필요가 없는 PUM (probabilistic union model) 방식 등이 연구되었다[4,7]. 현재 많은 인식 시스템들이 채택하고 있는 전대역 (full-band) 음성인식 방식에서는 전체 주파수 대역의 전반적인 스펙트럼 형태를 표현하는 p차의 직교화된 특징벡터를 사용하므로 스펙트럼의 일부 영역에 오류가 발생할 경우에도 p차의 벡터요소 전체에 오류가 전파된다. 반면에 다중대역 시스템은 다수의 부대역에 대해 독립적인 특징을 추출하여 독립적으로 인식하므로 손상되지 않은 주파수 영역의 음성정보를 최대한 활용할 수 있다[6].

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 좁은 경우, 해당 부대역 내부에 손상되지 않은 음성 정보가 존재함에도 불구하고 이를 효과적으로 활용할 수 없다. 둘째, 광대역 잡음의 경우 부대역 음성정보간의 상관 (correlation) 정보가 인식에 중요하게 사용된다. 그러나 각 부대역에서 독립적인 특징을 추출하는 다중대역 인식에서는 부대역 간 상관정보를 잘 활용하지 못하므로 광대역 잡음에 대해서는 기존의 전대역 인식방식에 비해 그다지 효과적이지 않다고 알려졌다[9]. 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.

III. 채널집중 멜 켑스트럼

본 장에서는 전대역 음성인식 방식을 유지하면서도 앞 절에서 언급한 다중대역 음성인식 방식의 한계를 극복할 수 있는 채널집중 멜 켑스트럼 (channel-attentive Mel frequency cepstral coefficient; CAMFCC) 특징추출 과정과 변형된 HMM 출력확률 계산식을 소개한다.

3.1. 채널집중 멜 켑스트럼 추출

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

$$x_i = \sum_{k=1}^{K} |X(k)|^2 \cdot \varphi_i(k) \tag{3}$$

$$x_i^l = \log(x_i),\ 1 \le i \le Q \tag{4}$$

의 식에서 k는 FFT 인덱스, $\varphi_i(k)$는 i번째 멜 대역통과 필터를 나타낸다. 위 첨자 l은 이 변수가 로그 스펙트럼 영역의 값임을 의미한다. 위 식에서 구한 Q차의 로그 필터뱅크 에너지 벡터를 $\mathbf{x}^l=(x_1^l, x_2^l, \cdots, x_Q^l)^t$라고 할 때, 이 벡터의 요소들은 채널의 신뢰도에 따라 음성인식에 유효한 정보를 다른 정도로 포함하게 된다. 제안한 CAMFCC 특징추출은 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 $(w_1, w_2, \cdots, w_Q)$이라 할 때, 이 값들로부터 대각 가중 행렬 $\mathbf{W}=diag(w_1, w_2, \cdots w_Q)$를 정의할 수 있다. 이 행렬을 이용하여 가중 로그 필터뱅크 에너지 벡터 $\hat{\mathbf{x}}^l$을 다음과 같이 구한다.

$$\hat{\mathbf{x}}^l = \mathbf{W}\cdot\mathbf{x}^l=\begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_Q \end{bmatrix}\begin{bmatrix} x_1^l \\ x_2^l \\ \vdots \\ x_Q^l \end{bmatrix} \tag{5}$$

본 연구에서는 각 채널의 신호대 잡음비 (signal-to-noise ratio; SNR)를 스펙트럼 차감법을 이용하여 계산하고, 다음 식과 같은 시그모이드 함수를 이용하여 SNR 0 dB에서 30 dB 사이를 0과 1사이의 값이 되도록 정규화하였다[5].

$$w_i=\frac{1}{1+\exp(-\alpha(\rho_i-15))} \tag{6}$$

위 식에서 α는 함수의 0에서 1사이 값의 기울기를 조절하는 변수로서 본 연구에서는 0.3을 실험적으로 결정하여 사용하였고, ρ_i는 채널의 SNR이다. 켑스트럼 추출의 마지막 단계로 이산 코사인 변환 (discrete cosine transform; DCT) 행렬을 $\mathbf{C}=\{c_{ij}\}$라고 할 때, c_{ij}는 다음과 같이 정의할 수 있다.

$$c_{ij}=\sqrt{\frac{2}{Q}}\cdot\cos\left(\frac{\pi(i-1)(j-0.5)}{Q}\right), 1\le i\le D, 1\le j\le Q \tag{7}$$

앞서 구한 가중 로그 필터뱅크 에너지 벡터 $\hat{\mathbf{x}}^l$에 대해 다음 식 (8)과 같이 DCT 변환을 적용하여 D차의 CAMFCC 특징벡터를 얻는다.

$$\hat{\mathbf{x}}=\mathbf{C}\cdot\hat{\mathbf{x}}^l \tag{8}$$

그림 2에서 지금까지 기술한 CAMFCC 추출과정을 도식화하였다.

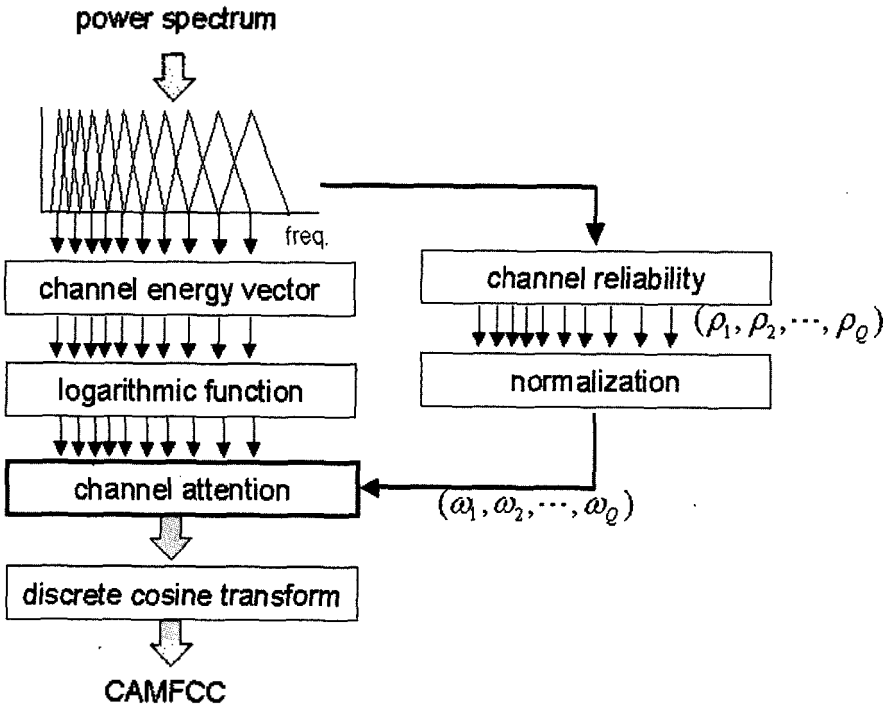

그림 2. 채널집중 멜 켑스트럼 특징추출
Fig. 2. Overview of channel-attentive MFCC feature extraction.

3.2. 변형된 HMM 출력확률 계산식

HMM 상태 s의 평균벡터 및 공분산 행렬을 각각 μ와 $\sum$로 표기할 때, 이 상태에 대한 멜 켑스트럼 특징벡터 $\mathbf{x}$의 로그 출력 확률은 식 (9)와 같다. 이 식에서 K는 상수항이며, 앞으로 수식의 간결한 표현을 위해 상수항 K는 제외하기로 한다.

$$\log \Pr(\mathbf{x} \mid s) = -0.5((\mathbf{x}-\mu)' \sum^{-1} (\mathbf{x}-\mu)) + K \tag{9}$$

MFCC 벡터 $\mathbf{x}$를 대신하여 식 (8)의 CAMFCC 벡터 $\hat{\mathbf{x}}$을 사용할 경우, 상태 s의 평균벡터 μ에 보정 행렬 $\mathbf{V}$를 적용하여 변형한 HMM 출력 확률식은 다음과 같다.

$$\begin{aligned}\log \Pr(\hat{\mathbf{x}} \mid s) &= -0.5((\hat{\mathbf{x}}-\mathbf{V}\mu)'\Sigma^{-1}(\hat{\mathbf{x}}-\mathbf{V}\mu)) \\ &= -0.5((\hat{\mathbf{x}}-\hat{\mu})'\Sigma^{-1}(\hat{\mathbf{x}}-\hat{\mu})) \\ &= -0.5((\mathbf{C}\mathbf{W}\mathbf{x}'-\mathbf{C}\mathbf{W}\mu')'\Sigma^{-1}(\mathbf{C}\mathbf{W}\mathbf{x}'-\mathbf{C}\mathbf{W}\mu')) \\ &= -0.5\sum_{j=1}^{p}\left[\sum_{k=1}^{p}\frac{c_{jk}w_k(x_k'-\mu_k')}{\sigma_j}\right]^2\end{aligned} \tag{10}$$

위 식에서 $\mathbf{C}$, $\mathbf{W}$, $\mathbf{x}'$은 앞 절과 동일한 의미를 갖는다. 또한, 식에서 k번째 채널이 잡음에 의해 심하게 손상되어 $w_k \simeq 0$ 일 때는 해당 채널의 정보가 인식계산에서 제외되며, $w_k \simeq 1$의 경우는 해당 채널에서의 입력과 인식모델 간의 차이가 인식에 중요하게 사용된다. 식 (10)에서 CAMFCC 벡터 $\hat{\mu}$을 구하기 위해 식 11과 같이 먼저 DCT 역변환에 의해 켑스트럼 평균벡터 μ를 로그 스펙트럼 벡터 μ'로 변환하고, 식 (12)에서 가중행렬 $\mathbf{W}$를 적용한 후, 식 (13)처럼 다시 DCT 변환을 수행한다.

$$\mu' = \mathbf{C}^{-1}\mu \tag{11}$$

$$\hat{\mu}' = \mathbf{W}\mu' \tag{12}$$

$$\hat{\mu} = \mathbf{C}\hat{\mu}' \tag{13}$$

식 (11), (12) 및 (13)에 의해 식 (10)에서 $\hat{\mu} \simeq \mathbf{V}\mu$를 만족하는 보정 행렬 $\mathbf{V}$는 $\mathbf{V}=\mathbf{C}\mathbf{W}\mathbf{C}^{-1}$로 구해진다.

3.3. 동적 파라미터의 추가

동적 특징벡터는 정적 특징벡터의 차분 (difference) 혹은 선형회귀 (linear regression)로 계산된다. 이들 연산자를 $\triangle$로 표기하면 동적 파라미터를 추가한 입력 CAMFCC 벡터는 $[\hat{\mathbf{x}}', (\triangle\hat{\mathbf{x}})']'$로 표현되며, 상태 s의 MFCC 평균 벡터 $[\mu', (\triangle\mu)']'$에 채널 가중을 적용한 CAMFCC 벡터 $[\hat{\mu}', (\triangle\hat{\mu})']'$는 다음 식처럼 μ와 $\triangle\mu$로부터 계산할 수 있다.

$$\begin{aligned}[\hat{\mu}', (\Delta\hat{\mu})']' &= [(\mathbf{V}\mu)', \Delta(\mathbf{V}\mu)']' \\ &= [(\mathbf{V}\mu)', (\mathbf{V}(\Delta\mu))']'\end{aligned} \tag{14}$$

끝으로 기존 MFCC를 사용하는 음성인식의 디코딩 단계에서 출력확률 계산의 전체 회수를 N, 정적 특징벡터의 차수를 d라고 할 때, 제안한 CAMFCC의 변형된 출력확률 계산은 $2d^2N$의 곱셈 연산 및 $2d(d-1)N$의 덧셈 연산이 추가적으로 필요하다.

IV. 실험 및 결과

4.1. 실험 환경

제안한 방법을 평가하기 위해 TIDIGITS 데이터베이스를 사용하였다. 이 데이터베이스는 "zero"부터 "nine" 및 "oh"의 11개 영어 숫자음 발성을 포함하고 있으며, 학습자료는 8623개의 발성음으로서 1자리부터 7자리의 연속 숫자음이 고르게 분포되어 있고, 평가자료는 8700개의 연속 숫자음 자료가 학습자료와 마찬가지로 분포되어 있다. 기본 음성인식 시스템은 12차의 MFCC 및 12차의 delta MFCC를 이용하여 숫자별로 연속 확률밀도 HMM 단어모델을 학습하였으며, HMM은 상태별로 8개의 가우시안 혼합밀도 함수를 갖는 10개의 상태로 구성하였다. 다중대역 시스템은 4개의 부대역으로 구분하였으며, 각 부대역의 SNR을 0에서 1사이의 값으로 정규화하여 부대역 통합시에 가중치로 사용하였다. 부대역 범위는 [0-950], [850-1860], [1691-3625], [3295-8000] Hz이다. 특징벡터로는 부대역마다 6개의 멜 필터뱅크에서 3차의 정적 MFCC를 추출하였고, 3차의 delta 특징을 추가하였다.

인식 성능의 평가를 위해서 TIDIGITS 평가자료 중 2486개의 단독 숫자음 및 단독 숫자음을 포함한 8700개의 전체 연속 숫자음 자료에 잡음을 추가하였다. 실험에 사용한 잡음 신호는 중앙 주파수가 각각 450, 900, 1350, 1770, 2650, 3460 Hz이고 대역폭이 100 Hz인 대역통과 필터들에 가우시안 백색잡음을 통과시켜 획득하였다. 이 신호들은 4개의 부대역을 갖는 다중대역 인식시스템에서 1개, 2개 및 3개의 부대역을 손상시킨다. 중앙 주파수가 450, 1350 및 2650 Hz인 잡음은 각각 부대역 1, 2 및 3을

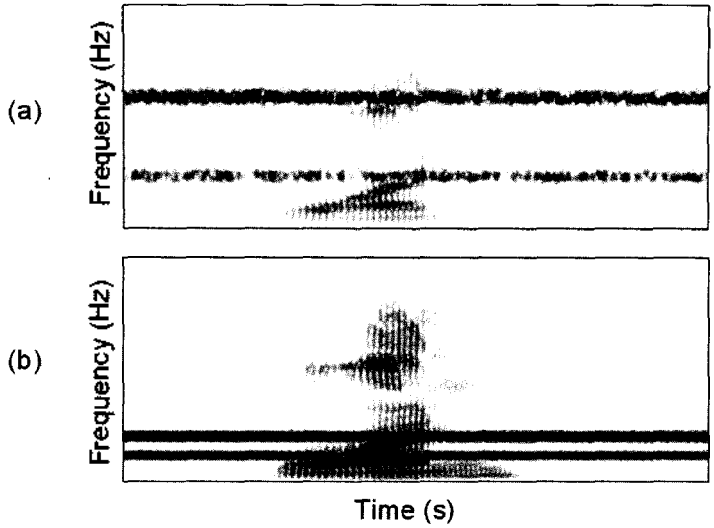

그림 3. SNR 5 dB 잡음에 의해 부분 손상된 음성의 스펙트로그램 (a) 1770 및 3450 Hz 주파수선택 잡음 (b) DTMF 신호
Fig. 3. Spectrograms of partially corrupted speech (at SNR 5 dB) (a) 1770 and 3450 Hz frequency-selective noise (b) DTMF tone.

손상시켜 하나의 부대역을 손상시킨다. 또한 중앙 주파수가 900, 1770 및 3460 Hz인 잡음은 부대역 1과 2, 부대역 2와 3, 그리고 부대역 3과 4사이의 경계선 상에 위치하여 두 개의 부대역을 손상시킨다. 세 개의 부대역을 손상시키는 잡음은 위의 두 종류의 잡음을 가산한 450과 1770 Hz 잡음 (부대역 1, 2, 3을 손상), 1350과 3460 Hz 잡음(부대역 2, 3, 4를 손상)을 사용하였다. 마지막으로 보다 실제적인 잡음에 대한 성능 평가를 위해 전화 번호 1, 5, 9에 해당하는 DTMF 신호음을 가산하였다. 모든 잡음과 평가음성은 SNR 10, 5, 0 dB로 조절되었다.

그림 3은 SNR 5 dB의 1770 및 3460 Hz의 주파수선택 잡음과 DTMF 신호에 의해 손상된 숫자음 "one"의 스펙트로그램을 나타낸다.

4.2. 실험 결과

표 1은 4-band 시스템의 손상된 부대역 개수 및 손상된 정도에 따른 각 방법의 인식률을 비교한 것이다. 이 결과에서 음성의 일부 주파수 대역이 손상된 경우에는 기존의 전대역 인식 (MFCC) 및 전대역 인식에 기존의 대표적 잡음처리 방법 중의 하나인 스펙트럼 차감법을 적용한 경우 (MFCC:SS)에 비해 4-band 다중대역 시스템 및 제안한 CAMFCC 방식이 월등한 인식성능을 나타냄을 알 수 있

표 1. 다양한 주파수선택 잡음에 대한 전대역 음성인식 (MFCC), 전대역 음성인식에 스펙트럼 차감법 적용 (MFCC:SS), 4-band 음성인식 및 제안한 CAMFCC의 성능비교: TIDIGITS 평가자료의 단독 숫자음 (1-dgt) 및 연속 숫자음 (c-dgt)에 대한 단어 인식률(%)
Table 1. Comparison of a full-band ASR (MFCC), a full-band ASR with the spectral subtraction (MFCC:SS), a 4-band ASR and a full-band ASR using CAMFCC under several frequency-selective noise situations: word accuracies (%) for isolated digits (1-dgt) and continuous digits of TIDIGITS database.

# Noisy Bands	SNR (dB)	Multi-band		Full-band					
		4-band		MFCC		MFCC:SS		CAMFCC	
		1-dgt	c-dgt	1-dgt	c-dgt	1-dgt	c-dgt	1-dgt	c-dgt
1	10	95.6	93.2	40.4	34.2	25.8	41.0	97.7	95.2
	5	93.1	89.8	22.3	25.0	19.5	34.4	94.3	91.5
	0	85.4	78.3	15.0	14.4	14.1	25.7	84.6	79.9
2	10	91.5	89.8	44.2	36.1	32.4	41.4	98.5	96.8
	5	88.0	84.8	30.5	30.0	27.5	35.2	96.3	94.0
	0	79.3	71.3	23.1	24.4	21.2	27.9	89.1	83.1
3	10	61.2	59.6	13.5	23.1	14.6	32.9	96.1	90.8
	5	62.7	52.2	11.5	13.5	13.2	24.1	83.1	80.0
	0	55.6	39.3	10.5	8.1	12.2	15.8	60.7	61.3

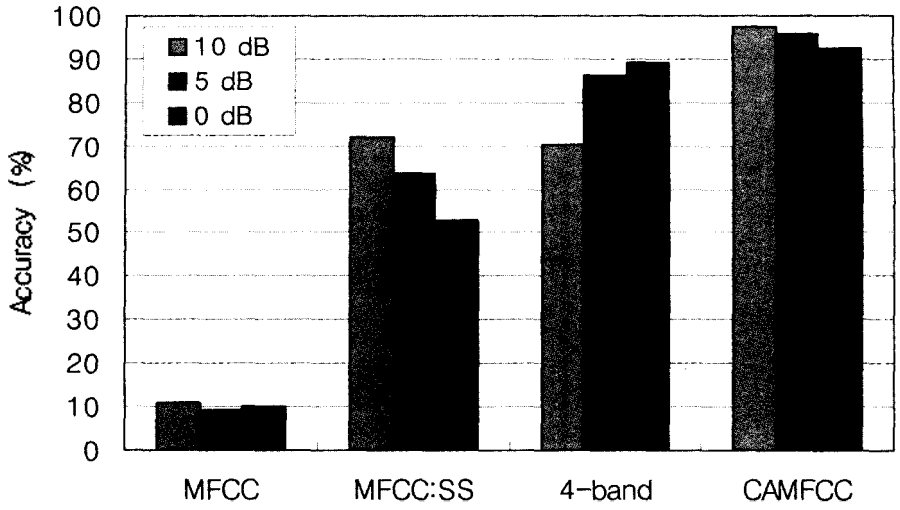

그림 4. 전화번호 1, 5, 9에 해당하는 DTMF 신호에 의해 손상된 단독 숫자음의 인식률 (%)
Fig. 4. Word accuracies (%) of isolated digit utterances corrupted by DTMF tones corresponding to the phone numbers one, five, and nine.

다. 그 이유는 기존 MFCC 특징추출을 사용하는 전대역 음성인식의 경우, 주파수 영역의 일부가 손상된 경우에도 전체 특징벡터요소가 영향을 받았기 때문이며, 4-band 및 CAMFCC의 경우 손상된 영역의 음성정보를 인식에 덜 중요하게 사용하므로 높은 성능을 보였다.

표 1의 4-band 시스템 인식결과에서 손상된 부대역의 개수가 1개인 경우보다 2개인 경우에 성능이 저하되었다. 이는 주파수선택 잡음의 대역폭은 동일하고 단지 주파수상의 위치만 다른 경우로서, 다중대역 시스템은 부대역 경계선이 고정되어 잡음에 대한 국부화 (localize)가 불가능하므로, 동일한 정도의 손상이 부대역의 경계선부근에서 발생할 경우 두 부대역의 인식성능이 동시에 저하되는 단점을 갖음을 알 수 있다. 이런 문제점을 극복하기 위해 보다 많은 부대역으로 구분하는 것을 고려해 볼 수 있으나, 제한된 전체 주파수를 많은 수의 부대역으로 나눌 경우, 각각의 부대역 인식성능이 매우 낮아진다는 한계가 있다. 반면에 표 1에서 제안한 CAMFCC 방법은 잡음의 주파수 위치에 무관하게 성능을 향상시킬 수 있음을 알 수 있다. 또한, 표에서 잡음처리를 하지 않은 MFCC에 비해 스펙트럼 차감법을 적용하였을 때 단독 숫자음 인식성능이 오히려 저하되었다. 이 경우, 잡음에 의해 손상되지 않은 주파수 대역에 존재하는 유효한 음성정보도 상당량이 손실되어 오히려 성능의 저하가 발생되었다고 판단된다. 단독 숫자음 및 연속 숫자음 인식결과를 모두 고려할 때 제안한 CAMFCC 방식은 다중대역 인식방식에 비해 평균 11.2%정도 높은 인식성능을 나타내었다.

그림 4는 세 종류의 DTMF 신호음에 대한 단독 숫자음 인식결과를 보인다. 이 경우에도 제안한 CAMFCC가 다중대역 및 그 외의 방법들에 비해 높은 성능을 보였다. 4-band 시스템의 경우 잡음의 크기가 클수록 인식률이 오히려 높아지는 것을 볼 수 있는데, 이는 손상된 부대역을 크게 감쇠시킬수록 그 외의 부대역들의 정보가 더욱 중요하게 사용되어 오히려 정답을 찾는 경우가 발생하기 때문이라 판단된다.

V. 결 론

본 논문에서는 주파수 영역의 일부분을 상대적으로 심하게 손상시키는 종류의 잡음에 강한 음성인식 방법을 제안하였다. 제안한 CAMFCC는 기존의 멜 켑스트럼 특징추출의 필터뱅크 분석 단계에 각 채널의 신뢰도에 따른 가중치를 도입한 것으로서 별도의 부대역 구분이 불필요하며 기존의 전대역 인식기 구조를 유지하면서도 다중대역 음성인식보다 높은 성능을 나타내었다. 추후 연구로는 채널의 신뢰도 혹은 정보량을 보다 적절히 표현할 수 있는 방법에 관한 연구가 필요하며, 주파수 선택 잡음뿐만 아니라 광대역 잡음에 대해서도 기존 잡음처리 방법과 함께 적용하여 성능을 향상시킬 수 있을 것으로 기대된다.

참 고 문 헌

1. S. Boll, "Suppression of acoustic noise in speech using

spectral subtraction," *IEEE Trans. On Speech and Audio Processing*, **27** (2), 113–120, 1979.
2. 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.
3. M. Cook, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, **34**, 267–285, 2001.
4. A. Morris, A. Hagen and H. Bourlard, "The full combination sub-bands approach to noise robust HMM/ANN based ASR," *Proc. of European Conference on Speech Communication and Technology*, 599–602, 1999.
5. J. B. Allen, "How do humans process and recognize speech?" *IEEE Trans. On Speech and Audio Processing*, **2**, (4), 567–577, 1994.
6. H. Hermansky, S. Tibrewala and M. Pavel, "Towards ASR on partially corrupted speech," *Proc. of International Conference on Spoken Language Processing*, **1**, 462–465, 1996.
7. J. Ming, P. Jancovic and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Trans. On Speech and Audio Processing*, **10** (6), 403–414, 2002.
8. H.-Y. Cho, *Robust Speech Recognition based on Partial Information Technique*, Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Division of Computer Science, KAIST, 2003.
9. S. Okawa, E. Bocchieri and A. Potamianos, "Multi-band speech recognition in noisy environments," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 641–644, 1998.
10. A. Morris, A. Hagen, H. Glotin and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, **34**, 25–40, 2001.
11. S. Okawa, T. Nakajima and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," *Proc. of European Conference on Speech Communication and Technology*, **2**, 603–606, 1999.
12. A. Hagen, H. Bourlard and A. Morris, "Adaptive ML-weighting in multi-band recombination of gaussian mixture ASR," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, **1**, 257–260, 2001.
13. 조훈영, 지상문, 오영환, "다중대역 음성인식을 위한 부대역 신뢰도의 추정 및 가중," 한국음향학회지, **21** (6), 552–558, 2002.

저자 약력

● 조 훈 영 (Hoon-Young Cho)

1995년 8월: 한국과학기술원 전자전산학과 (학사)
1998년 2월: 한국과학기술원 전자전산학과 (석사)
2003년 2월: 한국과학기술원 전자전산학과 (박사)
2003년 3월~현재: 한국과학기술원 정보전자연구소 Post Doc.
※ 주관심분야: 잡음에 강한 음성인식, 패턴인식, 기계학습

● 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 컴퓨터과학과 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