Homogeneous Centroid Neural Network에 의한 Tied Mixture HMM의 군집화

정회원 박 동 철*, 김 우 성**

Clustering In Tied Mixture HMM Using Homogeneous Centroid Neural Network

Dong-Chul Park*, Woo-Sung Kim** *Regular Members*

요 약

음성인식에서 TMHMM(Tied Mixture Hidden Markov Model)은 자유 매개변수의 수를 감소시키기 위한 좋은 접근이지만, GPDF(Gaussian Probability Density Function) 군집화 오류에 의해 음성인식의 오류를 발생시켰다. 본 논문은 TMHMM에서 발생하는 군집화 오류를 최소화하기 위하여 HCNN(Homogeneous Centroid Neural Network) 군집화 알고리즘을 제안한다. 제안된 알고리즘은 CNN(Centroid Neural Network)을 TMHMM상의 음향 특징벡터에 활용하였으며, 다른 상태에 소속된 확률밀도가 서로 겹쳐진 형태의 이질군집 지역에 더 많은 코드벡터를 할당하기 위해서 본 논문에서 새로 제안이 제안되는 이질성 거리척도를 사용 하였다. 제안된 알고리즘을 한국어 고립 숫자단어의 인식문제에 적용한 결과, 기존 K-means 알고리즘이나 CNN보다 각각 14.63%, 9,39%의 오인식률의 감소를 얻을 수 있었다.

Key Words : Speech Recognition, Tied Mixture, Unsupervised Learning, Hidden Markov Model

ABSTRACT

TMHMM(Tied Mixture Hidden Markov Model) is an important approach to reduce the number of free parameters in speech recognition. However, this model suffers from a degradation in recognition accuracy due to its GPDF (Gaussian Probability Density Function) clustering error. This paper proposes a clustering algorithm, called HCNN(Homogeneous Centroid Neural network), to cluster acoustic feature vectors in TMHMM. Moreover, the HCNN uses the heterogeneous distance measure to allocate more code vectors in the heterogeneous areas where probability densities of different states overlap each other. When applied to Korean digit isolated word recognition, the HCNN reduces the error rate by 9.39% over CNN clustering, and 14.63% over the traditional K-means clustering.

Ⅰ. 서 론

통계적인 음성인식론은 HMM의 상태관찰확률(State Observation Probability)을 모델화하여 Gaussian Mixture Density를 사용한 CDHMM(Continuous Density Hidden Markov Models)에 기초하고 있다. 그렇지만, CDHMM은 많은 자유 매개변수가 필요하여, 변수 측정과 관찰 확률의 측정에 많은 계산량을 요구하기 때문에, mixture 결박(tying)과 상태 결박을 함께 적용하는 TMHMM이 최근에 제안되었다[1]. TMHMM은 GPDF의

* 명지대학교 정보공학과 지능컴퓨팅 연구실, ** 호서대학교 컴퓨터공학부
논문번호 : KICS2005-12-486, 접수일자 : 2005년 12월 1일, 최종논문접수일자 : 2006년 8월 20일

전체집합을 공유함으로써, 코드벡터의 크기를 현저하게 감소시킴과 동시에 모델링의 정확도를 유지시키고자 한다. 그러나, GPDF를 공유하는데서 기인하는 군집화 오류에 의한 음성인식이 정확도 저하를 피할 수는 없으므로, 군집화 오류를 최소화하는 알고리즘의 개발이 요구되어 왔다[2-4].

많은 경우 K-means 알고리즘이 GPDF의 군집화에 이용되었는데[1], Rigazio등은 거리측정에 Bhattacharyya 거리를 사용한 Optimal Bhattacharyya Centroid 알고리즘을[2], Dermate등은 사전에 목표 오인식률을 정의하고, 분할과 병합 알고리즘을 사용해서 최적의 GPDF 수를 찾는 방법을 제안했다[3]. 많은 기존의 군집 알고리즘과 비교하여, 이 방법은 적은 수의 코드벡터 또는 낮은 오차율 측면에서 좋은 결과를 보여주고 있다. 비슷한 착상이 DCNN(Divergence-based Centroid Neural Network) 군집화 알고리즘이 제안 되었는데[4], DCNN은 CDHMM상에 음향 특징벡터를 군집화하기 위해 CNN(Centroid Neural Network)[5]를 활용하였으며, 승자뉴런을 결정하기 위해 발산 척도를 사용하였다.

본 논문은 TMHMM에서 나타나는 코드벡터의 공유성을 활용하기 위해서 이질성 거리(heterogeneous distance)라고 불리는 새로운 거리척도를 사용한 CNN을 제안하며, HCNN(Homogeneous CNN)라 부른다. HCNN에서 중요한 착상은 CDHMM과 달리, TMHMM상에서 모든 상태의 특징벡터를 상태에 상관없이 한 공간에 넣고 군집화하면서, 되도록이면 같은 상태에서 온 GPDF를 함께 군집화하며, 다른 상태에서 온 GPDF는 다른 군집에 속하도록 군집화를 진행시키고자 한다.

본 논문의 2장에서는 TMHMM을 간단히 요약하고, 이질성 거리를 정의한다. 3장에서는 HCNN을 제안하며, 4장에서는 한국어 고립 숫자단어 인식에 응용한 실험 결과와 다른 알고리즘들과 비교한 성능을 보여준다. 5장에서는 본 논문의 결론이 주어진다.

Ⅱ. TMHMM과 이질성 거리

2.1. Tied Mixture Hidden Markov Model

그림 1은 3개의 상태를 가지는 간단한 형태의 TMHMM의 예인데, j-번째 상태의 관찰확률을 다음과 같이 정의한다.

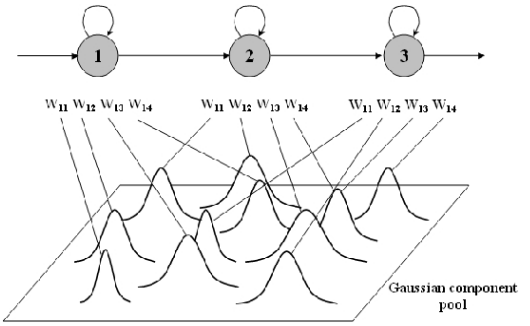


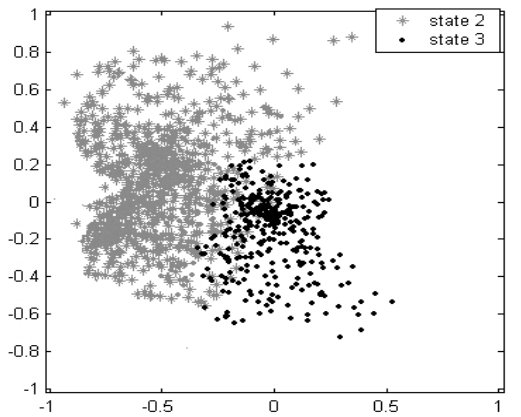
그림 1. 3개의 상태를 적용한 Tied Mixture HMM의 예

$$b_j(o_t)=\sum_{m=1}^{M_s} c_{jm} \cdot \aleph\left[o_t,\mu_m,\delta_m^2\right] \qquad (1)$$

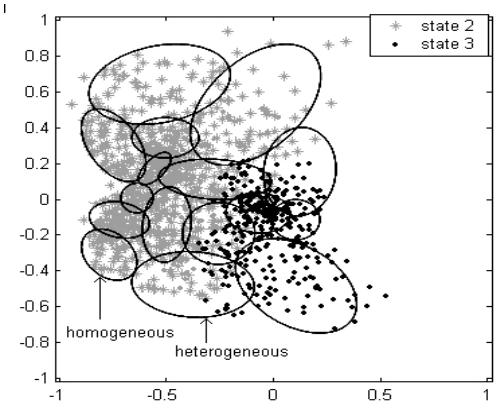
여기서 Ms는 Gaussian mixture 성분의 수, $\aleph$은 확률분포함수이고, c_{jm}은 m과 연관된 상태 j에 대응하는 연결강도를 나타낸다.

2.2. TMHMM 군집에러와 이질성 거리

TMHMM은 확률분포함수로 표현되는 특징벡터들을 모두 한 공간에 모아놓고, 비슷한 성질을 가지는 것들끼리 군집화 시킨 다음, 대표벡터들로 같은 군집의 특징벡터들을 대체하여 사용한다. 이 경우, 각 군집을 "동질군집"과 "이질군집"으로 나눌 수 있는데, 동질군집은 모든 특징벡터들이 같은 상태에 속하는 군집이고, 이질군집은 그 군집에 두개 이상의 상태에 속하는 특징벡터가 있는 것이다.



(a)



(b)

그림 2. 동질군집과 이질군집 (a) 2차원 평면에 투사한 특징벡터들 (b) 군집화의 결과

그림 2-(a)는 특징벡터들의 군집화를 2차원 평면에 투사시킨 예이다. 그림 2-(b)는 이들 특징벡터를 여러 개의 군집으로 군집화 결과인데, 왼쪽 위의 군집들은 한 개의 군 집 내부에 상태 2에서 온 특징벡터들로만 이루어진 "동질군집"이고, 오른쪽 중간의 군집들은 상태 2와 상태 3에서 온 특징벡터가 같은 군집에 속하는 "이질군집"이다.

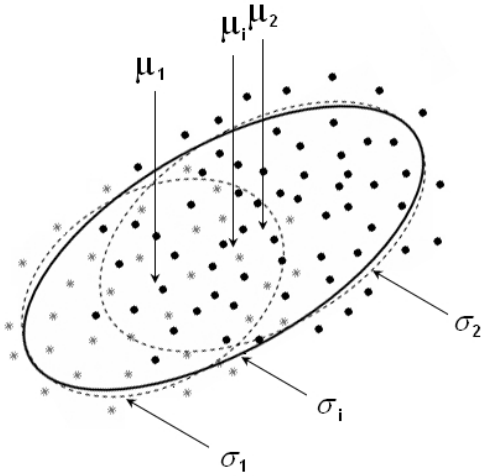


그림 3. 전형적인 이질군집

실제로, TMHMM에서 상태 관찰확률의 계산과 그 모델의 정확성은 군집결과에 영향을 받게 되는데, 인식기의 오류정도의 중요한 원인의 하나는 "이질군집"으로 추정된다. "이질군집"의 전형적인 예는 그림 3으로 표현될 수 있다.

이질군집의 내부에 존재하는 특징벡터의 분포를 분석하면, 2개의 상태가 있는 경우, i-번째 군집은 첫 번째 상태의 벡터 N_i^1개와 두 번째 상태의 벡터 N_i^2개로 이루어져 있다. 이때, 이질군집 i는 다시 이들 각 상태에서 온 특징벡터들로만 이루어진 2개의 하위 군집을 형성할 수 있는데, 이를 각각 $g_i^1(\mu_i^1, \Sigma_i^1)$과 $g_i^2(\mu_i^2, \Sigma_i^2)$로 표현한다. 여기서 μ_i는 평균, Σ_i는 공분산을 나타내는데, 평균 벡터와 공분산 행렬은 아래와 같이 계산된다.

$$\mu_i = \frac{1}{N_i} \sum_{x_j \subseteq X} x_j$$

$$\Sigma_i = \frac{1}{N-1} \sum_{x_j \subseteq X} (x_j - \mu_i)(x_j - \mu_i)^T$$

$$\mu_i^1 = \frac{1}{N_i^1} \sum_{x_j \subseteq X_i^1} x_j$$

$$\Sigma_i^1 = \frac{1}{N_i^1 - 1} \sum_{x_j \subseteq X} (x_j - \mu_i^1)(x_j - \mu_i^1)^T$$

$$\mu_i^2 = \frac{1}{N^2} \sum_{x_j \subseteq X_i^2} x_j$$

$$\Sigma_i^2 = \frac{1}{N_i^2 - 1} \sum_{x_j \subseteq X_i^2} (x_j - \mu_i^2)(x_j - \mu_i^2)^T$$

T: 전치행렬

여기서 X_i^1와 X_i^2는 각각 상태 1과 상태 2의 모든 벡터의 집합이므로, 이질군집에 모든 벡터의 집합은 $X_i = X_i^1 \cup X_i^2$이다.

2.3. TMHMM 군집 오류와 이질성 거리

군집의 이질성을 측정하기 위한 이질성 레벨은 상위 군집과 하위 군집들 사이 차이의 합으로 측정되어 진다.

$$H_i = \sum_{j=1}^{N_i} D(g_i, g_i^j) \qquad (2)$$

여기서 H_i는 i-번째 군집의 이질성 레벨, N_i는 HMM에 있는 상태의 수이고, $D(g_i, g_i^j)$는 군집 g_i와 하위 군집 g_i^j사이의 차이를 지칭한다. 군집 i에 j-번째 상태가 특징 벡터를 가지고 있지 않으면, $D(g_i, g_i^j)$=0이라고 가정한다. 어떤 2개의 상태의 거리를 표현하기 위해 본 논문에서는 Bhattacharyya 거리가 사용되었다. Bhattacharyya 거리는 2개 Gaussian 분포사이 분리척도(separability measure)이며, 다음과 같다.

$$D(g_i, g_i^j) = \frac{1}{8}(\mu_i - \mu_i^j)^T \left[\frac{\Sigma_i + \Sigma_i^j}{2}\right]^{-1}(\mu_i - \mu_i^j) + \frac{1}{2} ln \frac{\left|\frac{\Sigma_i + \Sigma_i^j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_i^j|}} \quad (3)$$

식 (3)의 첫 번째 항은 평균값의 차이로 인한 군집 가분성을 반영한 반면, 두 번째 항은 공분산 행렬의 차이로 인한 군집 가분성을 보여준다.

Ⅲ. Homogeneous CNN(HCNN)

3.1 CNN 알고리즘

CNN 알고리즘은 기존 K-means 알고리즘을 기초로, 주어진 데이터에 존재하는 군집의 중심을 찾는다[4, 5]. 기존 CNN은 승자, 패자의 연결강도의 갱신 방법에 의해 지역적으로 최적의 연결 강도를 설정함으로 주어진 데이터를 표현할 수 있는 중심을 설정하는데 탁월하며, 또한 사전에 학습계수나 전체 학습 Epoch수를 설정할 필요가 없다. 입력 벡터 x가 n 시간에 인가된 경우, 승자뉴런 j와 패자뉴런 i의 연결강도 갱신은 아래와 같다.

$$\boldsymbol{w_j}(n+1) = \boldsymbol{w_j}(n) + \frac{1}{N_j+1}[\boldsymbol{x}(n) - \boldsymbol{w_j}(n)]$$

$$\boldsymbol{w_i}(n+1) = \boldsymbol{w_i}(n) - \frac{1}{N_i-1}[\boldsymbol{x}(n) - \boldsymbol{w_i}(n)] \quad (4)$$

위 식에서 $\boldsymbol{w_j}$와 $\boldsymbol{w_i}$는 각각 승자뉴런과 패자뉴런의 연결강도를 표현하며, N_i와 N_j는 각 군집 i와 j의 데이터 수를 나타낸다. 출력뉴런의 연결강도는 아래의 식과 같이 총 거리를 최소화하는 방법을 선택한다.

$$\boldsymbol{w_j} = \min_{\boldsymbol{w}} \sum_{i=1}^{N_i} \| \boldsymbol{x_j}(i) - \boldsymbol{w} \|^2 \quad (5)$$

위 식에서 N_i는 i 군집에 속한 데이터 수를 나타내며, CNN에 관한 더 자세한 설명은 [4]-[6]에서 찾을 수 있다.

3.2 Homogeneous CNN(HCNN)

TMHMM에서 군집화 오류의 중요한 요인 중 하나가 다른 상태에서 온 특징벡터가 같은 군집으로 군집화 되는 "이질군집"의 지역인데, 이러한 이질군집에 더 많은 코드벡터를 할당하여, 동질군집으로 변화를 주고자 하는 것이 HCNN이다. 즉, 아래와 같은 새로운 거리척도를 CNN에 적용한다.

$$d_H(\boldsymbol{x}(n), \boldsymbol{w}(n)) = \frac{n_h}{H_{X(n)}} d(\boldsymbol{x}(n), \boldsymbol{w}(n)) \quad (6)$$

위 식에서 $H_{X(n)}$은 $\boldsymbol{x}(n)$의 이질성 레벨, $d(\boldsymbol{x}(n), \boldsymbol{w}(n))$은 특징벡터 $\boldsymbol{x}(n)$과 코드벡터 $\boldsymbol{w}(n)$ 사이 Euclidean 거리이고, $d_H(\boldsymbol{x}(n), \boldsymbol{w}(n))$는 이질성 거리라 부른다. 매 i학습에서, 각 군집의 이질성 거리를 계산하고, 계산된 이질성 거리는 군집내의 모든 특징벡터의 이질성 거리로 정해진다.

Ⅳ. 실험 결과

본 실험에서는 ETRI 음성 데이터베이스의 한국어 고립 숫자단어를 이용한 실험으로 6,451개 발성을 사용하였다. 음성의 발췌는 무선, 유선, PCS, 휴대전화 환경에서 남자 981명과 여자 1,019명을 통해 기록되었다. 데이터는 8kHz 샘플한 후, 16bit/mono signed raw PCM 포맷으로 저장되었고, $1-0.97z^{-1}$의 전달함수를 이용하여 전처리되었다. 파형은 30ms Hamming window로 매 10ms 블록으로 프레임화된 후, 26채널의 filter bank를 이용해 12 MFCC(Mel Frequency Cepstral Coefficients)를 구했다. 학습의 단계에서 각 단어를 모델링하기 위해 5상태, left-to-right HMM이 사용되었으며, 총 13개의 숫자의 인식을 위해 13개의 HMM이 사용되었다.

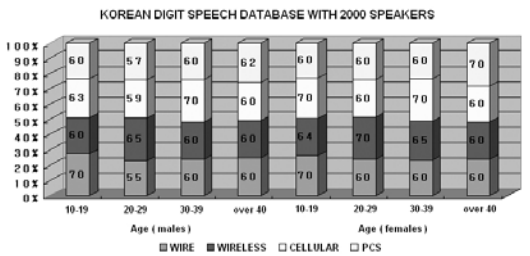


그림 4. 한국어 숫자 ETRI 음성 데이터베이스

5,451개의 데이터로 학습한 후, 1,000개의 데이터를 이용해서 3개 알고리즘(K-means, CNN, HCNN)에 적용하고, 정확한 비교를 위해 70~500개의 Gaussian 성분들을 바꿔가며 실험한 결과가 그림 5로 주어지는데, CNN과 HCNN의 성능이 K-means보다 월등함을 알 수 있다. 이 결과는 참고문헌 [4]의

실험결과와 개략적으로 일치한다. 두드러진 현상은 겹쳐진 지역에 더 많은 코드벡터를 할당함으로, HCNN은 CNN과 K-means보다 좋은 인식 성능 보여주고 있다. 특별히 Gaussian 성분의 수가 400개를 초과하는 경우 현저하게 나타난다. HCNN 평균 오인식률은 11.2%이다. K-means(평균 오인식률 13.12%)와 비교해서, HCNN는 (13.12-11.2)/13.12=14.63%의 오인식률 감소를 보였으며, CNN과 비교해서는 9.39% 평균 오인식률 감소를 가져왔다. 이러한 오인식률의 감소는 TMHMM에 이질군집/동질군집의 개념을 도입하여, 이질군집에 보다 많은 코드벡터를 할당하여 군집화를 실행하는 HCNN의 군집화능력에 기인한다고 할 수 있다.

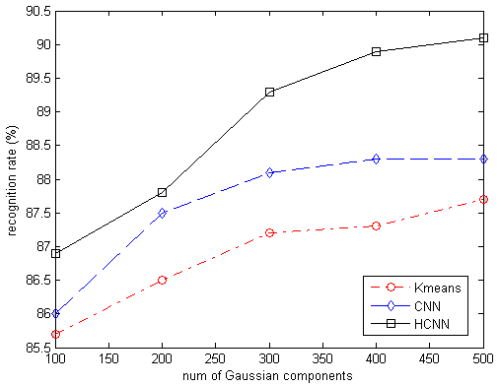
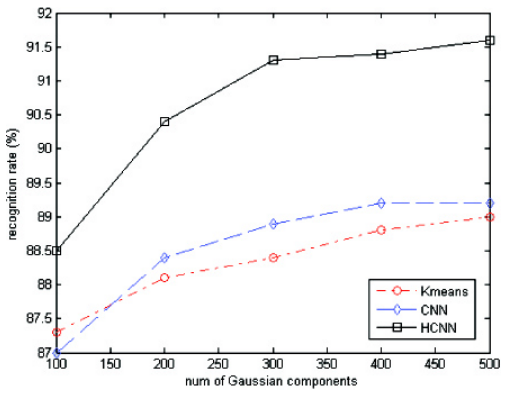
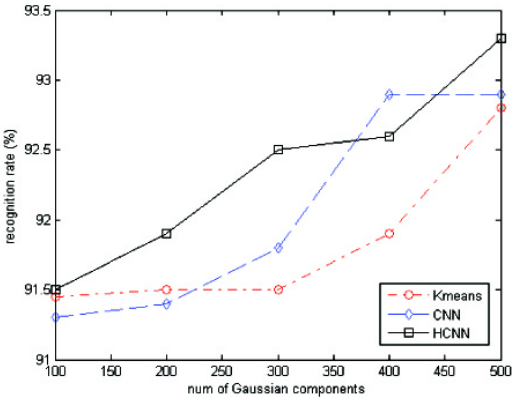


그림 5. K-means와 CNN과 비교한 HCNN 인식률

그림 6은 무선 환경과 PCS 환경 데이터에 대한 각각의 실험 결과를 보여준다. 일반적으로 무선 환경 데이터에 대한 인식은 높은 잡음 때문에 인식률의



(a)



(b)

그림 6. 다른 환경에서 인식률 (a) 무선(wireless) 환경, (b) PCS 환경

저하를 가져오며, 그림 6에서 보듯이 PCS환경에서 보다 약 3.5%의 인식률 하락이 관찰되었다. 그림 5의 경우에서 보다 그림 6의 개별 환경에서의 결과가 인식률 면에서 더욱 좋은 성능을 보이는데, 이는 다양한 환경에서 기록된 신호를 함께 섞는 경우에 더욱 인식하기가 어려운 것에 기인한 것으로 추정된다. 위의 결과에서도 역시 HCNN은 대부분의 경우 CNN이나, K-means의 성능을 개선시킨 결과를 보여준다.

V. 결 론

음성인식에서 TMHMM(Tied Mixture Hidden Markov Model)은 자유 매개변수의 수를 감소시키기 위한 좋은 접근이지만, GPDF(Gaussian Probability Density Function) 군집화 오류에 의해 음성인식의 오류를 발생시켰다. 본 논문은 TMHMM에서 발생하는 군집화 오류를 최소화하기 위하여 HCNN (Homogeneous Centroid Neural Network) 군집화 알고리즘을 제안한다. 제안된 알고리즘은 CNN(Centroid Neural Network)을 TMHMM상의 음향특징벡터에 활용하였으며, 다른 상태에 소속된 확률밀도가 서로 겹쳐진 형태의 이질군집 지역에 더 많은 코드벡터를 할당하기 위해서 본 논문에서 새로 제안이 제안되는 이질성 거리척도를 사용 하였다. 제안된 알고리즘을 한국어 고립 숫자단어의 인식문제에 적용한 결과, 기존 K-means 알고리즘이나 CNN보다 각각 14.63%, 9,39%의 오인식률 감소를 얻을 수 있었다.

음성인식을 위해 적용되는 TMHMM에서 GPDF의 군집화를 위하여 사용되어온 K-means 알고리즘은 HMM의 상태에 관계없이 군집화를 통해 코드북을 발생시키는 방법으로, 인식률의 저하를 피할 수 없었다. 본 논문에서는 TMHMM에서 발생하는 군집화 오류를 최소화하기 위하여 군집화의 과정에서 되도록 같은 상태에서 온 상태들이 한 군집을 이루도록 하는 알고리즘인 HCNN이 제안되었다. 이를 위해 각 군집에 다른 상태에서 온 GPDF가 얼마나 포함되었는 가의 정도를 나타내는 이질성 거리척도가 본 논문에서 새로이 제안되었다. HCNN은 새로이 제안된 이질성 거리척도와 CNN을 기반으로하는 새로운 군집화 알고리즘으로, TMHMM에서 다른 상태의 확률분포가 서로 겹쳐진 이질지역에 더 많은 코드벡터를 할당하는 역할을 하여, 궁극적으로 GPDF의 군집화에서 발생하는 인식률의 저하를 최소화하게 된다. 제안된 HCNN은 ETRI 음성 데이터베이스의 한국어 고립 숫자단어의 인식에 적용되었고, 그 결과, HCNN은 K-means에 비해 오인식률을 약14% 감소시켰으며, CNN과 비교해서는 오인식률을 9.39% 감소시켰다. 이러한 성공적인 응용 결과는 제안된 HCNN이 TMHMM의 GPDF의 군집화를 통한 음성인식기의 소형화에 지대한 영향을 기대할 수 있게 한다.

참 고 문 헌

〔1〕 Liu, Y. and Fung, P., "State dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition," IEEE Tr. on ASSP, vol.14, issue.1, pp. 89-102, Jul. 2004.

〔2〕 Rigazio, L., Tsakam B., and Junqua J., "An optimal Bhattacharyya centroid algorithm for Gaussian clustering with applications in automatic speech recognition," Proc. of ICASSP, vol.3, pp. 1599-1602, 2000.

〔3〕 Dermatas, E. and Kokkinakis, G., "Algorithm for clustering continuous density HMM by recognition error", IEEE Tr. on ASSP, vol.4, pp231-234, May.1996.

〔4〕 Park, D.C., Kwon, O.H., and Suk, M., "Clustering of Gaussian Probability Density Functions Using Centroid Neural Networks," IEE Electronic Letters, vol 49, no.4, pp.381-382, Feb 2003.

〔5〕 Park, D.C., "Centroid Neural Network for Unsupervised Competitive Learning", IEEE Tr. on Neural Networks, vol.11, no.2, pp520-528, Mar. 2000.

〔6〕 박동철, 우영준, "신경망에의한 테두리를 보존하는 영상압축," 한국통신학회 논문지, 24권, 10B호, pp. 1946-1952, 1999

박 동 철 (Dong-Chul Park) 정회원

1980년 2월 : 서강대학교 전자공학과(공학사)

1982년 2월 : 한국과학기술원 전기 및 전자공학과(공학석사)

1990년 6월 : Univ. of Wahsington (Seattle), Electrical Engineering (Ph.D.)

1990년 8월~1994년 2월 : 조교수, Florida Int'l Univ. Dept. of Eelct. and Comp. Eng.

1994년 3월~현재 : 명지대학교 정보공학과 교수

1997년~2000년 : IEEE Tr. on Neural Networks, Associate Editor

1999년~현재 : IEEE Senior Member

<관심분야> 신경망 알고리즘 개발, 음성인식, 신경망의 금융공학에의 응용

김 우 성 (Woo-Sung Kim) 정회원

1993년 2월 : 서강대학교(공학박사)

1984년~1987년 : 한국전자통신연구소 연구원

1999년~2000년 : Univ. of Wah sington 방문교수

1987년~현재 : 호서대학교 컴퓨터공학부 교수

2004년~현재 : 호서대학교 정보관리처장, 첨단정보기술대학원장

<관심분야> 영상처리, 임베디드 시스템, 인공지능