한국음향학회지 제22권 제1호 pp. 35~41 (2003)

주파수 부대역의 켑스트럼 해상도 최적화에 의한 특징추출

Feature Extraction by Optimizing the Cepstral Resolution of Frequency Sub-bands

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

*경성대학교 정보과학부, **한국과학기술원 전자전산학과 전산학전공

(접수일자: 2002년 9월 3일; 수정일자: 2002년 11월 15일; 채택일자: 2002년 11월 27일)

일반적인 음성인식 방법에서는 주파수 전대역에서 추출한 특징벡터를 사용하므로, 각 주파수 부대역은 최종인식 결과에 동등하게 기여한다. 본 논문에서는 주파수 부대역별로 독립적인 특징을 추출하고, 음성인식에 효과적이 되도록 부대역의 켑스트럼 해상도를 조절하는 방법을 제안한다. 주파수 부대역별로 독립적인 특징을 추출하는 멀티밴드 음성인식접근을 사용하여 부대역 특징벡터의 차원을 변화시킨다. 최적의 벡터 차원 조합을 찾기 위하여 음성인식률과 군집화 품질을 사용한다. TIDIGITS 연결 숫자음을 사용한 실험결과에서, 제안한 방법은 전대역 특징추출에 비해 적은 계산량으로도 숫자열 인식률은 99.12%, 백분율 정확도 (percent correct)는 99.775%, 백분율 정밀도 (percent accuracy)는 99.705%를 얻었으며 ,이는 전대역 특징벡터에 비해 상대적 오류율을 각각 38%, 32%, 37% 감소시킨 결과이다.

핵심용어: 켑스트럼 해상도, 부대역 특징 추출, 멀티밴드 음성인식, 최적의 벡터 차원 조합
투고분야: 음성처리 분야 (2.5)

Feature vectors for conventional speech recognition are usually extracted in full frequency band. Therefore, each sub-band contributes equally to final speech recognition results. In this paper, feature vectors are extracted indepedently in each sub-band. The cepstral resolution of each sub-band feature is controlled for the optimal speech recognition. For this purpose, different dimension of each sub-band ceptral vectors are extracted based on the muti-band approach, which extracts feature vector independently for each sub-band. Speech recognition rates and clustering quality are suggested as the criteria for finding the optimal combination of sub-band vector dimension. In the connected digit recognition experiments using TIDIGITS database, the proposed method gave string accuracy of 99.125%, 99.775% percent correct, and 99.705% percent accuracy, which is 38%, 32% and 37% error rate reduction relative to baseline full-band feature vector, respectively.

Keywords: *Cepstral resolution, Sub-band cepstral vector, Multi-band approach, The optimal combination of sub-band vector dimension*

ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성인식 특징추출 과정의 첫번째 단계는 주파수 분석으로서 멜 스케일 분석, 바아크 스케일 분석이 널리 사용되고, 최적필터뱅크 생성, 필터뱅크 대신에 퓨리에 분석의 결과를 직접 이용하는 방법, lapped transform (LT), wavelet transform (WT)을 이용하는 방법이 있다[1-6]. 멜 스케일과 바아크 스케일 분석은 인간의 청각특성을 반영하여 청각기의 인지 단위를 모의한 멜이나 바아크 단위로 주파수 스케일을 변환하는 방법이다. 이 방법은 선형적인 주파수 간격보다 음성인식에 유용하지만, 변환된 주파수 간격은 음성인식을 위한 정보량 간격은 아니다. 따라서 변환된 주파수 전대역에서 특징을 추출하면,

책임저자: 지상문 (smchiks@ks.ac.kr)
608-736 부산광역시 남구 대연동 110-1
경성대학교 정보과학부
(전화: 051-620-4677; 팩스: 051-622-1078)

음성인식 정보량과 비례하게 주파수 정보가 이용된 것이라 볼 수 없다. 멜이나 바아크 스케일 분석 이외의 주파수 분석방법인 최적 필터뱅크 생성은 인식률을 최대화하도록 필터뱅크의 간격을 최적화하는 방법이고, 필터뱅크의 에너지와 함께 퓨리에 분석의 결과를 직접 이용하는 방법은 부대역 스펙트럼 구조의 상세정보를 추출하기 위한 방법이다. 퓨리에 변환이나 DCT (discrete cosine transform) 같은 블록변환이 분석프레임의 양끝에서 불연속이 발생하는 블러킹 효과를 갖는 반면, LT는 불러킹 효과가 발생하지 않는다. 따라서 LT를 시간축 또는 주파수축 상에서 필터뱅크를 구성할 때 사용하여 왜곡이 없는 필터뱅크를 구성할 수 있다. WT의 경우는 고주파 대역에서는 높은 시간 해상도를 갖고, 저주파 대역에서는 높은 주파수 해상도를 갖도록 다중 해상도를 갖는 필터뱅크를 구성할 수 있다.

특징추출의 두번째 단계는 주파수 분석 결과를 판별력이 저하되지 않도록 하면서 차원을 감소시키는 변환과정으로 DCT, Karhunen- Loeve transform (KLT), linear discriminant analysis (LDA), heteroscedastic discriminant analysis (HAD) 등이 있다[7-9]. 가장 널리 사용되는 DCT는 AR (1) 신호의 자기상관 행렬의 고유벡터를 근사하므로 KLT와 점근적으로 유사하며, M개의 주파수 밴드에서 M개의 다른 주파수 성분을 얻을 수 있는 장점이 있다. 그러나 DCT는 최소자승의 관점에서 보면 최적 사영 변환하는 KLT에 비해 신호의 표현이 떨어지고, 통계적인 판별력의 관점에서 보면 LDA, HAD 등의 자료기반 변환에 비해 성능이 떨어진다. 그러나 자료기반 방법은 주어진 자료에 최적화할 수 있으나, 신호의 특성이 변할 때 음성인식의 경우에는 음운적 특성, 음향환경이 변할 경우에 성능이 저하되므로 다시 학습하여야 하며, 변환행렬이 일관성 있는 구조를 갖지 않아서 분석이 용이하지 않다는 단점이 있다.

본 논문에서는 부대역에 포함된 음성인식 정보를 효과적으로 이용하기 위해, 주파수 분석 단계보다 최적화할 파라미터의 개수가 적고 분석이 용이한 특징추출의 두번째 단계인 변환과정에서 최적화를 한다. 즉 최적화를 위해 상세한 음성정보를 추출하면 음성인식률의 향상에 기여할 것으로 여겨지는 주파수 부대역의 특징 파라미터 차원을 증가시킨다. 변환방법으로 자료기반 변환을 사용하는 대신에 신호에 독립적인 DCT를 사용하여 부대역별로 상이한 차원의 켑스트럼 특징벡터의 효과를 알아 본다. 그러나 어떠한 변환이라도 부대역별로 상이한 차원의 특징벡터를 생성할 수 있으므로 DCT 이외의 변환도 사용이 가능하다. 본 논문의 2장에서는 부대역별로 독립적인 특징 추출 방법을 설명하고, 3장에서는 음성인식에 기여도가 큰 주파수 부대역을 선택하는 부대역별로 켑스트럼 특징벡터의 차원을 최적화하기 위한 기준을 제시한다. 4장에서 제안한 특징추출방법을 음성인식 실험에 적용한 결과를 기술하고, 5장에서 결론 및 향후 연구에 대해서 살펴보기로 한다.

II. 주파수 부대역별 특징 추출

음성인식에 기여하는 정도에 비례하여 부대역을 상세분석하려면 부대역별로 독립적인 특징추출이 가능하여야 한다. 전대역에서 특장을 추출하게 되면 모든 필터뱅크의 에너지가 동등한 기여를 하게 된다. 본 논문에서는 부대역별로 독립적인 특징추출을 하기 위한 방법으로 멀티밴드 접근음성인식 방식을 사용한다. 멀티밴드 접근음성인식은 인간이 음성을 인식할 때, 전체 주파수 대역을 대상으로 하지 않고 다수의 부대역으로 나누고 독립적으로 인식한다는 연구결과에 기반을 두고 있다[10]. 이를 음성인식에 적용하려는 시도가 멀티밴드 접근방법으로, 주 연구분야는 부대역의 정의 및 구성 방법, 부대역별 효과적인 특징추출 및 인식 단위 선정, 부대역간의 비동기적 인식, 부대역의 인식 모델, 인식 결과의 통합 방법 등이다[11-13].

부대역별로 독립적인 특징 벡터를 추출하기 위하여, 각 부대역에 포함된 필터뱅크 에너지를 로그변환한 후 DCT를 사용하여 켑스트럼 특징벡터로 변환한다. 필터뱅크의 구성은 퓨리에 변환을 통하여 얻은 파워 스펙트럼 $P(f)$를 청각기의 인지단위를 모의한 바크 스케일 주파수축 Ω로 변환한다[2].

$$\Omega(f) = 6\log\{f/600 + [(f/600)^2 + 1]^{1/2}\} \quad (1)$$

변환된 파워 스펙트럼은 주대역 매스킹 곡선 $\Psi(\Omega)$와 컨벌루션하여 k번째 필터뱅크 에너지 $\Theta(\Omega_k)$를 얻는다.

$$\Theta(\Omega_k) = \sum_{\Omega=-1.3}^{2.5} P(\Omega_k - \Omega)\Psi(\Omega) \quad (2)$$

단, $\Psi(\Omega)$는 Schroeder의 청각필터를 근사한 것이다.

$$\Psi(\Omega) = \begin{cases} 10^{2.5(\Omega+0.5)}, -1.3 \le \Omega \le -0.5 \\ 1, -0.5 \le \Omega \le 0.5 \\ 10^{-1.0(\Omega-0.5)}, 0.5 \le \Omega \le 2.5) \end{cases} \quad (3)$$

부대역 b에 포함된 필터뱅크 에너지를 $\{\Theta(\Omega_{s(b)}),$

$\Theta(\Omega_{s(b)+1}), \cdots, \Theta(\Omega_{e(b)})\}$라 할 때, 부대역 b에서 켑스트럼 $C_{b,n}$을 얻는다.

$$C_{b,n} = \sum_{k=s(b)}^{e(b)} \log(\Theta(\Omega_k)) \cos((k+0.5-s(b))*n*\pi / (e(b)-s(b)+1)) \quad (4)$$

부대역의 개수를 B라 하고, 부대역 b에서 켑스트럼의 최고 차원을 d(b)라면 최종적인 특징 벡터는 이들을 연결한 것이다.

$$C = (C_{1,0}, C_{1,1}, \cdots, C_{1,d(1)}, C_{2,0}, C_{2,1}, \cdots, C_{2,d(2)}, \cdots, C_{B,0}, C_{B,1}, \cdots, C_{B,d(B)})^t \quad (5)$$

부대역을 상세분석하기 위한 방법으로 부대역의 켑스트럼 최고 차원을 증가시킨다. 따라서 부대역마다 다른 최고 차원의 켑스트럼을 가지게 된다.

일반적인 멀티밴드 방법에서는 부대역별로 추출된 특징벡터에 대해 인식기를 독립적으로 구성한다. 따라서 전체 대역내에서의 스펙트럼 구조나 부대역간의 상관관계를 이용하지 않으므로 인식률이 전대역 특징에 비해 감소하는 경향이 있다. 실험결과[14]에서 보듯이 전대역 특징벡터를 보조적으로 사용할 경우 성능이 더욱 향상된다. 이는 부대역과 전대역이 제공하는 음성인식을 위한 정보에 중복되지 않는 부분이 존재하고, 전대역의 스펙트럼 구조가 멀티밴드에서 주장하는 인간에 의한 음성인식과는 달리, 현재 컴퓨터에 의한 음성인식에 중요한 정보가 되기 때문이라 판단된다. 본 논문에서는 주파수 부대역별로 특징을 추출하지만, 부대역 특징에 대해 독립적인 인식기를 구성하지 않고 각 부대역의 특징을 하나로 통합하여 이용하므로 부대역간의 상관관계를 유지할 수 있다.

부대역별 특징추출을 계산량의 관점에서 살펴보면 전대역에서 켑스트럼을 추출할 경우에는 식 (4)에서 로그 필터뱅크 에너지와 코사인값을 곱하는 개수 (e(b)-s(b)+1)이 전체 필터뱅크의 개수로 바뀐다. 따라서 필터뱅크의 수를 N이라 하고, 전대역을 B개의 부대역으로 분할할 경우에는 대략 N번의 곱셈에서 N/B번의 곱셈으로 계산량이 감소한다.

III. 부대역별 켑스트럼 해상도 결정을 위한 기준

최적 음성인식 성능을 얻도록 각 부대역의 켑스트럼 해상도를 결정하기 위한 적당한 기준은 음성인식률이다. 따라서 학습자료를 사용하여 가능한 모든 부대역별 상세분석켑스트럼 차원의 조합 중에서 최고의 인식률을 갖는 최적 조합을 결정하여 이후의 인식에 적용한다. DCT를 사용하여 최종 특징 벡터를 추출하므로, 각 부대역에 존재하는 필터뱅크의 개수가 이들로부터 최대한 추출할 수 있는 특징벡터의 차원이다. 모든 가능한 경우의 수는 $\prod_{b=1}^{B}$(e(b)-s(b)+1)개이나 너무 작은 차원의 특징벡터는 인식에 적합하지 않으므로 제외한다.

음성인식률 기준 이외에 음성인식에 효과적인 특징벡터는 음성인식 단위를 보다 더 잘 구분할 수 있다고 판단되므로, 이를 최적조합을 찾는데 이용할 수 있다. 음성인식 단위의 집합을 C개의 클래스 $D_1, D_2, \cdots, D_C$라 하자. 또 특정 상세분석 켑스트럼 해상도 조합에 의해 추출된 특징 벡터의 집합을 $D=\{x_1, x_2, \cdots, x_n\}$이라 하자. 이들 특징벡터가 각 클래스를 얼마나 잘 구분할 수 있는지를 알아내기 위해 군집화 품질을 조사한다.

D_k에 속한 특징벡터의 개수를 n_k, 평균벡터를 m_k, 전체 자료의 개수를 n, 전체 평균벡터를 m이라 하자. 전체 스캐터 (scatter) 행렬 S_T, 클래스 내의 스캐터 행렬 S_W와 클래스간의 스캐터 행렬 S_B 사이에는 $S_T = S_W + S_B$의 관계가 성립한다. 단, $S_T = \sum_{x \in D}(x-m)(x-m)^t$, $S_W = \sum_{k=1}^{C}\sum_{x \in D_k}(x-m_k)(x-m_k)^t$, $S_B = \sum_{k=1}^{C} n_k(m_k-m)(m_k-m)^t$ 이다. 일반적인 군집화 과정에서는 S_T가 고정이므로 S_W를 최소화하면 S_B가 최대화되어, 클래스 내의 분산은 최소화되고, 클래스간의 거리는 최대화되는 군집화를 수행할 수 있다. 그러나 본 논문에서 구해지는 S_T는 켑스트럼 해상도 조합마다 달라지므로 S_W만으로는 각 부대역의 군집화 품질을 판단할 수 없고 S_T를 고려한 다음의 값으로 부대역의 판별력을 정의한다.

$$\chi(d(1), \cdots, d(B)) = \frac{|S_W(d(1), \cdots, d(B))|}{|S_T(d(1), \cdots, d(B))|} \quad (6)$$

여기서 $|\cdot|$는 행렬식이며, 군집화의 품질 척도 χ, S_T 및 S_W는 부대역 켑스트럼 해상도 조합 $(d(1), \cdots, d(B))$의 함수이다.

행렬식은 주축 방향의 분산의 곱과 비례한다. 따라서 식 (6)은 행렬 S_W과 S_T의 흩어짐 체적 (scattering volume)의 제곱에 대한 비율에 해당하므로 군집화 품질의 척도로 사용할 수 있다. 이 값이 작을수록 군집화의 품질이 우수다고 할 수 있다.

IV. 실험 및 결과

TIDIGITS 자료를 이용하여 인식실험을 수행하였다 [15]. 남성 55명, 여성 57명이 발성한 음성을 학습에 사용하였고, 학습에 사용되지 않은 남성 56명, 여성 57명이 발성한 음성을 평가자료로 사용하였다. 각 화자는 11개의 영어 숫자 "zero", "oh", "one", "two", "three", "four", "five", "six", "seven", "eight", "nine"을 고립형태로 각각 두 번씩, 두 자리부터 일곱 자리까지의 연결된 숫자를 55번 발성하였다. 평가자료는 28489개의 숫자로 구성되어 있다. 예비실험을 통해서 최적화된 상태랑 연속분포함수 (mixture)의 수, 상태수, 분석프레임의 길이를 결정하였다. 16개의 Gaussian mixture와 10개의 상태를 가지는 CHMM (continuous density hibden Markov model)을 인식모델로 사용하였고 분석프레임은 28 ms와 32 ms에서 최적의 결과가 발생하므로 두 경우에서 발생한 인식률의 평균을 사용하였다. 모든 실험에서 주파수 전대역을 식 (1)의 바아크 스케일을 사용하여 균등한 19개의 간격으로 분할한 19개의 필터뱅크를 사용하였다. 이로부터 켑스트럼과 델타 켑스트럼을 추출하여 인식에 사용하였다. 실험에는 연결 숫자를 인식한 결과를 표시하였는데, 숫자열 인식률은 숫자열을 이루는 숫자가 모두 정확히 인식되었을 때의 비율이고, 백분율 정확도 (% correct)와 백분율 정밀도 (% accuracy)는 숫자열을 이루는 개별 숫자들이 정확히 인식된 비율인데, 백분율 정확도는 삽입 오류를 오류로 취급하지 않았을 때이고, 백분율 정밀도는 삽입 오류를 오류에 포함한 것이다.

제안한 방법과 비교를 위해 주파수 전대역에서 추출된 켑스트럼을 사용하여 연결숫자 인식을 수행하였다. 그림 1은 벡터의 최고 차수를 10부터 16까지 변화시키고, 켑스트럼 0차를 포함한 특징 파라미터를 사용한 숫자열 인식률이다. 표 1은 최적의 결과가 발생한 최고 차수 11차일때의 인식 결과를 나타내었다. 그림 1에서 보듯이 10차부터 14차까지 인식률이 비슷하였고, 더 이상 차원을 확장하여도 인식률의 향상되지는 않았다. 이론적으로 모델링이 정확하다면 차원이 클수록 성능이 향상되어야 하지만, 모델링의 가정이나 학습이 적절하지 않았을 때는 오히려 벡터의 일부만 사용하는 것이 성능이 좋다. 이러한 현상은 현재의 음성인식 모델링에서 흔히 발생하는 문제로서 전대역 특징 파라미터의 경우에도 최적의 차수가 존재한다.

표 1. 전대역 켑스트럼 최고 차수 11일때의 연결 숫자 음성 인식률 (%)

Table 1. Connected digit recgonition rates using full-band 11^{th} order cepstral coefficients (%).

String accurracy	%Correnr	%Accuracy
98.585	99.670	99.535

그림 2 (a), (b)는 최적의 부대역별 상세분석 조합을 결정하기 위하여 학습자료를 대상으로 한 음성 인식 실험의 결과이다. 세 개의 주파수 부대역 $\{\Theta(\Omega_0),\cdots,\Theta(\Omega_6)\}$, $\{\Theta(\Omega_6),\cdots,\Theta(\Omega_{12})\}$, $\{\Theta(\Omega_{12}),\cdots,\Theta(\Omega_{18})\}$에서 독립적인 켑스트럼을 추출하였다. 대역별 분석의 조합 (i,j,k)는 첫번째 부대역부터 세번째 부대역까지 차례로 켑스트럼 벡터의 최고차수가 i,j,k일 때를 표시한다. (3, 3, 3)부터 (5, 5, 5)까지 $3\times3\times3=27$가지 경우에서 최적의 조합을 구한다. i**는 첫번째 부대역의 켑스트럼 최고 차수가 i이고 두번째와 세번째는 최고차수가 3,4,5인 모든 경우의 평균이며, * j*, ** k도 마찬가지로 정의된다. 그림 1의 결과에서 보듯이 특징 벡터의 차원의 확장만으

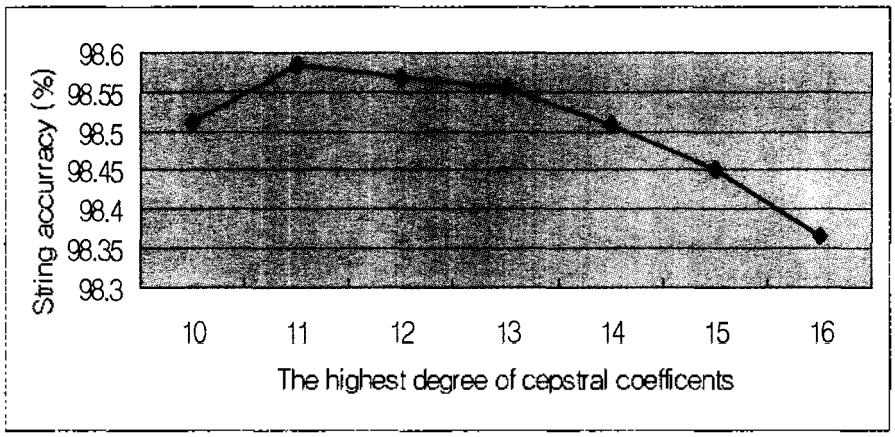
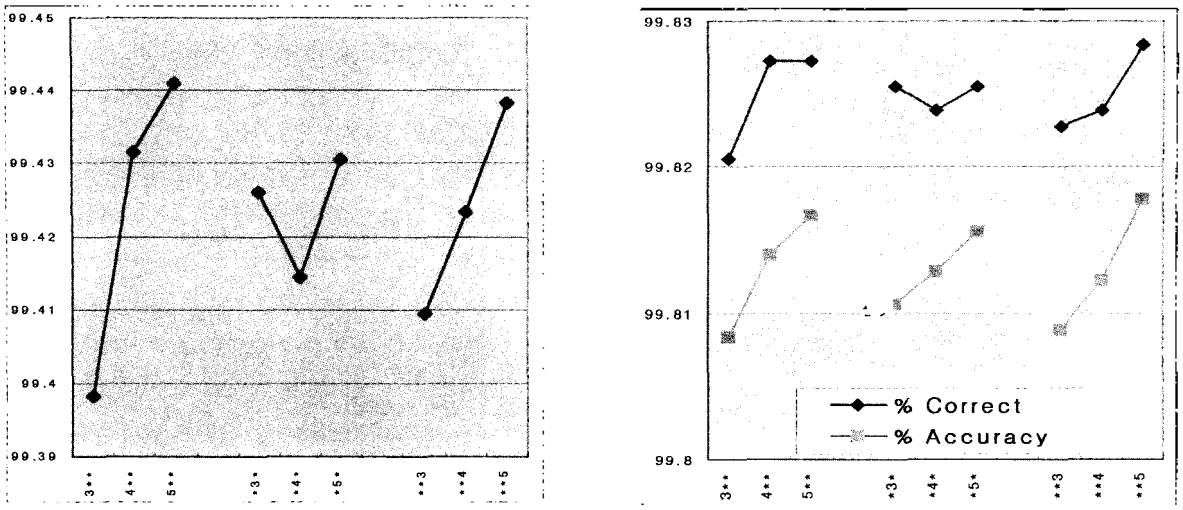


그림 1. 여러 차원의 전대역 켑스트럼 벡터를 사용하였을 때의 숫자열 인식률

Fig. 1. String accurracy using several dimensions of full-band cepstral coefficients.



(a) String accuracy (b) % correct and % accuracy

그림 2. 부대역 특징벡터 차원의 여러 조합에서의 음성인식률 (a) 숫자열 인식률 (b) % correct와 % accuracy
Fig. 2. Speech recognition rates using several combinations of sub-band feature vector dimension.

로는 인식률의 향상을 얻을 수 없으나, 학습자료는 모델을 학습하는데 사용한 것이므로 벡터의 차원이 커질수록 인식률이 증가하는 특징이 있다. 이러한 점을 고려하여 숫자열 인식률을 나타낸 그림 2 (a)와 백분율 정확도와 백분율 정밀도를 나타낸 그림 2 (b)를 관찰하였다. 첫번째 대역은 그림 2 (a), (b)에서 최고차수가 3에서 4로 증가할 때 인식률이 크게 증가하였으나, 5로 증가할 경우에는 향상이 적거나 없었다. 따라서 최적 차수는 4이다. 두번째 대역은 그림 2 (a), (b)에서 최고 차수가 4일 경우는 인식률이 감소하거나 증가가 작았고, 5로 증가할 경우에도 인식률의 증가가 작으므로 최적 차수는 3이다. 세번째 대역은 차원의 증가에 따라 비례하게 인식률이 증가하였으므로 최적차수는 5이다.

그림 3은 그림 2와 같은 조건에서 인식률 대신에 식 (6)의 통계치를 사용한 결과이다. 인식단위로 숫자를 사용하였고, 음성자료가 인식단위별로 분할되어 있지 않으므로 비터비 알고리즘을 사용하여 숫자열을 인식단위로 분할하여 실험에 사용하였다. 그림 2와 비교하면 전체적으로는 비슷한 결과를 보임으로서, 음성인식률을 기준으로 선택한 최적조합이 군집화의 품질로도 최적조합이 됨을 알 수 있다. 세번째 대역에서 최고차수가 3에서 4로 증가할 때의 군집화 품질이 크게 증가한 점이 그림 2의 결과와 다르다. 그러나, 본 논문에서는 군집화 품질을 인식률 기준에 대한 타당성을 확인하기 위한 수단으로 사용하므로, 그림 2에서 얻은 결과를 기준으로 최적 조합을 결정한다.

표 2는 평가자료를 사용하여, 여러가지 부대역 벡터의 차원조합에 대하여 인식실험을 수행한 결과이다. 인식률이 제일 높은 것은 (4, 5, 5)였으나, 최적조합으로 예상되는 (4, 3, 5)가 두번째로 인식률이 높았다. 두 인식방법을

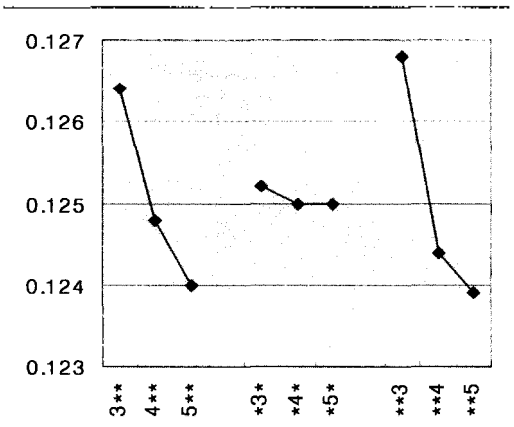


그림 3. 부대역 특징벡터 차원의 여러 조합에서의 군집화 품질 비교
Fig. 3. Comparison of the clustering qualities using several combinations of sub-band feature vector dimension.

표 2. 부대역의 캡스트럼 차원의 여러 조합에 따른 연결 숫자 음성 인식률 (%) 비교
Table 2. Comparison of connected digit recognition rates using several combinations of the dimension of sub-band cepstral coefficients.

	String accuracy	%Correct	%Accuracy
(4, 3, 5)	99.120	99.775	99.705
(4, 5, 5)	99.125	99.810	99.715
etc.	98.988	99.761	99.668

표 3. 정적인 특징을 사용한 연결 숫자 음성 인식률 (%) 비교
Table 3. Comparison of connected digit recognition rates using static feature.

	String accuracy	%Correct	%Accuracy
Optimal full band feafure	94.425	98.420	98.110
(4, 3, 5)	95.355	98.755	98.455

표 4. 동적인 특징을 사용한 연결 숫자 음성 인식률 (%) 비교
Table 4. Comparison of connected digit recognition rates using dynamic feature.

	String accuracy	%Correct	%Accuracy
Optimal full band feafure	96.505	99.435	98.830
(4, 3, 5)	97.640	99.585	99.210

비교할 경우에 상대적인 오류율의 감소를 자주 사용한다. 오류율은 (100 - 인식률)로 정의되는 오인식되는 비율이며, 상대적인 오류율 감소는 (제안방법의 오류율 / 기준방법의 오류율)로 정의된다. 제안한 방법인 최적 조합 (4, 3, 5)와 전대역 켑스트럼의 인식률 표 1을 비교하면 숫자열 인식률로는 오류를 38%, 백분율 정확도는 오류를 32%, 백분율 정밀도는 오류를 37% 감소시켰다

표 2는 정적인 특징과 동적인 특징을 동시에 사용한 실험 결과이나, 표 3은 동적인 파라미터를 제외하고 정적인 파라미터만을 사용하여 음성인식기의 기준 성능을 낮추어서 제안한 방법의 효과를 알아 본 결과이다. 최적 조합 (4, 3, 5)와 최고차수가 11인 최적 전대역 특징벡터를 사용한 인식률을 비교하면, 숫자열 인식률 기준으로 오류를 17%, 백분율 정확도는 오류를 21%, 백분율 정밀도는 오류를 18% 감소시켰다. 동적인 파라미터를 동시에 사용하는 표 2보다는 오류의 감소율이 적었지만 인식률이 향상되는 결과를 나타내었다.

표 4에서는 정적인 파라미터를 사용하지 않고, 동적인 파라미터만을 사용하여 제안한 방법의 효과를 알아보았다. 제안한 방법은 숫자열 인식률 기준으로 오류를 32%, 백분율 정확도는 오류를 27%, 백분율 정밀도는 오류를 32% 감소시켰다. 정적인 파라미터를 동시에 사용하는 표 2보다는 오류의 감소율이 적었지만 정적인 파라미터만을 사용하는 표 3의 경우보다는 인식률이 향상되는 결과를 나타내었다.

V. 결 론

본 논문에서는 주파수 부대역별로 독립적인 특징을 추출하고, 음성인식에 효과적이도록 부대역의 켑스트럼 해상도를 조절하는 방법을 제안하였다. 부대역의 켑스트럼 해상도를 조절하기 위해서, 주파수 부대역별로 독립적인 특징을 추출하는 멀티밴드 접근을 사용하였다. 부대역의 특징벡터를 독립적으로 인식기에 사용하고, 마지막 단계에서 인식결과들을 통합하는 멀티 밴드 인식방법 대신에, 특징벡터에 통합하여 이용하여 부대역간의 상관관계를 유지하였다. 음성인식률과 군집화 품질을 사용하여 부대역 특징벡터의 최적의 차원 조합을 찾았고, 이 조합을 TIDIGITS 연결 숫자음을 음성인식 실험에 사용하여 전대역에서 특징을 추출하는 방법과 비교하였다. 각각 숫자열 인식률, 백분율 정확도, 백분율 정밀도 기준으로 오류를 38%, 32%, 37% 감소시킴으로서 제안한 방법의 유효성을 확인할 수 있었다.

제안한 방법은 주어진 학습자료에서 최적의 부대역 특징추출을 결정한다. 따라서 본 연구의 실험자료인 영어 연결 숫자음 이외에 한국어 자료와 연속음성에 대해서 검증할 필요가 있다. 또한 주파수 부대역별 최적화를 위해 특징벡터의 해상도만을 조절하였으나, 부대역별로 최적화된 변환이나 특징추출을 사용하면 더욱 효과적일 것으로 판단된다.

감사의 글

본 논문은 2002학년도 경성대학교 특별과제연구비에 의하여 연구되었습니다.

참 고 문 헌

1. S. B. Davis and P. Mermelstain, "Comparison of parametric representations for monosyllable word recognition," *IEEE Trans. ASSP*, 28 (4), 357-366, 1990.
2. H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc. Am.*, 1738-1752, 1990.
3. E. Choi, D. Hyun and C. Lee, "Optimizing feature extraction for english word recognition," *Proc. ICASSP*, 813-816, 2002.
4. J. Lei and X. Bo, "Including detailed information feature in MFCC for large vocabulary continous speech recognition," *Proc. ICASSP*, 805-808, 2002.
5. Z. Tufekci and J. Gowdy, "Subband feature extraction using lapped orthogonal transform for speech recognition," *Proc. ICASSP*, SPEECH-P11.10, 2001.
6. R. Gemello, D. Albesano, L. Moisa and R. Mori, "Integration of fixed and multiple resolution analysis in a speech recognition system," *Proc. ICASSP*, SPEECH-P11.3, 2001.
7. K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, New York, 403, 1990.
8. K. Demuynck, J. Duchateau and D. V. Compernolle, "Optimal

feature sub-space selection based on discriminant analysis," *Proc. EUROSPEECH*, Budapest, Hungary, 1311-1314, 1999.
9. N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, **26**, 283-297, 1988.
10. J. B. Allen, "How do humans process and recognize speech," *IEEE Trans. On Speech and Audio Processing*, 2 (4), 567-577, 1994.
11. 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.
12. H. Hermansky, S. Tibrewala and M. Pavel, "Towards ASR on partially corrupted speech," *Proc. Int. Conf. on Spoken Language Processing*, **1**, 462-465, 1996.
13. N. N. Mirghafori, "A multi-band approach to automatic speech recognition," ICI TR-99-04, 1999.
14. C. Cerisara and D. Fohr, "Multi-band automatic speech recognition," *Computer Speech and Language*, **15**, 151-174, 2001.
15. R. G. Reonard, "A database for speaker-independent digit recognition", *Proc. ICASSP*, **3**, 42.11//1-4, 1984.

저자 약력

● 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

● 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

● 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