論 文

가중 거리 개념이 도입된 HCNN을 이용한 화자 독립 숫자음 인식에 관한 연구

正會員 金 度 錫* 正會員 李 壽 永*

Speaker-Independent Korean Digit Recognition Using HCNN with Weighted Distance Measure

Do Seok Kim*, Soo Young Lee* *Regular Members*

要 約

HCNN(Hidden Control Neural Network)은 신경회로망에 의한 비선형 예측과 HMM의 segmentation 기능을 접합시킨 신경회로망 모델로서, 시간에 따라 입출력 사상 함수를 변화시킴으로써 음성 신호를 잘 모델링할 수 있도록 되어 있다. 본 논문에서는 첫째, HCNN의 성능이 HMM보다 우수함을 보이고, 둘째로, HCNN에서의 예측 오차 측정에 적절한 거리 측도를 이용하기 위해 가중거리가 도입된 HCNN을 제안하여, 화자 독립 음성 인식에 있어 그 성능이 우수함을 보였다. 여기서 가중거리는 음성 특징 벡터 각 구성 성분의 분산도 차이를 고려한 거리이다. 화자 독립 숫자음 인식 실험 결과, 유클리드 거리를 이용한 HCNN에 대해 95%의 인식율을 얻었는데, 이는 HMM에 비해 1.28% 높은 결과로서, 확률적인 제한이 가해진 HMM에 비해 시스템의 동작인 모델링을 이용한 HCNN이 더 우수함을 알 수 있다. 또한 가중거리를 이용한 HCNN에 대해서는 97.35%의 인식율을 얻었는데, 이는 유클리드 거리를 이용한 HCNN에 비해 2.3%가 향상된 결과이다. 가중 거리를 도입한 HCNN의 경우에 더 높은 인식율을 얻은 이유는, 오인식이 많이 되는 화자의 인식율을 높임으로써 화자간의 인식율차가 감소하게 되기 때문임을 알 수 있었고, 따라서 화자 독립 음성인식에 가중거리를 도입한 HCNN이 보다 적합함을 알 수 있다.

ABSTRACT

Nonlinear mapping function of the HCNN(Hidden Control Neural Network) can change over time to model the temporal variability of a speech signal by combining the nonlinear prediction of conventional neural networks with the segmentation capability of HMM. We have two things in this paper. First, we showed that the performance of the HCNN is better than that of HMM. Second, the HCNN with its prediction error measure given by weighted distance is proposed to use suitable

*韓國科學技術院 電氣·電子工學科
Dept. of Electrical and Electronics Eng., KAIST
論文番號 : 93-144

distance measure for the HCNN, and then we showed that the superiority of the proposed system for speaker-independent speech recognition tasks. Weighted distance considers the differences between the variances of each component of the feature vector extraced from the speech data. Speaker-independent Korean digit recognition experiment showed that the recognition rate of 95% was obtained for the HCNN with Euclidean distance. This result is 1.28% higher than HMM, and shows that the HCNN which models the dynamical system is superior to HMM which is based on the statistical restrictions. And we obtained 97.35% for the HCNN with weighted distance, which is 2.35% better than the HCNN with Euclidean distance. The reason why the HCNN with weighted distance shows better performance is as follows : it reduces the variations of the recognition error rate over different speakers by increasing the recognition rate for the speakers who have many misclassified utterances. So we can conclude that the HCNN with weighted distance is more suitable for speaker-independent speech recognition tasks.

Ⅰ. 서 론

신경회로망은 대단위 병렬성, 적응학습의 특징을 지니며, 패턴 인식, 적응 제어, 최적화 등에서 매우 우수한 성능을 보여 주고 있다[1]. 최근 들어 이러한 신경회로망을 음성 인식에 적용하는 연구가 매우 활발해지고 있으며, 기존의 방법보다 비교적 우수한 성능을 얻은 결과들이 발표되고 있다[2]. 그러나 대개의 이러한 연구들은 음성을 정적인 패턴으로 간주하여 음성 인식을 패턴 인식 문제로 변환시키는 접근 방법이며, 이러한 방법은 입력되는 음성을 인식 단위로 미리 분할한 다음에 인식해야 하는 어려움이 있다. 그것은 음성이 문자와는 달리 연속적으로 변화하는 신호이기 때문에 그 경계 구분을 정확히 하기가 쉽지 않기 때문이다. 한편, 전통적 기법으로써 많이 쓰이는 HMM(Hidden Markov Model)[7]은 그 자체 내에서 시간의 변동을 흡수할 수 있기 때문에, 위에 기술한 것과 같은 패턴 분류에 기초한 신경회로망의 문제점이 발생하지 않으며, 연속음성 인식에의 확장이 용이한 장점이 있다[3,4]. 하지만 HMM은 관측 기호들이 서로 확률적으로 독립이라는 제한이 가해져야 하므로 음성을 정확히 모델링하지 못하는 단점이 있다.

HCNN(Hidden Control Neural Network)[5,6]은 패턴 분류가 아니라, 예측에 기반을 둔 동적인 구조의 신경회로망 모델로서, 신경회로망의 우수한 시스템 모델링 능력과 HMM의 상태(state)에 의한 segmentation 기능을 모두 갖추고 있으므로, 음성과 같이 시간에 따라 변화하는 신호를 잘 모델링할 수 있다. 그런데, HCNN에서의 예측 오차는 시간 t에서의 특징 벡터가 입력으로 들어갔을 때 HCNN의 출력과, 시간 t+1에서의 실제 벡터간의 유클리드 거리로 정의되는데, 특징 벡터들의 분포가 특징 벡터 전 공간에 균일하게 분포되어 있는 경우에는 이 거리 측도가 적절하겠지만, 만약 특징 벡터의 분포가 어떤 특정한 분포함수를 따른다면, 그에 따른 거리를 이용하는 것이 타당할 것이다.

본 논문에서는 HCNN의 성능이 HMM보다 우수함을 보이고, HCNN에 유클리드 거리보다 적절한 거리 측도를 도입하기 위해, 가중거리가 도입된 HCNN을 제안하여, 유클리드 거리를 사용한 HCNN이나 HMM보다 화자 독립 음성 인식 성능이 우수함을 보이고자 한다. 여기서 가중거리는 특징 벡터 각 구성 성분의 분산도를 규준화 시키는 거리 측도이다. 이를 이용하기 위해서는 특징 벡터의 분포가 Gaussian이어야 하며, 특징 벡터 공분산 행렬에서 대각선 성분 이외의 값을 무시할 수 있다는 가정이 성립해야 하는데, 본 논문의 Ⅲ-1에서 실제 음성 데이타로부터 위의 가정이 타당하다는 것을 입증하겠다. 또 Ⅲ-2에서는 HMM과 유클리드 거리를 이용한 HCNN, 그리고 가중거리가 도입된 HCNN의 인식율을 비교, 검토하겠다.

Ⅱ. 가중 거리 개념이 도입된 HCNN

1. 신경회로망에 의한 비선형 예측과 HCNN

시스템 모델링이나 신호의 예측 등의 응용 분야에서는 과거 몇 개의 신호로부터 미래의 신호를 추측하게 되며 기존에 많이 쓰이던 방법은 선형 예측이다. 선형 예측은 식 (1)에 표현된 바와 같이, 과거 신호들

의 선형조합으로 미래의 신호를 예측할 수 있다고 가정하고 예측 계수들을 구하는 방법이다.

$$\hat{x}(t+1) = \sum_{i} \omega_i x(t) \tag{1}$$

여기서 $\hat{x}(t+1)$는 시간 t에서의 예측된 신호를 말하며 예측 계수 ω_i는 학습 데이타로부터 구하게 된다.

최근 신경회로망은 패턴 인식뿐 아니라 비선형 예측, 시스템 모델링 등의 분야에서도 매우 좋은 성능을 보여 주고 있다. 식 (1)에서 선형 조합을 비선형 함수로 대치하면 비선형 예측이 되며 이는 신경회로망의 비선형성에 의해 구현될 수 있다. 즉, 신경회로망에 과거 샘플을 입력시키면, 미래의 신호가 출력되도록 학습시키는 것이다. 예를 들어 예측에 하나의 과거 벡터만을 이용하는 1차 예측의 경우를 보면, 신경회로망 출력층의 뉴런값은 다음과 같은 식으로 표현된다.

$$y = \hat{x}(t+1) = F_{\omega}(x(t)) \tag{2}$$

여기서 $x(t)$는 신경회로망의 입력으로 인가되는 시간 t에서의 특징 벡터를, $\hat{x}(t+1)$은 신경회로망에 의해 예측되는 출력 벡터를, $F_{\omega}(\cdot)$는 신경회로망의 파라미터 ω에 의해 결정되는 신경회로망의 사상함수 나타내는데, 신경회로망의 파라미터인 연결 가중치 ω는 학습 데이타 $\{x_t,\ t=0,\ \ldots,\ T\}$으로부터, 예측 오차가 최소가 되도록 구하게 된다. 여기서의 예측 오차란 예측된 벡터 $\hat{x}(t+1)$과 실제의 벡터 $x(t+1)$과의 거리의 차를 의미하는데, 유클리드 거리를 이용해 정의하면

$$E(\omega) = \sum_{t=0}^{t-1} \| x(t+1) - F_{\omega}(x(t)) \|^2 \tag{3}$$

로 표현되며, 신경회로망을 학습시켜 얻는 파라미터 $\hat{\omega}$는 식 (4)와 같다.

$$\hat{\omega} = \arg\min_{\Omega} E(\omega) \tag{4}$$

여기서 Ω는 ω의 전 공간이다.

그러나 식 (2)와 같이 신경회로망의 연결 가중치가 고정되어 있으면, 시간에 따라 그 특성이 변화하는 시변 시스템(time-varying system)을 모델링 할 수 없다. 특히 음성은 시간의 흐름에 따라 그 특성이 바뀌는 비선형 시변 시스템으로부터 출력되는 신호라고 볼 수 있으므로, 음성의 이러한 시간 변동을 잘 흡수하기 위해서는 신경회로망의 연결 가중치를 시간에 따라 바꿔줌으로써 신경회로망의 사상함수를 변화시키는 것이 바람직하다고 볼 수 있다.

HCNN(Hidden Control Neural Network)은 이와 같은 문제를 해결하기 위한 신경회로망 모델이며, 그림 1에 그 구조를 나타내었다. 그림의 구조에서 제어 입력층을 제외하면 일반적인 다층 신경회로망(Multilayer Perceptron)과 동일하게 되며, 이는 시간 t에서의 벡터 $\boldsymbol{x}(t)$가 신경회로망에 인가되었을 때 시간 $t+1$에서의 벡터의 예측치인 $\hat{x}(t+1)$가 출력되는 비선형 예측기가 된다.

그런데 그림과 같이 제어 입력층을 두어 시간에 따라 제어 입력 벡터 $\boldsymbol{c}_t$를 변화시켜 주면, 동일한 $\boldsymbol{x}(t)$에 대해서도 은닉층으로 전달되는 신호는 다르게 될 것이고 결국 출력층에서도 다른 값이 나올 것이다. 즉, HCNN은 시간에 따라 연결가중치를 변화시키는 대신, 제어 입력층을 두어 시간에 따라 제어 입력 $\boldsymbol{c}_t$를 다르게 함으로써 연결 가중치의 변화 없이도 신경회로망의 입-출력 사상을 변화시킬 수 있는 구조이다. 이는 HCNN의 사상함수가 연결가중치 뿐 아니라 제어 입력에 의해 결정되는 것을 의미하며, HCNN의 출력은

$$y = \hat{x}(t+1) = F_{\omega,\, \boldsymbol{c}_t}(x(t)) \tag{5}$$

와 같은데, 식 (5)의 $F_{\omega,\, \boldsymbol{c}_t}(x(t))$는 연결 가중치 ω와 $\boldsymbol{c}_t$에 의해 결정되는 HCNN의 출력벡터를 의미한다.

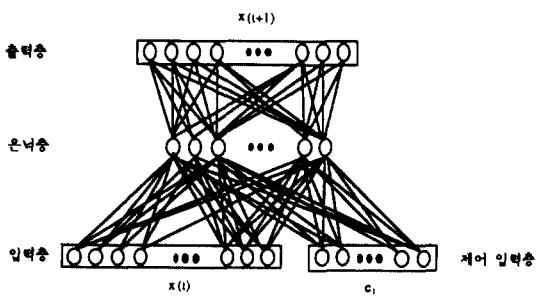


그림 1. HCNN의 구조
Figure 1. The Structure of the HCNN

여기서 제어 입력에 두 가지 제한이 가해진다. 첫째, 제어 입력의 종류를 유한한 개수 N으로 제한한다. 이렇게 되면 HCNN은 제어 입력의 종류 $\{C_1, C_2, \ldots, C_N\}$에 따라 입-출력 사상함수가 $\{F_1, F_2, \ldots, F_N\}$로 변화하는 유한 상태망이 된다. 예를 들어 제어 입력이 C_1에서 C_2로 바뀌면 network의 사상함수도 F_1에서 F_2로 바뀌게 되는데, 이것은 HCNN의 상태가 천이하는 것으로서, HMM에서의 상태천이와 동일한 개념이라고 볼 수 있다. 둘째, 단어 단위의 모델링을 위해 상태의 천이를 left-to-right 모델로 제한한다. 즉, 어떤 순간의 제어 입력이 C_i라면, 그 다음 순간에서의 제어 입력은 C_i이거나 C_{i+1}이어야만 한다는 제약을 두는 것이다. 위와 같이 제어 입력을 고려한 HCNN의 상태 천이도를 그림 2에 나타내었다.

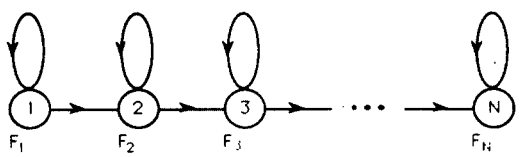


그림 2. N state Left-to-right HCNN의 상태 천이도
Figure 2. The state transition diagram of the N-state left-to-right HCNN

HCNN의 동작 원리를 이해하기 위해 음성을 인식하는 과정을 먼저 살펴보겠다. 그림 3은 HCNN을 이용한 간단한 단어 인식 시스템의 block diagram을 나타낸 것이다. M개의 단어를 인식하기 위한 M개의 HCNN이 구성되어 있으며, 각 HCNN 모델은 그에 해당되는 단어에 대해 학습이 되어 있다고 가정하자.

이제 시간축 길이가 T+1인 어떤 단어의 특징 벡터열 x(t)(t=0, 1, ..., T)이 있을 때, 각각의 HCNN 모델에 대해서 예측 오차가 최소가 되도록 하는 최적의 제어 입력열을 찾아야 한다. 그런 다음 예측 오차가 가장 작은 HCNN 모델을 인식된 결과로 출력하게 된다. 여기서 최적의 제어 입력열을 찾는다는 것은 단어를 최적으로 segmentation하는 HCNN의 상태열을 찾는다는 것을 의미하며, 이를 도식적으로 나타낸 것이 그림 4이다. 다시 말하면, HCNN은 시간에 따라 그 특성이 변화하는 음성 신호를 시불변 특성을 갖는 여러 구간의 열로 나누어, 그 각각의 구간을 연결가중치가 고정되어 있는 서로 다른 신경회로망들

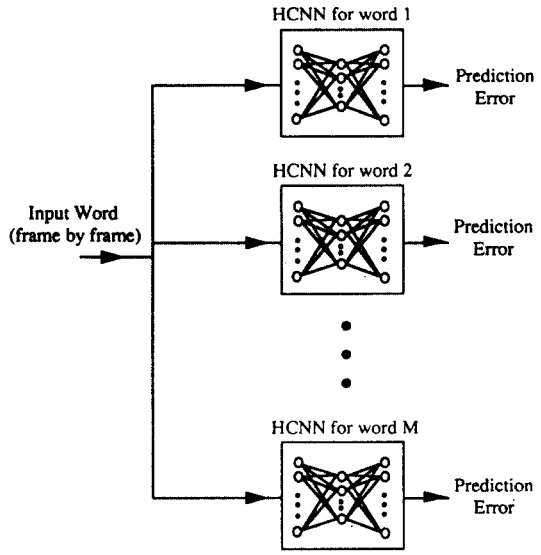


그림 3. 단어 단위로 구성된 HCNN을 이용한 단어 인식블럭도
Figure 3. The block diagram of the word recognition system consists of word-level HCNN

이 담당해서 모델링 하도록 되어있는 구조라고 볼 수 있다. 예측 오차는 순간 순간의 예측 오차를 단어 전체에 대하여 누적한 값인데, c_t가 시간 t에서 HCNN에 인가되는 제어 입력을, 그리고 c_0^{T-1}가 시간 t=0부터 T−1 까지의 제어 입력열 $c_0, c_1, \ldots, c_{T-1}$를 각각 나타낸다고 하면 예측 오차는

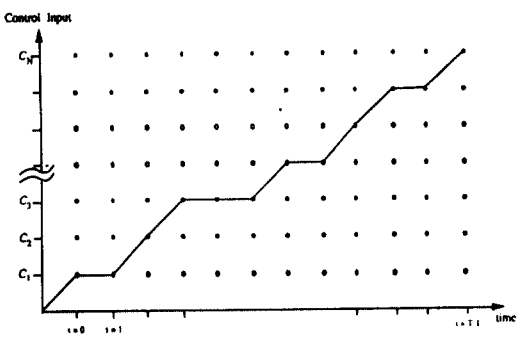


그림 4. HCNN의 상태열에 의한 단어 segmentation
Figure 4. Word segmentation by the state sequence of the HCNN.

$$E(\omega, c_0^{T-1}) = \sum_{t=0}^{T-1} \| x(t+1) - F_{\omega, c_t}(x(t)) \|^2 \qquad (6)$$

와 같이 표현된다. 그런데 학습이 완료된 HCNN은 연결가중치가 결정되어 있으므로 이 예측 오차는 제어 입력열만의 함수이다. 그러므로 최적의 제어 입력열을 찾기 위해서는 N가지 모든 제어 입력의 조합에 따른 예측 오차값을 비교해야 하겠지만, Viterbi 알고리즘[9]에 의해 예측 오차를 최소로 하는 최적의 제어 입력열 $\hat{c}_0^{T-1}$를 효율적으로 찾을 수 있게 된다.

2. 가중 거리(Weighted Distance Measure)가 도입된 HCNN

음성 인식, 패턴 인식 등의 분야에서는 저장 패턴과 입력 패턴을 비교할 때 거리의 개념이 중요하게 이용된다. 즉, 입력 패턴과 저장 패턴과의 거리를 측정하여 가장 가까운 저장 패턴을 인식된 결과로 간주하게 되는 것이다. 많이 사용되고 있는 거리로는 유클리드 거리, Mahalanobis 거리 등이 있다. 유클리드 거리는

$$d_E = \| x - y \|^2 \qquad (7)$$

와 같이 정의할 수 있으며, Mahalanobis 거리는

$$d_M = (x-y)^t V^{-1} (x-y) \qquad (8)$$

로 정의되며, 측정하고자 하는 패턴의 분포가 Gaussian일 때 많이 사용된다. 여기서 $\boldsymbol{x}$는 저장 패턴의 특징 벡터이며 $\boldsymbol{y}$는 입력 패턴의 특징 벡터이다. 또 $\boldsymbol{V}$는 특징 벡터의 공분산 행렬(covariance matrix)이다. 실제로 음성에서 켑스트럼을 특징 벡터로 추출했을 때 각 계수들의 분포는 거의 Gaussian에 가깝게 된다는 것이 알려져 있으므로, 유클리드 거리보다는 Mahalanobis 거리가 보다 적절한 거리 측정의 척도가 될 것이다. 하지만 특징 벡터의 차원이 N일 때, 유클리드 거리의 경우 N 번의 곱셈이 필요함에 반하여, Mahalanobis 거리의 경우는 공분산 행렬의 역행렬을 구해야 하며, 실제 문제에 적용할 때에도 $N(N+1)$번의 곱셈이 필요하게 되어 구현이 어려워 진다.

일본 ATR의 Tohkura[8]는 이런 문제점을 해결하기 위해 공분산 행렬의 대각선 성분만 이용하는 가중 거리(weighted distance)를 제안하였다. x_i와 y_i가 각각 N-차원 벡터 $\boldsymbol{x}$와 $\boldsymbol{y}$의 i 번째 성분이고, σ_{ii}는 i번째 성분의 표준편차일 때, 두 벡터 $\boldsymbol{x}$와 $\boldsymbol{y}$간의 가중 거리는

$$d_W = \sum_{i=1}^{N} \frac{1}{\sigma_{ii}^2} (x_i - y_i)^2 \qquad (9)$$

이다. 이 가중 거리는 벡터를 구성하는 성분들 사이의 분산도 차이를 규준화 시키는 효과가 있으므로, 보다 타당한 거리 측정의 지표로 간주될 수 있다. 또한, $N+1$ 번의 곱셈으로 구현이 가능하다.

이 가중 거리 개념을 HCNN의 예측 오차에 도입하면 식 (6)은 식 (10)로 표현된다.

$$E(\omega, c_0^{T-1}) = \sum_{t=0}^{T-1} \| A[x(t+1) - F_{\omega, c_t}(x(t))] \|^2$$
$$= \sum_{t=0}^{T-1} \sum_{i=1}^{N_y} \frac{1}{\sigma_{ii}^2} [x_i(t+1) - y_i(t)]^2 \qquad (10)$$

여기서 $\boldsymbol{A}$는 대각선 성분이 학습 데이타로부터 구한 특징 파라미터의 분산의 역수이고, 그 외의 값은 0인 행렬이다. 또 $y_i(t)$는 시간 t에서의 i번째 출력 뉴런 값이고 N_y는 출력 뉴런의 갯수이다. 그리고 입력되는 단어를 최적으로 segmentation하는 제어 입력열은 Ⅱ-1절에서와 마찬가지로

$$\hat{c}_0^{T-1} = \arg\min_{C^T} E(\omega, c_0^{T-1}) \qquad (11)$$

와 같이 추정할 수 있으며, 이는 Viterbi 알고리즘[9]에 의해 추정한다. 여기서 C^T는 제어 입력열의 전 공간을 의미한다.

3. 학습 알고리즘

HCNN의 학습이란 주어진 관측열, 즉 학습 데이타를 발생시킨 시스템을 잘 모델링할 수 있도록 HCNN의 연결 가중치를 결정하는 것을 말한다. 일반적인 다층 신경회로망의 경우 오차 역전달법(Error Back Propagation Algorithm)[10]으로 식 (3)의 예측 오차가 최소가 되도록 연결 가중치를 조정해 나가지만, HCNN의 경우는 예측 오차가 연결 가중치뿐만 아니라 제어 입력열의 함수이므로 학습은 joint minimization[5,6]에 의해 수행되어야 한다. Joint minimization이란 반복적인 방법으로 연결 가중치와 함께 제어 입력열을 추정하는 것을 말하는데, 임의의 k번째 반복시의 학습은 연결 가중치를 추정하는 reestimation 과정과

최적의 제어 입력열을 추정하는 segmentation 과정으로 나뉜다.

한편, 학습에 의해 추정되어야 할 연결 가중치는

$$\hat{\omega} = \arg\min_{\Omega} \{ \min_{C^T} E(\omega, c_0^{T-1}) \} \tag{12}$$

와 같이 표현할 수 있는데, reestimation은 고정되어 있는 제어 입력열에 대해서 예측 오차가 최소가 되도록 연결 가중치를 추정하는 것을 말한다. k번째 반복 학습시에 추정되는 연결 가중치는

$$(\hat{\omega})_k = \arg\min_{\Omega} E(\omega, (\hat{c}_0^{T-1})_{k-1}) \tag{13}$$

와 같으며, 오차 역전달법에 의해 구하게 된다. Segmentation은 reestimation 과정에서 얻어진 연결 가중치를 이용해서 예측 오차가 최소가 되도록 제어 입력열을 추정하는 것을 말하는데, k번째 반복시에 추정되는 제어 입력열은

$$(\hat{c}_0^{T-1})_k = \arg\min_{C^T} E((\hat{\omega})_k, c_0^{T-1}) \tag{14}$$

이며 Viterbi 알고리즘에 의해 구한다.

이 학습 알고리즘은 HMM의 학습 알고리즘의 하나인 근사적 MLE(Approximated Maximum Likelihood Estimation)와 등가임이 밝혀져 있다[5,6]. 그러나 HMM은 관측 기호가 현재 상태와 확률적인 분포함수에 의해 발생되며, 이러한 관측 기호들이 서로 확률적으로 독립이라는 제한을 두지만, HCNN의 경우는 예측을 통한 시변 시스템 모델링이라는 관점에서 음성을 인식하게 되므로 현재의 관측 기호가 과거의 관측 기호에 의존한다고 보게 되어, HMM과 같은 확률적인 제한 조건이 필요없게 된다.

Ⅲ. 실험 결과 및 고찰

본 연구에서 인식 대상으로 삼은 단어는 한국어 단음절 숫자음 10개로 선택하였는데, 이는 숫자음("영", "일", ..., "구")이 모두 단음절로 이루어져 있으며, 단어간의 구분이 비교적 어려운 편이기 때문이다. 예를 들면 "일"과 "칠", "삼"과 "사", "오"와 "구" 등은 사람이 인식할 때에도 오인식되는 경우가 많다. 음성 데이타는 방음 장치가 되어 있지 않은 실험실 환경에서 녹음된 12명의 남자가 5번씩 발음한 600개의 데이타와 또다른 2명의 남자가 2번씩 발음한 40개의 데이타로 구성하였다. HCNN의 학습은 5번씩 발음한 12명의 화자 중에서 임의로 6명을 선택하여, 그 6명이 3번씩 발음한 180개의 데이타로 수행하였고, 인식 실험은 학습에 사용되지 않은 사람들의 데이타를 사용하였다. 위와 같이 학습과 테스트에 사용될 데이타 추출을 조합을 달리하여 한번 더 반복함으로써 데이타군 I과 데이타군 II의 두 집합을 구성하였는데, 이는 한정된 음성 데이타로 화자 독립 음성 인식 성능 평가를 보다 객관적으로 하기 위함이다.

전처리 부분은 음성 인식에 있어서 최근 좋은 성과를 얻고 있는 켑스트럼과 차분 켑스트럼[11], 대수 에너지, 그리고 차분 대수에너지를 특징 파라미터로 사용하였다. 먼저 입력되는 음성 신호를 8kHz 표본화 비율의 A/D 변환기로 양자화한 후 음성 신호에 포함되어 있는 직류 성분이나 저주파 hum등을 제거하기 위해 0.95의 비율로 preemphasis 한다. 그 다음, 끝점 검출 알고리즘[12]을 사용하여 음성 부분만 골라낸 뒤 10차 선형 예측 분석(LPC analysis) 과정을 거쳐 켑스트럼 계수를 구하게 된다. 켑스트럼은 14차까지 구하며 이 벡터의 norm이 1이 되도록 정규화한 후 전후 각각 2 time-frame 만큼의 켑스트럼을 이용하여 차분 켑스트럼을 구한다. 대수 에너지 성분은 단어내의 최대값으로 정규화한 값을 사용하는데, 이는 발음할 때마다 달라지는 강약의 차이를 흡수하기 위함이다. 차분 대수 에너지는 차분 켑스트럼과 마찬가지로

표 1. 실험에 사용된 제어 입력값
(한 행의 각 열은 한 state에서 제어입력층의 각 뉴런에 인가되는 값을 나타낸다.)

Table 1. Control input values used in the experiment.
(Each column for one row indicates input values at one state assigned to each neuron in the control input layer of the HCNN).

C1	1.0	1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
C2	-1.0	1.0	1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
C3	-1.0	-1.0	1.0	1.0	-1.0	-1.0	-1.0	-1.0	-1.0
C4	-1.0	-1.0	-1.0	1.0	1.0	-1.0	-1.0	-1.0	-1.0
C5	-1.0	-1.0	-1.0	-1.0	1.0	1.0	-1.0	-1.0	-1.0
C6	-1.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0	-1.0	-1.0
C7	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0	-1.0
C8	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0

40 msec만큼의 대수 에너지 성분을 이용하여 구한다. 이렇게 하여 한 time-frame당 30차원의 특징 벡터가 생성된다.

실험에 사용된 HCNN은 출력층의 뉴런수 30개, 은닉층 뉴런수 40개인 구조이다. 또한 제어 입력의 종류를 8 가지로 제한하였으며 제어 입력층의 뉴런에 할당되는 값을 표 1에 나타내었다.

제어 입력의 종류가 8가지인 이유는 한 음절을 초성, 중성, 종성 그리고 초성과 중성간의 천이, 중성과 종성간의 천이의 5 부분으로 나뉘어 HCNN의 5개의 상태가 음절 내의 각 부분을 담당하도록 하되 여기에 약간의 여유분을 두어 8개의 상태를 갖도록 한 것이다. 또한 뉴런의 비선형 함수는 기울기 $\alpha=0.3$인 식 (15)와 같은 시그모이드 함수를 사용하였다.

$$f(x)=\frac{2}{1+e^{-\alpha x}}-1 \tag{15}$$

1. 특징 벡터의 분석

가중 거리를 HCNN의 예측 오차 측정에 이용하기 위해서는 학습 데이타로부터 특징 벡터의 각 성분에 대해 분산을 구해야 한다. 그런데 분산을 이용한 가중 거리를 실제 문제에 적용하기 위해서는 특징 벡터에 대한 공분산 행렬에서 주 대각선 성분 이외의 값은 무시될 수 있다는 가정이 성립해야 한다. 그림 5는 이 가정의 타당성을 검증하기 위한 것으로서, 본 실험에 사용된 데이타로부터 구해진 특징 벡터 각 성분간의 공분산을 나타낸 것이다. 각 그림에서 $\sigma(i, j)$는 특징 벡터의 i번째 성분과 j번째 성분과의 공분산을 의미하며, i와 j가 같은 경우에는 분산을 나타내게 된다. 그림으로부터 알 수 있듯이, 거의 모든 경우에 있어서 공분산 행렬의 주 대각선 성분이 우세한 값을 나타내고 있으므로, 그 외의 값은 무시할 수 있다.

그림 6은 데이타군 I 중에서 학습에 사용될 데이타로부터 특징 벡터를 추출했을 때, 벡터의 각 성분에 대한 도수분포표를 그린 것이다. 그림에서 C_i와 delta-C_i는 각각 i번째 켑스트럼 계수와 차분 켑스트럼 계수를 나타는데, 계수의 차수에 관계없이 거의 Gaussian에 가까운 분포를 보이지만, 각 계수에 대한 분산도는 다름을 알 수 있다. 즉, C1의 경우 −1과 1 사이의 영역 전체에 그 값이 분포하지만, 차수가 점점 증가함에 따라 그 분산도가 감소하게 된다. 차수가 증가할 수록 분산도가 작아진다는 것은 식 (9)에 나타나는 거리 측정시에 켑스트럼의 높은 차수에 대해서 더 비중을 두어 측정한다는 것을 의미한다. 그리고, 그

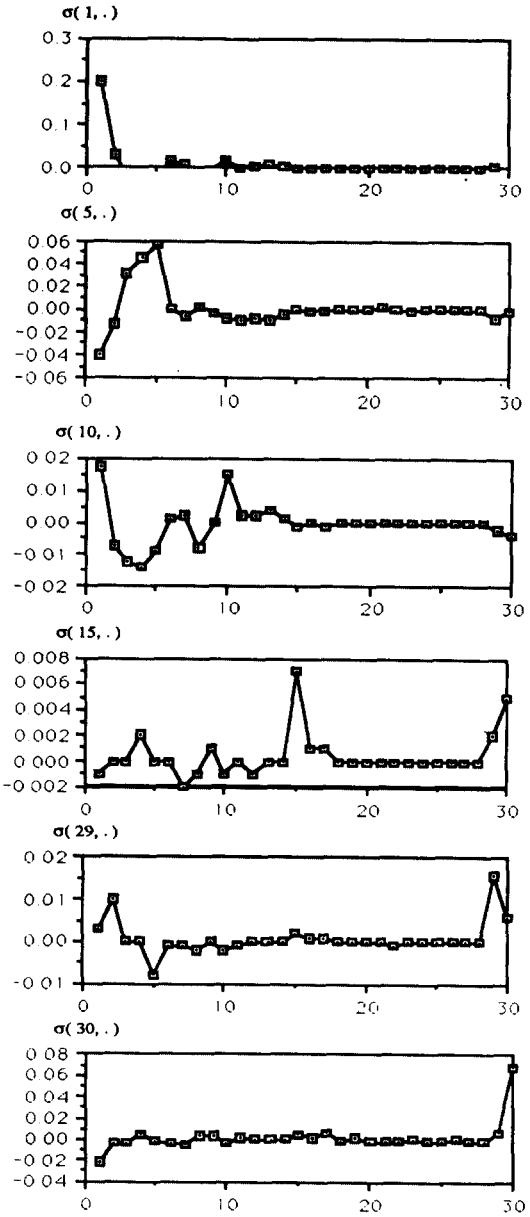


그림 5. 특징 벡터를 구성하는 각 성분간의 공분산
Figure 5. Covariances of the feature vector components.

림 7은 각각 데이타군 I과 데이타군 II 중에서 학습에 사용된 데이타로부터 측정한 특징 벡터 구성 성분의 index에 따른 분산과, 가중 거리 측정시 곱해지는 요소인 가중치, 즉 분산의 역수를 나타낸 것이다.

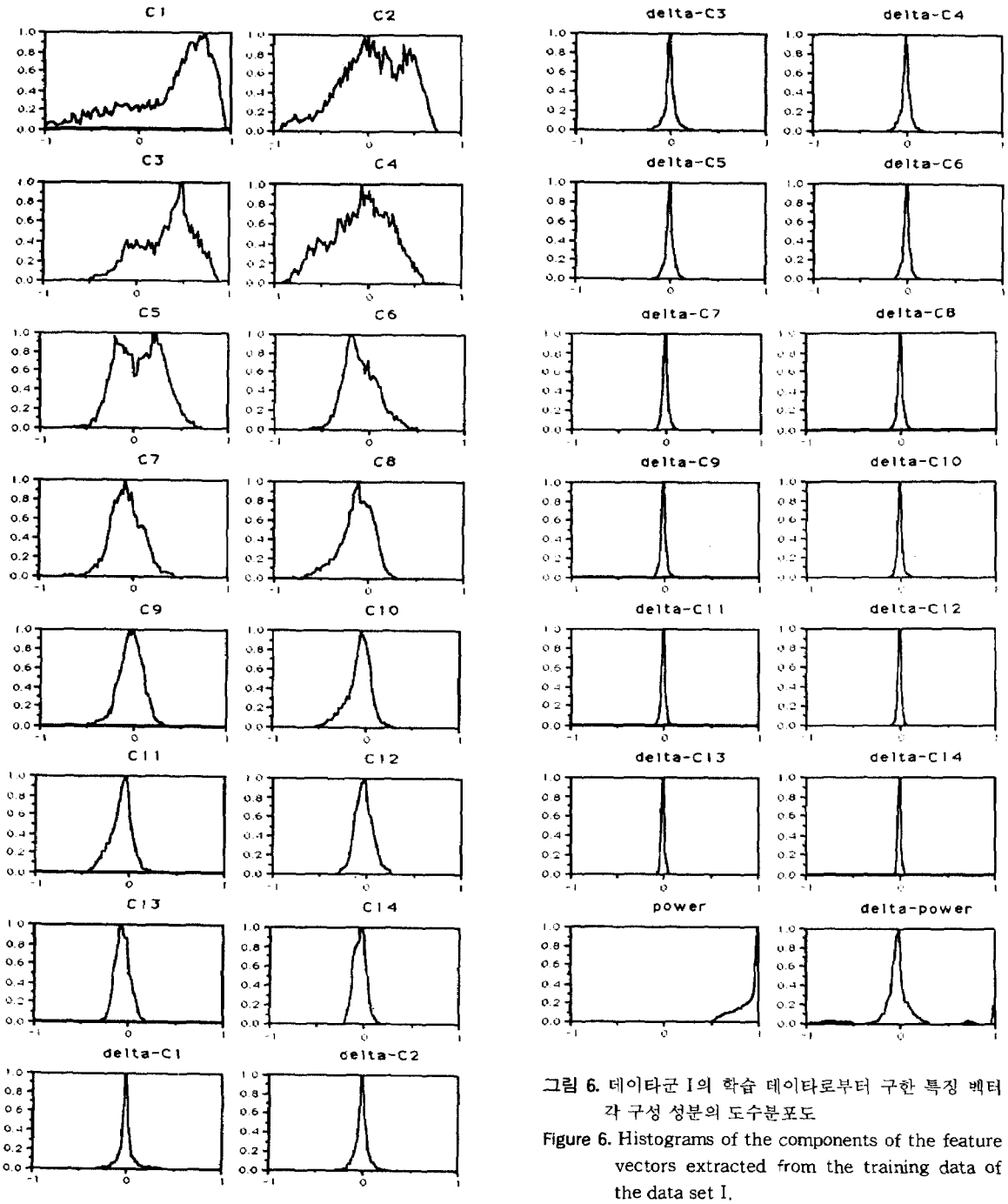
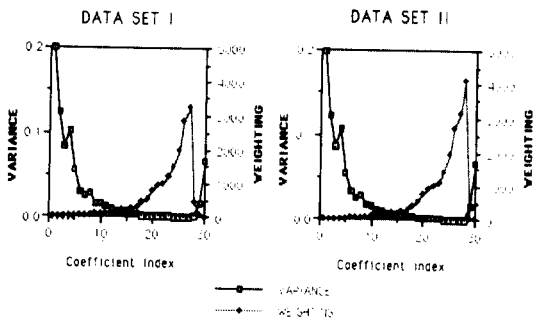


그림 6. 데이타군 I의 학습 데이타로부터 구한 특징 벡터 각 구성 성분의 도수분포도

Figure 6. Histograms of the components of the feature vectors extracted from the training data of the data set I.



(a) (b)

그림 7. 특징 벡터 구성 성분의 분산과 가중 거리의 가중치
(a) 데이타군 I의 학습 데이타
(b) 데이타군 II의 학습 데이타

Figure 7. Feature vector components variances and weighting values of the weighted distance extracted from the raining data of the (a) data set I and (b) data set II.

2. 인식 성능 비교

본 절에서는 HMM, 유클리드 거리를 이용한 HCNN, 분산의 역수로 가중치가 가해진 가중 거리를 도입한 HCNN의 성능을 비교·검토하겠다.

학습 조건은 다음과 같다. 유클리드 거리를 이용한 HCNN은 학습율을 0.18로 하였고, 가중 거리를 이용한 HCNN의 경우는 특징 벡터 성분의 차수가 증가함에 따라 가중값이 급격하게 커져서 식 (10)의 예측 오차가 매우 커지게 되므로, 이를 보상해주기 위해 학습율을 0.009로 작게 하였다. HMM은 HCNN과 같은 조건 하에서 비교하기 위해 8 state, left-to-right model discrete HMM을 사용하였다. 그리고 HMM에서의 벡터 양자화를 위한 codebook은 그 크기를 64로 하였으며 LBG 알고리즘[13]을 이용해 작성하였다.

가중 거리를 이용한 HCNN의 경우는 데이타군 I에 대해서 95.29%의 인식율을 얻었다. 이것은 유클리드 거리를 이용한 HCNN에 비해 2.35%, HMM에 비해 3.82% 향상된 결과였다. 또한 데이타군 II에 대해서는 HMM이 93.82%, 유클리드 거리를 이용한 HCNN이 95%의 인식율을 보여 준 반면, 가중 거리를 이용한 HCNN은 97.35%의 인식율을 나타내었다. 이것을 표 2에 나타내었으며, HCNN에 의한 음성의 동적인 모델링이 확률적인 전제 조건 하에서 출발한 HMM보다 우수함을 알 수 있다. 또한 가중 거리를 HCNN에 도입하는 경우에 더욱 향상된 인식 결과를 얻을 수 있었다. 이것은 유클리드 거리가 특징 벡터 각 계수간의 변이를 전혀 고려하지 않은 거리임에 반하여, 가중 거리는 학습 데이타로부터 미리 얻은 각 계수의 분산도 차를 거리 측정에 고려하기 때문이다.

표 2. 인식율 비교
Table 2. Comparison of recognition rate.

	HMM	HCNN(Euclidean distance)	HCNN(weighted distance)
SET I	91.47%	92.94%	95.29%
SET II	93.82%	95.00%	97.35%

한편, 그림 8은 화자별 오인식율의 차이를 상기 기술한 세 가지 방법에 대해 각각 나타낸 것이다. 그림을 보면 HMM과 유클리드 거리를 이용한 HCNN은 화자간의 인식율 차이가 매우 심함을 알 수 있다. 특히 몇몇 화자의 오인식율이 상대적으로 매우 커서 전체의 오인식율에 큰 영향을 미치고 있다. 이에 반해 가중 거리를 도입한 HCNN의 경우는 오인식이 많이 되는 화자의 인식율이 높아져서 화자간의 인식율차가 줄어들게 되었으며, 이로 인해 전체적인 인식율의 향상을 가져오게 되었음을 알 수 있다. 여기서 화자간의 인식율차는 한정된 데이타로 화자 독립 음성 인식 성능을 평가하는 한 척도가 되므로, 가중거리가 도입된 HCNN은 다른 방법에 비해 화자간 인식율차가 적어 화자 독립 음성 인식 성능이 우수함을 알 수 있다.

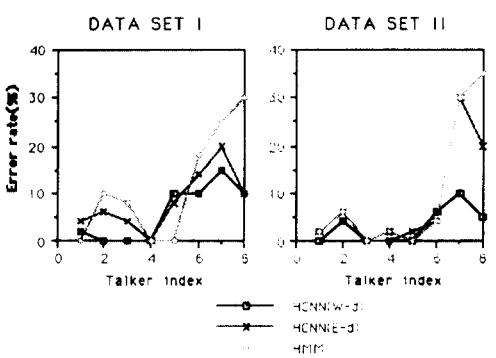


그림 8. 화자별 오인식율 비교

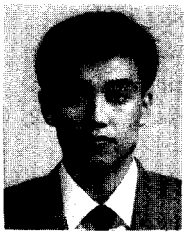
Figure 8. Comparison of recognition error rate as a function of talkers.

Ⅳ. 결 론

예측에 기반을 둔 HCNN은 HMM의 상태 개념을 신경회로망에 도입하여 시간에 따라 network의 사상 함수를 변화시킴으로써, 모델링하고자 하는 신호의 nonstationary한 특성을 잘 포착할 수 있는 구조로 되어 있다. 본 연구에서는 먼저 HCNN에 의한 시스템의 동적인 모델링이 확률적인 제한 조건하에서 성립하는 HMM보다 우수함을 보였으며, 또 특징 벡터 각 성분의 분포 변이를 고려한 가중 거리를 HCNN의 예측 오차 측정에 도입하고, 비교적 구분하기 어려운 한국어 숫자음에 대해 화자 독립으로 인식 실험을 수행하여, 그 성능이 유클리드 거리를 이용한 HCNN보다 우수함을 보였다. 실험 결과 유클리드 거리를 이용한 HCNN의 경우 데이타군 II에 대하여 95.0%의 인식율을 얻었는데, 이는 HMM에 비해 1.18% 향상된 결과로서, HCNN에 의한 시스템의 동적인 모델링이 HMM보다 우수함을 알 수 있었다. 또한 가중거리를 HCNN에 도입했을 때 97.35%의 인식율을 얻게되어, 특징 벡터 각 계수간의 분포 변이를 전혀 고려하지 않은 유클리드 거리에 비해, 각 계수의 분산도 차를 거리 측정에 고려한 가중거리를 HCNN에 도입했을 때 더욱 우수한 성능을 보임을 알 수 있었다. 가중거리를 도입한 HCNN이 높은 인식율을 나타내는 것은, 오인식율이 높은 화자의 인식율을 높임으로써 화자간 인식율차를 줄이는 효과가 있기 때문이며, 따라서 화자 독립 음성 인식에 가중거리를 도입한 HCNN이 보다 적합함을 알 수 있었다.

참 고 문 헌

1. R.P. Lippman, "An Introduction to Computing with Neural Nets," *IEEE ASSP magazine*, pp. 4-22, April 1987.
2. R.P. Lippman, "Review of Neural Networks for Speech Recognition," *Neural Computation*, Vol. 1, pp. 1-38, 1989.
3. H. Bourlard, "Neural Nets and Hidden Markov Models: Review and Generalizations," *Eurospeech 91*, Vol. 2, pp. 363-369, September 1991.
4. H. Bourlard and C.J. Wellekens, "Links Between Markov Models and Multilayer Perceptrons," *IEEE Trans. on PAMI*, Vol. 12, No. 12, pp. 1167-1178, December 1990.
5. E. Levin, "Word Recognition Using Hidden Control Neural Architecture," *Proc. of ICASSP*, pp. 433-436, April 1990.
6. E. Levin, "Modeling Time Varying Systems Using Hidden Control Neural Architecture," *In Advances in Neural Information Processing Systems*, NIPS-3, pp. 147-154, June 1991.
7. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of IEEE*, Vol. 77, No. 2, pp. 257-286, February 1989.
8. Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. on ASSP*, Vol. ASSP-35, No. 10, pp. 1414-1422, October 1987.
9. G.D. Forney, "The Viterbi Algorithm," *Proc. of IEEE*, Vol. 61, pp.268-278, March 1973.
10. D. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Internal Representation by Error Propagation," *Parallel Distributed Processing, D.Rumelhart and J. McClelland (Eds.)*, Vol. 1, MIT Press, 1986.
11. S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, 80, 4, pp. 1016-1025.
12. L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. on ASSP*, Vol. 29, pp. 777-785.
13. Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. on COMM*, Vol. COM-28, No. 1, January 1980.



金 度 錫(Do Seok Kim) 정회원
1968년 2월 23일생
1991년 2월 : 한양대학교 전자공학과(공학사)
1993년 2월 : 한국과학기술원 전기 및 전자공학과(공학석사)
1993년 3월~현재 : 한국과학기술원 전기및 전자공학과 박사과정
1993년 1월~현재 : 시스템공학연구소 연구원
※주관심분야 : 신경회로망, 음성인식 등

李 壽 永(Soo Young Lee) 정회원
1952년 1월 15일생
1975년 : 서울대학교 전자공학과(학사)
1977년 : 한국과학기술원 전기및 전자공학과(석사)
1984년 : Polytechnic Institute of New York(박사)
1977년~1980년 : 대한엔지니어링 주식회사 근무
1985년 : 미국 General Physics Corp. 근무
1986년~현재 : 한국과학기술원 전기및 전자공학과 근무, 교수.
※주관심분야 : 신경회로망, 수치해석 등