論文95-32B-5-12

화자적응 신경망을 이용한 고립단어 인식

(Isolated Word Recognition Using a Speaker-Adaptive Neural Network)

李 基 熙 *, 林 寅 七 *

(Kee Hee Lee and In Chil Lim)

요 약

본 논문은 MLP(MultiLayer Perceptron)에 기초한 HMM(Hidden Markov Model) 음성 인식기의 인식성능을 개선하기 위한 화자적응 방법을 기술한다. 이 방법에서는 새로운 화자의 데이터를 MLP에 적응시키기 위해 1차 선형변환망을 사용한다. 변환 파라미터의 값은 분류오차를 변환망으로 역전파하여 조정하며, 이때의 MLP 분류기는 변하지 않는다. 인식시스템은 MLP를 퍼지 벡터양자화기로 사용하는 준연속 HMM에 기초한다. 인식실험의 결과는 이 방법이 빠른 속도로 화자에 적응하며, 높은 인식성능을 가짐을 보여준다. 즉, 지도적응일 경우, 오인식률은 기본시스템의 9.2%에서 화자적응 후의 5.6%로 크게 감소되었다. 또, 무지도적응일 경우, 새로운 화자에 대한 정보가 없어도 오인식률은 5.1%로 감소되었다.

Abstract

This paper describes a speaker adaptation method to improve the recognition performance of MLP(multiLayer Perceptron) based HMM(Hidden Markov Model) speech recognizer. In this method, we use 1st-order linear transformation network to fit data of a new speaker to the MLP. Transformation parameters are adjusted by back-propagating classification error to the transformation network while leaving the MLP classifier fixed. The recognition system is based on semicontinuous HMM's which use the MLP as a fuzzy vector quantizer. The experimental results show that rapid speaker adaptation resulting in high recognition performance can be accomplished by this method. Namely, for supervised adaptation, the error rate is significantly reduced from 9.2% for the baseline system to 5.6% after speaker adaptation. And for unsupervised adaptation, the error rate is reduced to 5.1%, without any information from new speakers.

I. 서 론

음성인식 시스템은 특정화자의 음성만을 인식하는 화자종속 시스템과 모든 화자의 음성을 인식할 수 있는 화자독립 시스템으로 나누어진다.

화자종속 시스템은 특정한 화자의 음성 데이타로부터 훈련되고, 또한, 이 화자의 음성을 인식하므로 고립단어 인식이나 연속음성 인식에서 우수한 성능을 보인다. 반면, 화자독립 시스템은 훈련에 참가하지 않은 불특정화자의 음성을 인식하므로, 비교적 저조한 인식성능을 나타내며, 이를 개선하기 위한 여러가지 방법이 발표되고 있다. 화자종속 및 화자독립 음성인식기에서 인식성능을 저하시키는 원인으로는 1)화자의 발성습관, 성문 및 성도의 길이와 모양의 차이, 2)인접한 단어간

* 正會員, 漢陽大學校 電子工學科
(Dept. of Elec. Eng., Hanyang Univ.)

接受日字:1994年12月5日, 수정완료일:1995年4月20日

의 상호 조음효과(coarticulary effect), 그리고 3)주변 잡음, 반향(reverberation) 및 마이크 특성과 같은 주변환경의 차이 등을 들 수 있다. 이와 같은 여러가지 원인으로 음성신호의 피치, 포만트 및 스펙트럼 기울기(spectral tilt)와 같은 음성 특징은 화자에 따라 다양한 차이를 보이게 된다.[1-3]

화자적응(speaker adaptation) 기법은 화자종속 시스템을 새로운 화자의 음성에 적응시켜 인식성능을 개선하는 기술이지만, 최근에는 화자독립 시스템에도 이를 적용하여 인식기의 성능을 개선시키고 있다. 화자적응 인식기는 새로운 화자로부터 얻은 약간의 음성정보를 이용하여 인식시스템을 새로운 화자에 적응시킴으로써, 그 화자의 나머지 음성을 보다 신뢰성있게 인식하도록 해준다. 화자적응은 크게 지도적응(supervised adaptation)[4-9]과 무지도적응(unsupervised adaptation)[10,11]으로 나눌 수 있다. 전자는 화자적응을 위한 교정데이타(calibration data)를 인식시스템에 알려주고 적응시키는 방법으로, 새로운 화자는 적응과정 동안 미리 지정된 어휘를 발음한다. 이 방법은 교정데이타가 지정되어 있으므로 알고리듬이 비교적 간단하고, 적응도(adaptation degree)가 높으므로 많이 연구되고 있다. 한편, 후자의 무지도적응은 새로운 화자가 임의의 어휘를 발음하여도 이에 적응하는 방법이나 현재로서는 만족할 만한 성능을 얻지 못하고 있다. 적응기법은 인식기의 종류에 따라 여러가지가 있으며, 대부분 HMM인식기에 기초를 둔다. 이산 HMM(discrete HMM)에 기초한 인식기인 경우 코드북 적응(codebook adaptation)과 HMM 파라메타 적응[4,5]을 사용하여 원래의 HMM을 새로운 화자에 적응시키고, 연속 HMM(continuous HMM) 인식기인 경우에는 평균벡터 적응(mean vector adaptation)과 공분산행렬 적응(covariance matrix adaptation) 기법[6,7] 등의 방법이 사용된다. 최근에는 지도 스펙트럼 사상(supervised spectral mapping)을 통해 특징벡터를 변환하는 방법[8], 신경망의 비선형 사상(nonlinear mapping)에 기초한 LVQ(Learned Vector Quantization)를 이용하는 방법[9] 등 다양한 적응방법이 연구되고 있다.

여러가지 적응방법에서 공통의 문제점으로는 교정데이타의 부족과 과다한 적응시간을 들 수 있다. 교정데이터의 수가 충분히 많으면 인식 시스템은 훈련을 통해 새로운 화자에 적응되므로 인식성능은 개선되나 적응시간이 매우 길어지게 된다. 일반적으로 교정데이터의 수는 극히 제한적일 수 밖에 없다. 이 경우 새로운 화자의 음성에 대한 적응시간은 빠르지만, 적응된 시스템의 파라미터는 신뢰도가 낮고 경우에 따라서는 잘못된 교정데이터로 인해 인식성능이 오히려 악화될 수도 있다.

본 논문에서는 다층신경망을 이용한 지도적응 및 무지도적응에 의한 화자적응 기법을 제시한다. 인식시스템에서는 1차 선형변환망을 이용하여 음성신호를 다층퍼셉트론에 적응하게 하고, 신경망의 출력값으로 부터 준연속 HMM(semicontinuous HMM)을 이용하여 음성인식을 수행한다. 선형변환망은 입력음성을 주변환경에 적응시키고, 스펙트럼을 정규화하여 다층퍼셉트론의 입력층으로 보내게 된다. 이 변환망의 파라미터는 다층퍼셉트론의 출력층에서 인식오차를 변환망으로 역전파하여 계산한다. 이와 같은 적응방법은 모든 HMM에 공통으로 사용되는 다층신경망의 전처리 과정에서 화자적응을 수행하므로 교정데이타의 수가 적을 경우에도 인식성능을 개선할 수 있고, 잘못된 교정데이타로 인한 성능 악화를 방지할 수 있다. 또, 변환망의 파라미터값은 출력층의 오차로부터 계산되므로 적응과정에 소요되는 시간을 단축할 수 있다. 제안한 방법의 성능은 화자독립 단어인식 실험을 통해 보인다.

Ⅱ. 인식 시스템의 개요

이 장에는 MLP(MultiLayer Perceptron)와 HMM을 이용한 기본적인 인식시스템을 기술한다.

1. MLP 확률추정기(probability estimator)

다층퍼셉트론은 하나의 입력층과 몇개의 은닉층, 그리고 하나의 출력층을 가지는 구조의 신경망으로 다양한 분야에서 응용되고 있다. MLP는 입력벡터를 원하는 출력벡터에 연관되도록 학습된다. M가지의 음소를 분류인식하는 MLP인 경우, 각 음소에 대응하는 하나씩의 출력노드가 있으므로 출력노드의 수는 M개가 된다. MLP는 자기의 음소에 대한 출력노드의 값을 1로, 다른 음소에 대해서는 0이 되도록 학습된다. 이러한 MLP의 출력값은 입력벡터의 확률을 추정하는 데 이용할 수 있다. 즉, 입력 X에 대해 음소 class c_i에 대응하는 출력노드의 값은 주어진 입력 X에 대한 사후확률(posterior probability) $p(c_i \mid X)$의 추정값이 된다.[12] MLP 확률추정기를 HMM과 결합할 때, 음소 class c_i에서 X의 확률 $p(X \mid c_i)$ 즉, 유사성(likelihood)은 다음의 식으로 구할 수 있다.

$$p(X \mid c_i) = \frac{P(c_i \mid X)}{P(c_i)} p(X) \qquad (1)$$

여기서 $P(c_i \mid X)$는 MLP의 출력값으로 X의 사후

확률이고, $P(c_i)$는 음소 c_i의 확률로서 c_i의 상대빈도가 되며, $P(X)$는 X의 확률이다.

2. 준연속 HMM

이산 HMM에 기초한 인식 시스템에서 VQ(Vector Quantizer)는 입력벡터에 가장 잘 정합되는 코드워드의 심볼을 발생시킨다. 벡터 양자화기는 음성공간을 Voronoi cell로 분할하므로 원래의 음성 공간을 충분히 표현할 수 없다. 이는 VQ 오류를 일으키고 인식기의 성능을 떨어뜨리는 원인이 된다. 한편, FVQ(Fuzzy VQ)[13]는 이러한 단점을 줄이는 방법으로서, 입력벡터가 각 코드워드에 정합되는 정도를 나타내는 퍼지 score를 발생시킨다. 또, MLP-FVQ[14]는 MLP를 FVQ로 사용하여 HMM의 최적화와 더불어 밀접하게 결합된다. HMM은 다음과 같이 구성된다. 먼저 초기확률 벡터를 $\pi=(\pi1,\pi2,\ldots,\pi S)$로, 상태천이 행렬을 $A=[a_{ij}]\ (1\le i,\ j\le S)$로, 또 관측확률 행렬을 $B=[b_{jm}]\ (1\le j\le S,\ 1\le m\le M)$로 각각 표현하기로 하자. 여기서 S와 M은 각각 상태의 수 및 FVQ 출력레이블의 수이며, b_{jm}은 상태 j에서 출력레이블 v_m의 확률을 나타낸다. HMM은 파라미터의 집합 $\lambda=(\pi, A, B)$로 표현된다. MLP는 입력벡터열 $X=(X_1, X_2, \ldots, X_T)$ (T는 입력음성의 길이)를 출력벡터열 즉, 퍼지관측열 $O=(O_1, O_2, \ldots, O_T)$로 변환한다.

$O_t=(O_{t1}, O_{t2}, \ldots, O_{tM})$로서 O_{tm}은 MLP의 m번째 출력노드의 값이고, M은 노드의 수이다. HMM λ의 상태 j에서 O_t의 관측확률은

$$\begin{aligned} W_t(j) &= P(O_t \mid s_j, \lambda) \\ &= \sum_{m=1}^{M} b_{jm} O_{tm} \end{aligned} \tag{2}$$

로 표현되며, 여기서 s_j는 상태 레이블이고 O_{tm}은

$$O_{tm} = \overline{O}_{tm} \ / \ \sum_{m=1}^{m} \overline{O}_{tm} \tag{3}$$

으로 정규화한 출력값으로

$$\sum_{m=1}^{M} O_{tm} = 1 \tag{4}$$

이 되도록 제한된다. 이와 같이 입력벡터열 X를 MLP-FVQ로써 퍼지관측열 O로 변환하고 상태 j의 새로운 관측확률 $W_t(j)$를 구하면, HMM의 파라미터 행렬 A와 B는 Baum-Welch 재추정식[15,16]을 이용하여 구할 수 있다. 기본 인식시스템은 훈련과정을 통해 모든 기준 HMM을 만들어 두고, 인식과정에서는 모든 HMM에서 입력음성의 관측확률을 구하여 가장 높은 확률을 가지는 HMM의 단어를 인식된 단어로 판정한다.

3. 입력벡터열의 관측확률

HMM의 λ는 S개의 상태를 가지는 단순 좌우(simple left-to-right)구조로서 입력음성 $O=(O_1, O_2, \ldots, O_T)$의 관측확률은

$$P(O \mid \lambda) = \sum_{\theta \in \Omega_S} \pi_{\theta O} \prod_{t=1}^{T} a_{\theta t-1,\ \theta t} W_t(\theta_t) \tag{5}$$

이다. 여기서 $\theta=(\theta_0, \theta_1, \ldots, \theta_T)$는 Markov chain의 상태열이고, Ω_s는 상태열의 공간을 나타낸다. 한편, 상태지속시간은 음성세그먼트의 길이를 나타내는 중요한 정보로서 HMM 인식기에 이를 고려할 필요가 있다. HMM의 특정한 상태에서 상태지속시간의 확률은 기하분포(geometric distribution) 형태가 되므로 음성세그먼트의 길이 정보를 충실히 표현하지 못하는 단점이 있다. 이를 해결하는 방안으로는 지속시간의 확률분포를 구해두고 인식시 관측확률에 포함시키는 방법[17]과 지속시간의 확률을 HMM에 구조적으로 통합시키는 방법[16,18]이 있다. 본 논문에서는 알고리듬을 단순히 하기 위해서 지속시간의 확률분포를 가우스 분포로 근사화 하였다. 즉, 상태 j에서 지속시간이 τ_j일 확률은

$$d_j(\tau_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-(\tau_j - m_j)^2 / 2\sigma_j^2\right] \tag{6}$$

로서, m_j와 σ_j^2은 각각 상태지속시간의 평균 및 분산이다. 인식시 지속시간을 고려한 관측확률은

$$P_D(O \mid \lambda) = P_D(O \mid \lambda)\ [\prod_{j=1}^{S} d_j(\tau_j)]^{\gamma} \tag{7}$$

로 계산되며, γ는 지속시간 확률의 상대적인 비중을 나타내는 상수로서 이 논문에서는 $\gamma=2.0$ 이다.

Ⅲ. 다층신경망을 이용한 화자적응 음소인식

이 장에서는 변환망 및 다층퍼셉트론을 이용한 화자적응 음소인식 방법에 대해 기술한다.

1. 변환망과 다층퍼셉트론

화자적응을 위한 다층신경망은 그림 1과 같이 하나의 은닉층을 갖는 MLP와 1차의 변환망으로 구성된다.

각 음성 프레임의 특징벡터는 N차의 LPC 켑스트럼 C(i)과 로그 에너지 D로서, 시간 t에서 입력벡터 $X_t = \{C(1), C(2), \ldots, C(N), D\}^T$로 표현된다. 그림에서

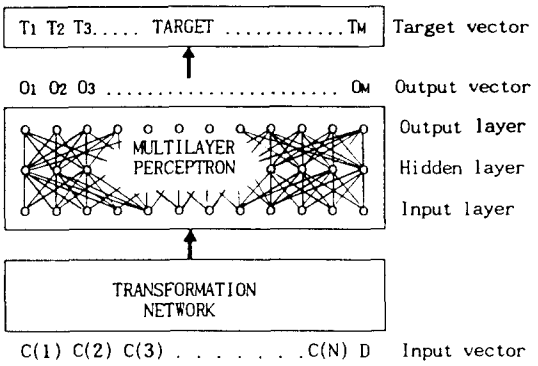


그림 1. 화자적응 다층신경망
Fig. 1. Speaker-adaptive multilayer neural network.

아래쪽의 망은 화자적응을 위한 1차 선형변환망으로서 입력벡터를 적절히 변환하여 신경망에 입력하는 역할을 한다. 입력벡터의 변환은

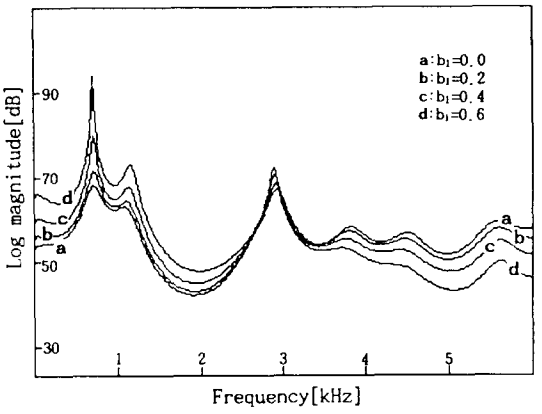
$$\overline{X}_t = \mathbf{P}X_t + \mathbf{B} \qquad (8)$$

로 이루어지며, 여기서 $P=[p_{il}]$ $(1 \le i, l \le N)$와 $B=[b_i]$ $(1 \le i \le N+1)$는 새로운 화자의 특정벡터 X_t를 기존의 MLP에 적응시키는 행렬이고, 스펙트럼이 변환된 특징벡터 $\overline{X}_t$는 MLP의 입력값이 된다. 화자의 성문(glottis)특성으로 인한 스펙트럼의 기울기는 화자간 스펙트럼의 차이를 일으키는 주된 원인이 된다.[6] 또, 음성신호에 주변잡음이 더해질 때도 스펙트럼의 기울기가 변화된다. 켑스트럼 계수의 낮은 차수부분은 스펙트럼 기울기에 영향을 주므로 행렬 B의 b_1과 b_2 같은 낮은 차수의 값은 스펙트럼의 기울기를 조정할 수 있다. 이의 예로서, 그림 2는 음소 "아"의 특징벡터에 b_1과 b_2의 값만을 변화시킨 스펙트럼을 나타낸다. 그림 2(a)에서는 b_1의 증가에 따라 스펙트럼의 고주파 성분은 감소되고, 저주파 성분은 증가되어 전체적인 스펙트럼의 기울기가 변화됨을 알 수 있다. 그림 2(b)에서는 b_2의 증가에 따라 고주파 및 저주파 성분은 증가되고, 중간주파수 영역은 감소됨을 알 수 있다. 한편, P는 입력신호의 주파수 변환을 위한 N×N행렬로서 새로운 화자의 스펙트럼을 MLP에 적응시키는 역할을 한다.

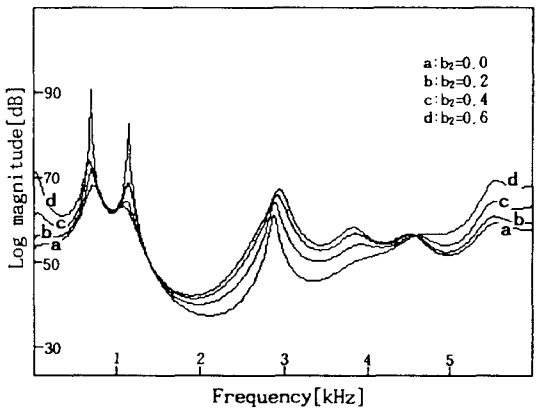
선형변환망의 파라미터인 P와 B는

$$\mathbf{P} = \begin{bmatrix} 1\,0 \cdots\cdots 0 \\ 0\,1\;0 \cdots 0 \\ \cdots\cdots 1 \cdots\cdots \\ \cdots\cdots\cdots 0 \\ 0 \cdots\cdots 0\,1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

로 초기화된다. 신경망의 학습시 P와 B는 식(9)의 단위행렬 및 영벡터로 고정되며, 신경망의 가중치만 학습된다.



(a)

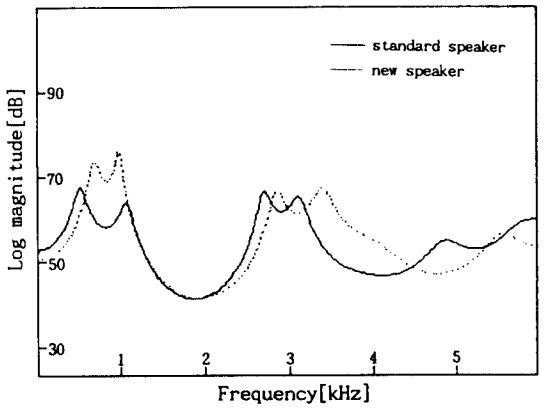


(b)

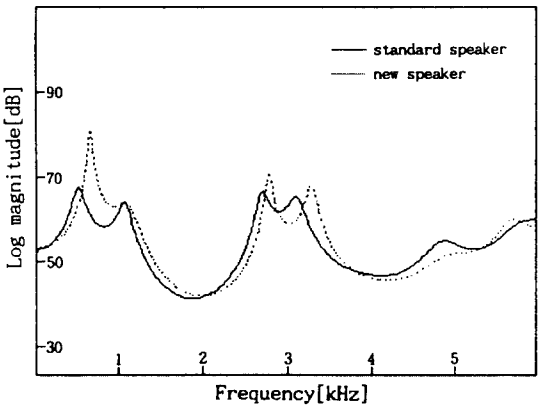
그림 2. 스펙트럼 바이어스 (a) b_1와 (b) b_2의 효과
Fig. 2. Effects of spectral biases (a) b_1 and (b) b_2.(b)

그림 3은 표준화자의 음소 "어"(실선)에 새로운 화자의 음소 "어"(점선)를 적응시킨 예로서, 그림 3(a)는 원래의 두 스펙트럼이고, 그림 3(b)는 B와 P를 조정

하여 적응시킨 후의 스펙트럼을 나타낸다. 점선으로 표시된 새로운 화자의 스펙트럼은 전체적으로 표준화자의 스펙트럼에 근접하고 있으나 두 스펙트럼은 약간의 차이를 보이고 있다. MLP는 입력음소에 대응하는 노드의 출력값을 1에 접근시키고, 다른 노드의 축력값은 0으로 억제하는 특성이 있다. 따라서 변환망은 새로운 화자의 스펙트럼을 표준화자에 유사하도록 변환하기도 하지만, 다른 음소와의 변별력도 증대시켜야 하므로, 음소 고유의 특징점을 제외한 나머지 부분에서는 오차가 생길 수 있다.



(a)



(b)

그림 3. 음성 스펙트럼 (a) 적응전 (b) 적응후
Fig. 3. Speech spectrum. (a) Before adaptation, (b) After adaptation.

2. 화자적응 음소인식

그림 1의 다층신경망을 새로운 화자에 적응시킬 때는 먼저 P는 단위행렬로 초기화하고, B의 요소 값은

b_i = 모든 교정데이타의 i번째 켑스트럼의 평균
- 학습데이타의 i번째 켑스트럼의 평균 (10)

으로 둔다. 교정데이타를 변환망에 입력하여 MLP의 출력노드 값을 구한다. MLP 목표값은 자기 음소에 대응하는 출력노드에서는 1로, 다른 출력노드에서는 0으로 둔다. MLP의 출력값과 목표값 간의 오차로 부터 역전파 알고리듬[19]을 이용하여 P와 B만 수정한다. 이때 MLP의 가중치(weight)는 변화시키지 않는다. 이와 같은 과정은 MLP의 출력의 평균 오차가 더 이상 감소되지 않을 때까지 반복 수행한다. 인식과정에서는 신경망의 출력값이 가장 큰 노드에 해당되는 음소를 인식된 음소로 판정한다.

Ⅳ. 화자적응 단어인식

이 장에서는 다층신경망을 이용한 화자적응 방법과 인식 시스템에 대해 기술한다.

1. 화자적응 과정

화자적응 과정의 블럭도는 그림 4와 같다. 지도적응에서는 교정음성의 내용을 시스템에게 알려주어야 한다.

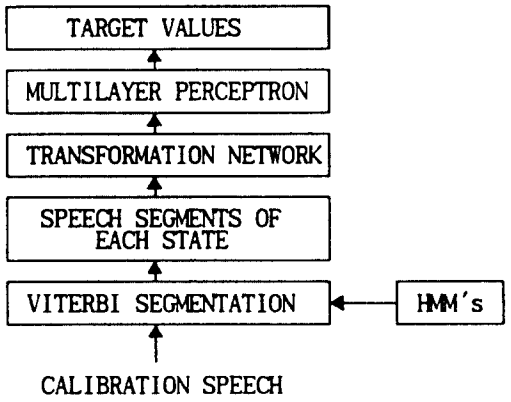


그림 4. 화자적응 과정의 블럭도
Fig. 4. Block diagram of speaker adaptation process.

새로운 화자의 교정음성은 이미 만들어 둔 HMM과 Viterbi 알고리듬[14]에 의해 상태별로 분할된다. 각 상태의 음성세그먼트에 대해 HMM의 관측 확률을 최대로 하기 위해 다음과 같이 변환망의 파라미터를 조정한다. 먼저, 상태 j에 속한 음성 세그먼트를 프레임

단위로 변환망에 입력하고 다층신경망의 출력 O_{tm}을 구한다. 출력값과 목표값 t_m의 오차 E_t를 계산한다. 즉,

$$E_t = \frac{1}{2}\sum_{m=1}^{M}(O_{tm} - t_m)^2 \qquad (11)$$

이다. 또한, 음성 세그먼트의 전체오차는

$$E = \sum_{t=0}^{T} E_t \qquad (12)$$

로 구해지며, T는 음성세그먼트의 길이이다. 오차 E를 MLP를 통해 변환망으로 역전파하고 최급하강법(gradient descent method)으로 P와 B의 값은 구한다. 즉, p_{il}과 b_i의 증분변화량은

$$\triangle p_{il} = -\eta_p \frac{\partial E}{\partial p_{il}} \qquad (13)$$

$$\triangle b_i = -\eta_b \frac{\partial E}{\partial b_i} \qquad (14)$$

으로 계산되며, 여기서 η_p와 η_b는 상수이다. 화자적응 과정은 다음과 같은 순서로 수행된다.

1) P는 단위행렬로 초기화하고, B는 식(10)으로 초기화한다.
2) 모든 교정데이터를 FHMM과 Viterbi 알고리듬을 이용하여 상태별로 분할한다.
3) 각 상태별로
   a) 음성세그먼트의 특징벡터열 X를 변환망에 입력하여 다층퍼셉트론의 출력값을 구한다.
   b) 출력값과 목표값 간의 오차를 구하고, 이를 맨 아래층의 변환망까지 역으로 전파한다.
   c) P와 B의 값를 수정한다. 이때 다층신경망의 가중치는 변하지 않는다.
4) 모든 교정데이터와 모든 상태에 대해 3)의 과정을 수행한다.
5) 변환망이 수렴할 때까지 3)에서 4)까지의 과정을 반복한다.

이와 같은 적응과정은 상태분할의 방법에 따라 두가지 방법으로 수행할 수 있다.

[방법 1] 초기화된 P와 B의 값과 HMM을 이용하여 모든 교정데이터를 상태별로 분할하고, 각 상태별 음성세그먼트에서 변환망의 P와 B를 반복 수정하여 적응시킨다.

[방법 2] 먼저, 방법 1과 같이 모든 교정데이터를 상태별로 분할하고, 각 상태별 음성세그먼트에서 변환망의 P와 B를 학습한다. 학습된 P와 B에 근거하여 모든 교정데이터를 다시 상태별로 분할하고 변환망의 P와 B를 학습하는 과정을 반복하여 적응시킨다. 이 방법은 변환망의 파라미터를 미소하게 조정하면서 이에 따라 상태분할도 병행해주는 방법이다.

위의 적응과정이 끝나면 인식과정이 시작되며, 입력음성은 적응과정에서 학습된 P와 B의 값에 의해 변환되어 MLP에 입력된다.

2. 목표값의 추정

각 상태의 목표값 T_m은 다음과 같이 구할 수 있다. 어떤 HMM λ의 상태 j에 속한 음성세그먼트 $X=(X_1, X_2, \cdots, X_T)$(T는 음성세그먼트의 길이)는 변환망과 MLP를 거쳐 퍼지 관측열 $O=(O_1, O_2, \cdots, O_T)$로 변환된다. 또한 식(3)으로 정규화되므로 식(4)의 조건을 만족한다. HMM λ의 상태 j에서 $\boldsymbol{O}$의 관측확률 $P(\boldsymbol{O} \mid s_j = j, \boldsymbol{\lambda})$을 최대로 하기 위한 $\boldsymbol{O}$의 값은 다음과 같은 목적함수 J로부터 추정한다. 목적함수

$$J = \sum_{t=1}^{T} \log P(O_t \mid s_j = j, \lambda) + L\left(\sum_{m=1}^{M} O_{tm}\right) \sum_{t=1}^{T} \log\left(\sum_{m=1}^{M} b_{jm} O_{tm}\right) + L\left(\sum_{m=1}^{M} O_{tm}\right) \qquad (15)$$

이며, 여기서 L은 Lagrange 승수(multiplier)이다. J를 O_{tm}에 대해 편미분을 취하면

$$\frac{\partial J}{\partial O_{tm}} = \frac{b_{jm}}{\sum_{m=1}^{M} b_{jm} O_{tm}} + L \qquad (16)$$

이 된다. $\partial J / \partial O_{tm} = 0$ 으로 두고 양변에 O_{tm}을 곱한 다음, 모든 m에 대해 더하면 L=-1 이 된다. 관측확률 $P(O \mid s_j = j, \lambda)$을 최대로 하기 위한 O_{tm}의 값은

$$1 = \frac{b_{jm}}{\sum_{m=1}^{M} b_{jm} O_{tm}} \qquad (17)$$

을 만족시킨다. 이 식은 O_{tm}에 대한 비선형 방정식이므로 다음과 같이 반복적인 방법으로 구한다.

$$O_{tm'}^{(n+1)} = \frac{b_{jm} O_{tm'}^{(n)}}{\sum_{m=1}^{M} b_{jm} O_{tm'}^{(n)}} \quad (18)$$

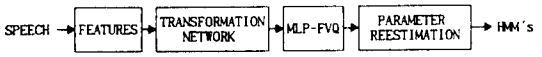
$O_{tm'}$은 시간 t에 무관하고 b_{jm}에만 의존되는 상수이다. 다층신경망의 목표값 T_m은 상태 j에서 O의 확률을 최대로 하는 $O_{tm'}$과 같게 둔다. 즉,

$$T_m = O_{tm'} \quad (19)$$

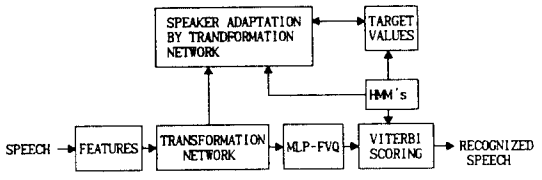
이 된다. 화자적응시 변환망은 MLP의 출력값이 목표값에 접근하도록 적응하므로 HMM의 관측확률은 증가하게 된다.

3. 지도적응(supervised adaptation) 단어 인식 시스템

인식시스템의 전체적인 구성은 그림 5와 같다. 새로운 화자는 먼저 지정된 교정데이터를 입력하여 시스템을 적응시키고, 이 과정이 완료되면 인식과정이 시작된다.



(a)



(b)

그림 5. 화자적응 인식 시스템의 구성도
(a) 학습시스템 (b) 인식시스템

Fig. 5. Block diagram of speaker adaptive recognition system. (a) Training system, (b) Recognition system.

입력음성은 적응과정에서 학습된 **P**와 **B**의 값에 의해 변환되어 다층신경망에 입력된다. 이와 같은 화자적응 방법은 HMM과 다층신경망의 파라미터를 수정하지 않고, 변환망의 파라미터를 수정하여 기존의 HMM과 다층신경망에 적응시키는 방법이다. 화자적응에 사용되는 교정데이터의 수는 극히 제한적이므로 이들로부터 파라미터를 추정하는 것은 매우 어렵다. 본 논문에서는 교정데이터로부터 변환망의 파라미터만 수정하도록 하여 제한된 데이터를 좀 더 효율적으로 사용하고, 적응에 소요되는 시간도 줄였다.

4. 적응 정도

화자적응에서 적응의 정도는 매우 중요하다. 일반적으로 교정데이터의 수는 학습 데이타에 비해 매우 작으므로 교정데이타만으로는 새로운 화자의 음성특징을 충분히 표현할 수 없다. 그러므로 인식 시스템이 교정데이타에 충분히 적응하더라도 추정된 파라미터는 신뢰도가 낮게 되고, 경우에 따라서는 잘못된 교정데이터로 인해 인식성능이 오히려 악화될 수도 있다. 본 논문에서는 이러한 단점을 해결하는 한 방안으로 새로운 화자에 적응된 파라미터 B와 P는

$$\mathbf{B} = \mathbf{B}_0 + \alpha(\mathbf{B}_a - \mathbf{B}_0) \quad (20)$$

$$\mathbf{P} = \mathbf{P}_0 + \alpha(\mathbf{P}_a - \mathbf{P}_0) \quad (21)$$

으로 두었다. 여기서 $\mathbf{B}_0$와 $\mathbf{P}_0$는 적응전의 초기값이고, $\mathbf{P}_a$와 $\mathbf{B}_a$는 변환망이 완전히 적응한 후의 값이다. 또, α는 적응정도의 척도로서, 0에서 1사이의 값이며,
$\alpha=0$은 적응을 시키지 않는 경우를, $\alpha=1$은 완전히 적응시킨 경우를 각각 나타낸다.

5. 무지도적응(unsupervised adaptation) 방법

지도적응 인식기에서는 시스템을 새로운 화자에 적응시킬때 교정데이타의 내용을 시스템에 알려주어야 하는 불편이 따른다. 무지도적응은 교정데이타의 내용을 일일이 시스템에 알려주지 않더라도 시스템을 적응시킬 수 있는 방법이다. 무지도적응은 임의로 발음된 교정데이타의 내용을 먼저 추정한 다음, 이로부터 지도적응 과정을 통해 새로운 화자에 적응한다. 교정데이타의 내용은 다음과 같이 추정한다. 먼저, V개의 단어 W_i로 구성된 어휘에 대해, 각 HMM λ_i에서 교정데이타 O의 관측확률 $P(O \mid \lambda_i)$을 구한다. HMM 인식기에서는 관측확률이 가장 높은 HMM λ_M에 해당하는 단어 WM을 인식된 단어로 판정한다. 교정데이타 O의 단어가 WM과 같을 확률을 확신도 (confidence)라 하면 이의 평균값은 HMM 인식기의 인식률과 같아 진다. 무지도적응은 확신도에 근거하여 교정데이타의 단어를 추정하며, 확신도의 척도 O는 다음과 같이 정의한다. 이 식은 HMM λM에서 O의 관측확률 $P(O \mid \lambda_i)$과 모든 HMM λ_i에서 O의 관측확률의 비로 정의된다.

여기서 η는 각 관측확률의 상대적인 크기를 조정하기 위한 상수

$$C = \frac{P(O|\lambda_M)^{\eta}}{\sum_{i=1}^{V} P(O|\lambda_i)^{\eta}} \qquad (22)$$

이다. HMM에서 O의 관측확률은 다층신경망의 변별력 있는 특성으로 인해 큰 차이를 보인다. η는 이러한 차이를 줄이기 위한 상수로 여기서는 $\eta=1/3$으로 둔다. 무지도적응에서는 교정데이타의 확신도 C를 구한 다음, 이 값이 지정된 문턱값 C_{THR}보다 큰 데이터는 적응에 사용하고 그렇지 않은 데이터는 제거시킨다. 이후의 적응과정 및 인식과정은 지도적응의 경우와 동일하다.

V. 실험결과및 검토

인식실험은 9명의 남성화자가 표 1에서 보인 26개의 단어음을 각각 20회씩 발음한 단어를 사용하였다. 인식시스템의 성능평가를 위해 모든 음성데이타의 단어음에 포함된 음소에 대하여 신경망을 학습하고, 음소인식 실험을 행하여 인식률을 확인하였다. 또, 학습된 다층신경망과 HMM을 이용하여 화자적응 인식실험을 수행하여 기존의 화자독립 HMM을 기반으로 하는 기본 인식시스템의 결과와 비교하였다.

표 1. 음성 데이타
Table 1. Speech data.

단어명	일	이	삼	사	오	육	칠	팔	구	
	공	영	서울	부산	대구	대전	인천	광주	강릉	
	청주	전주	제주	김천	성남	경주	포항	춘천		
음소	ㅏ	ㅓ	ㅐ	ㅣ	ㅡ	ㅗ	ㅜ			
	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅇ	ㅂ	ㅈ	ㅊ	ㅅ	ㅍ	ㅎ

음성데이타는 차단 주파수가 5.2㎑인 저역통과 필터를 거친 음성신호를 12㎑로 샘플링하고, $1-0.97z^{-1}$인 필터로 고주파 성분을 강조하도록 전처리하여 추출하였다. 각 프레임의 길이는 20msec이고 10msec의 겹침구간을 정하여 해밍윈도우 함수를 이용하였으며, 특징벡터로는 16차 LPC켑스트럼과 로그에너지를 사용하였다. HMM은 단어별 하나씩의 기준 HMM을 만들었으며, 이때 HMM의 상태수를 5로 하였다.

1. 다층신경망의 학습 및 음소인식

숫자음과 도시명으로 구성된 26개의 고립단어음에 내포된 19개의 음소에 대해 다층신경망을 학습하고, 이를 이용하여 음소인식을 수행하였다. 표 2는 9명의 화자중 4명이 19음소를 20회씩 발성한 1520개의 음소데이타에 의해 학습된 신경망에서 다른 5명의 화자가 20회씩 발성한 1900개의 음소를 대상으로 수행한 음소인식의 결과를 나타낸다. 적응인식은 인식화자의 음소데이타중 5명×19음소×5회=475개의 교정데이타를 이용하여 변환망의 파라미터인 B와 P행렬의 적응여부에 따라 3가지로 나누어 5명×19음소×15회=1425개 음소에 대해 수행했다. 변환망은 다층퍼셉트론의 출력오차가 더 이상 감소되지 않을 때까지 반복적으로 적응시켰다. 표 2에서 MLP인식은 B와 P를 초기값으로 두고 MLP만으로 인식한 결과이다. 적응인식에서 MLP/B는 B만 적응시키고 P는 초기값으로 둔 경우이며, MLP/P는 B는 초기화 값으로 고정하고 P만 적응시킨 경우의 오인식률이다. 그리고 MLP/B+P는 B와 P를 모두 고려한 경우의 오인식률이다.

표 2. MLP 음소 인식기에서 여러 적응방법에 대한 오인식률(%) 비교
Table 2. Comparison of error rates(%) for various adaptation method from MLP phone classifier.

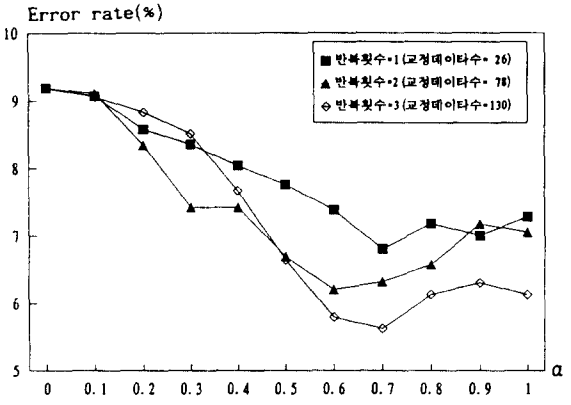
인식방법 \ 반복횟수		1	2	3	4	5	평균
MLP 인식		16.3					
적응인식	MLP/B	12.6	12.4	12.2	11.1	11.5	12.0
	MLP/P	11.3	10.7	11.1	10.5	10.7	10.9
	MLP/B+P	11.0	10.6	10.2	10.3	10.3	10.5

음소인식의 실험결과는 교정데이타의 수에 따라 적응인식의 오인식률이 전반적으로 감소되고 있음을 알 수 있다. 화자독립 인식(MLP인식)과 비교하면, 적응인식은 오인식률을 각각 MLP/B에서 평균 26.4%, MLP/P에서 평균 33.1%, 이들을 결합한 MLP/B+P에서 평균 35.6%씩 감소시킴을 알 수 있다. MLP/B+P는 MLP/B와 MLP/P보다 인식률을 약간 개선시킬 수 있음을 알 수 있다. 또, 각 음소별 교정데이타가 불과 2개씩만 있는 경우 즉, 반복횟수가 2일 때 인식률은 크게 개선됨을 알 수 있다.

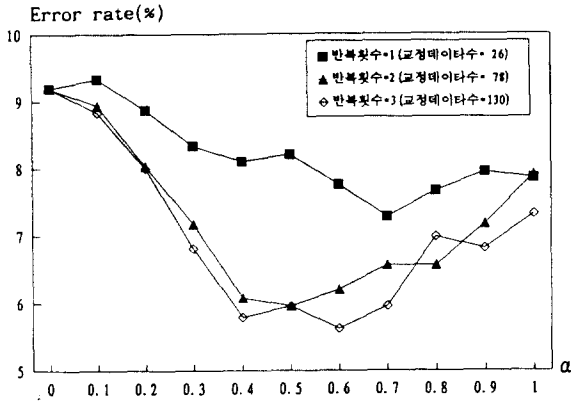
2. 지도적응에 의한 고립단어 인식실험

고립단어음의 화자적응 인식실험은 9명의 화자가 20회씩 발음하여 얻어진 9명×26단어×20회=4680개 단어음의 음성데이타를 사용하여 수행하였다. 이중 4명의 화자에 대한 음성데이타 4명×26단어×20회=2080개는 HMM의 학습데이타로 사용하였고, 나머지 5명의 화자에 대한 5명×26단어×20회=2600개의 단어음은 인식에 사용하였다. 화자적응을 위한 교정데이타는 인식에 사용된 데이타중 5명×26단어×5회=650개의 단

어음을 이용하고, 나머지 1950개의 단어음은 인식실험에 사용하였다.



(a)



(b) 방법2

그림 6. 적응정도에 따른 지도적응 단어인식기의 오인식률(%)
(a) 방법 1 (b) 방법 2

Fig. 6. Error rates(%) of word recognizer using supervised adaptation as a function of adaptation degree.
(a) Method 1, (b) Method 2.

그림 6은 적응정도 α에 따른 인식률을 보인 것이다. 여기서 변환망의 파라미터 **B**와 **P**는 모두 새로운 화자에 적응하도록 했고, 이들의 수렴여부는 식(12)로 주어지는 신경망출력의 평균오차값으로 판정했다. 매 반복적인 적응과정에 따른 오차값의 변화분이 0.0005의 이하이면 수렴된 것으로 간주했다. 그림에서 α가 0인 경우는 화자적응을 하지 않은 기본 인식시스템의 인식률과 같은 결과이고, α가 1인 경우는 새로운 화자의 교정데이타에 완전히 적응했을 때의 인식률을 나타낸다. 대체적으로 $\alpha=0.5$에서 $\alpha=0.7$ 범위의 값에서 가장 좋은 인식성능을 보임을 알 수 있다. $\alpha<0.3$인 경우는 새로운 화자에 대한 적응정도가 낮으므로 인식률이 저하되고 있고, $\alpha>0.8$인 경우는 변환망의 파라미터가 과도하게 적응하여 인식률을 떨어뜨리는 것으로 생각된다. 방법 1과 방법 2는 비슷한 인식결과를 보이고 있으나 방법 2가 방법 1에 비해 더 빨리 적응하였다. 이와 같은 적응방법은 변환망의 파라미터만을 조작하여 신경망출력의 오차를 감소시켜 HMM의 관측확률을 높일 수 있기 때문에 적응에 소요되는 시간을 크게 줄일 수 있었다.

표 3. 여러 지도적응 방법에 대한 오인식률(%) 비교

Table 3. Comparisons of error rates(%) for various supervised adaptation methods.

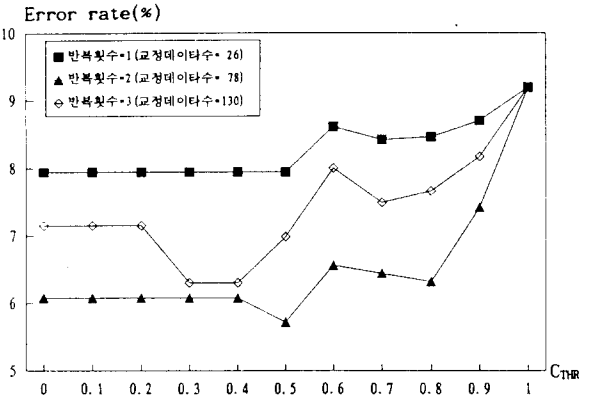
인식방법 \ 반복횟수		1	2	3	4	5	평균
화자종속 인식기		2.2					
화자독립 인식기		2.2					
방법 1	HMM/B	8.7	8.2	8.1	8.1	8.1	8.2
	HMM/P	7.8	7.0	6.8	7.1	6.3	7.0
	HMM/B+P	7.4	6.5	6.2	6.1	5.8	6.4
방법 2	HMM/B	8.7	8.1	8.1	8.0	7.9	8.2
	HMM/P	8.9	6.3	6.7	6.4	5.8	6.8
	HMM/B+P	7.8	6.5	6.2	6.8	5.6	6.6

표 3은 α를 0.6으로 하여 수행한 인식결과로서, 반복횟수는 화자적응을 위해 각 단어별로 반복 발음한 횟수를 의미한다. 여기서 화자독립 인식기는 **B**와 **P**를 초기값으로 두고 HMM만으로 인식한 결과이고, HMM/B는 **B**만 적응시키고 **P**는 초기값으로 둔 경우, HMM/P는 **B**는 초기화 값으로 고정하고 **P**만 적응시킨 경우, 그리고 HMM/B+P는 **B**와 **P**를 모두 고려한 경우의 오인식률이다. 방법 1인 경우, HMM/B는 기존 인식기의 오인식률에 비해 5.4%에서 12.0%까지, HMM/P는 15.2%에서 31.5%까지, HMM/B+P는 19.6%에서 37.0%까지 각각 개선시킬 수 있었다. 방법 2인 경우, HMM/B는 5.4%에서 14.1%까지, HMM/P는 3.2%에서 37.0%까지, HMM/B+P는 15.2%에서 39.1%까지 각각 개선시킬 수 있었다. 화자종속 인식기의 오인식률 2.2%에 비교하면 아직도 더 개선해야 할 부분이 남아 있다. 제안한 적응인식 방법은 방법 1과 방법 2에서 거의 비슷한 인식결과를 보

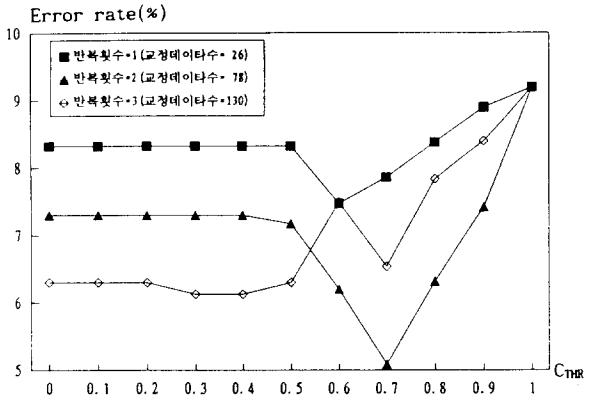
였으며, 각 방법에서 교정데이타를 화자별 단어음당 3회 반복으로 했을 때에 오인식률을 1/3정도 감소시킴을 알 수 있다. 또, HMM/B+P의 경우가 B 또는 P만 적응시킨 경우에 비해 개선된 인식률을 보이고 있다.

3. 무지도적응에 의한 고립단어 인식실험

그림 7은 확신도의 문턱값 C_{THR}에 따른 오인식률을 나타낸다.



(a) 방법1



(b) 방법2

그림 7. 확신도의 문턱값에 따른 무지도적응 단어인식기의 오인식률(%)
(a) 방법 1 (b) 방법 2

Fig. 7. Error rates(%) of word recognizer using unsupervised adaptation as a function of confidence threshold.
(a) Method 1, (b) Method 2.

그림 7(a)의 방법 1에서는 C_{THR}이 작을수록 오인식이 감소되고 있으며, 특히, C_{THR}의 값이 0.4 이하인 경우는 적응에서 제거되는 교정데이타가 거의 없으므로 일정한 값을 나타내고 있다. 그림 7(b)의 방법 2에서는 C_{THR}이 0.7 부근에서 가장 낮은 오인식률을 보이고 있으며, 방법 1에 비해 대체로 개선된 성능을 보인다. 반복횟수가 2와 3인 경우, C_{THR}>0.8일때는 교정데이타의 1/3이상이 제외되므로 교정데이타의 부족으로 인해 오인식률이 증가되고, C_{THR}<0.6일때는 제외되는 교정데이타가 거의 없으므로 잘못된 교정데이타로 인해 오인식률이 증가하는 것으로 생각된다. 반복횟수가 5인 경우는 대체로 C_{THR}이 작을 수록 오인식률이 감소되고 있다.

표 4은 α를 0.6으로 두고 수행한 무지도적응 인식의 결과를 나타낸다. 방법 1은 C_{THR}=0.0으로 두고, 반복횟수에 따른 오인식률의 결과로서, 반복횟수가 3인 경우의 오인식률 6.0%는 지도적응의 최저 오인식률과 거의 같은 결과를 보인다. 방법 2는 C_{THR}=0.7인 경우의 결과이며, 특히 반복횟수가 3일 때의 오인식률 5.1%는 지도적응의 최저 오인식률보다 더 우수한 결과를 보인다. 지도적응에서는 잘못된 교정데이타가 있어도 이를 모두 사용하여 적응하므로 오인식률이 높아질 수 있다. 반면, 무지도적응에서는 확신도에 의해 잘못된 데이타를 제거하므로 비록 교정데이타의 수는 약간 줄지만 새로운 화자의 올바른 데이타만 사용하게 되므로 인식성능이 좋아지는 것으로 생각된다.

표 4. 여러 무지도적응 방법에 대한 오인식률(%) 비교

Table 4. Comparisons of error rates(%) for various unsupervised adaptation methods.

반복횟수 / 인식방법	1	2	3	4	5	평균
방법 1 (HMM/B+P)	7.9	7.3	6.0	6.1	7.1	6.9
방법 2 (HMM/B+P)	7.8	6.6	5.1	6.7	6.5	6.5

VI. 결 론

본 논문에서는 다층퍼셉트론을 FVQ로 사용하는 준연속 HMM 음성 인식기의 인식성능을 향상시키기 위하여 화자적응 방법에 의한 단어인식 시스템을 구성하였다. 이 시스템의 적응방법은 1차 선형변환망을 이용하여 음성신호를 주변환경에 적응하게 하고, 주파수변환을 행하여 새로운 화자의 스펙트럼을 적응하도록 하

였다. 화자적응은 변환망의 파라미터를 조정하여 신경망 출력노드의 오차를 감소시켜 HMM의 관측확률을 높인다. 변환망의 파라미터는 다층신경망의 출력오차를 역전파하여 조정하므로 적응속도는 매우 빨라진다. 또한, 다층신경망의 전단에서 파라미터를 변경함으로써 전체적 HMM의 파라미터를 조정해주기 때문에 교정데이타의 수가 제한적일 경우에도 성능 개선을 기대할 수 있다. 제안한 방법의 인식성능을 입증하기 위해 화자독립 단어인식 실험을 수행하여 인식률이 향상됨을 보였다. 5명의 화자가 발성한 26개의 단어음을 대상으로 수행한 인식실험 결과, 지도적응일 경우의 오인식률은 기본시스템의 9.2%에서 화자적응 후의 5.6%로 크게 감소되어 오인식 감소률은 39.1%였다. 또한, 무지도적응일 경우의 오인식률은 5.1%로서 기본 인식기의 오인식률을 44.6% 줄였다.

참 고 문 헌

[1] L.R.Rabiner and B.H.Juang, *Fundamentals of speech recognition*, Prentice Hall International, 1993.

[2] R.L.Watrous, "Source decomposition of acoustic variability in a modular connectionist network," in *Proc. ICASSP*, pp.129-131, 1991.

[3] 이종석, 이상욱, "신경망과 구문분석을 이용한 한국어 연결 숫자음 인식," 전자공학회 논문지-B, 제 30권, 제 12호, pp.21-30,1993년 12월

[4] X.D.Huang and K.F.Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," in *Proc. ICASSP*, pp.877-880, 1991.

[5] W.A.Rozzi and R.M.Stern, "Speaker adaptation in continuous speech recognition via estimation of correleted mean vectors," in *Proc. ICASSP*, pp.865-868, 1991.

[6] Y.Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," IEEE Trans. *Speech Audio Processing*, vol.2, no.3, pp. 380-394, July 1994.

[7] B.F.Necioglu, M.Ostendorf,and J.R.Rohlicek, "A bayesian approach to speaker adaptation for the stochastic segment model," in *Proc. ICASSP*, pp.437-440, 1992.

[8] H.Matsukoto and H.Inoue, "A piecewise linear spectral mapping for supervised speaker adaptation," in *Proc. ICASSP*, pp. 449-452, 1992.

[9] O.Schmidbauer and J.Tebelskis, "An LVQ based reference model for speaker-adaptive speech recognition," in *Proc. ICASSP*, pp.441-444, 1991.

[10] R.M.Stern, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. ASSP*, vol. ASSP-35, pp.751-763, Jun.1987.

[11] S.Furui, "Unsupervised speaker adaptation based on hierarchical spectral clustering," *IEEE Trans. ASSP*, vol.37, no.12, pp.1923-1930, Dec.1989.

[12] S.Renals, N.Morgan, H.Bourlard, M. Cohen, and H.Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans. Speech* Audio Processing, vol.2, no.1, part II, pp. 161-173, Jan.1994.

[13] H.P.Tseng, M.Sabin, and E.Lee, "Fuzzy vector quantization applied to hidden Markov modeling," in *Proc. ICASSP*, pp. 641-644, 1987.

[14] P.L.Cerf, W.Ma, and D.V.Compernolle, "Multilayer perceptrons as lablers for hidden Markov models," *IEEE Trans. Speech* Audio Processing, vol.2, no.1, part II, pp.185-193, Jan.1994.

[15] L.E.Baum, T.Petrie, G.Soules, and N.Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol.41, no.1, pp.164-171, 1970.

[16] X.D.Huang, "Phoneme classification using semicontinuous hidden Markov models," *IEEE Trans. Signal* Processing, vol.40, no.5, pp.1062-1067, May 1992.

[17] B.H.Jung and L.R.Rabiner, "Mixture autoregressive hidden Markov models

for speech signals," *IEEE Trans. ASSP*, vol.ASSP-33, pp.1404-1413, Dec.1985.

[18] M.J.Russel and R.K.Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. ICASSP*, pp.5-8, 1985.

[19] Y.H.Pao, *Adaptive pattern recognition and neural networks*, Addison-Wesley, 1989.

저 자 소 개

李 基 熙(正會員)

1957年 2月 12日 生. 1984年 3月 서울산업대학 전자공학과(공학사). 1984年 3月 ~ 1986年 8월 한양대학교 산업대학원 전자공학과(공학석사). 1988年 3月 ~ 현재 한양대학교 대학원 전자공학과(박사과정). 1992年 3月 ~ 현재 대유공업전문대학 사무자동화과 전임강사. 주관심 분야는 음성인식, 영상처리 등임

林 寅 七(正會員) **第 30卷 B編 第 2號 參照**

현재 한양대학교 전자공학과 교수