論文95-32B-5-11

제한적 상태지속시간을 갖는 HMM을 이용한 고립단어 인식

(Isolated Word Recognition Using Hidden Markov Models with Bounded State Duration)

李 基 熙 *, 林 寅 七 *

(Kee Hee Lee, and In Chil Lim)

요 약

본 논문에서는 고립단어 음성인식을 위하여 MLP를 기반으로한 제한적 상태지속시간을 갖는 HMM을 제안한다. HMM의 각 상태에서 상태지속시간의 최소 및 최대시간은 훈련단계 동안 추정되고, 인식 단계에서는 상태의 천이를 제한하는 파라미터로서 사용된다. 또한, 이들 파라미터를 추정하는 과정과 제안된 HMM을 이용한 인식 알고리듬도 기술된다. 10개의 도시명과 11개의 숫자음으로 구성된 어휘를 사용하여 수행한 화자독립 고립단어 인식실험은 최소 상태지속시간을 조정함으로써 인식률이 개선됨을 보인다.

Abstract

In this paper, we proposed MLP(MultiLayer Perceptron) based HMM's(Hidden Markov Models) with bounded state duration for isolated word recognition. The minimum and maximum state duration for each state of a HMM are estimated during the training phase and used as parameters of constraining state transition in a recognition phase. The procedure for estimating these parameters and the recognition algorithm using the proposed HMM's are also described. Speaker independent isolated word recognition experiments using a vocabulary of 10 city names and 11 digits indicate that recognition rate can be improved by adjusting the minimum state durations.

I. 서 론

음성 인식기를 구현하려는 연구는 그동안 많은 발전을 하여 현재 수십에서 수백단어의 어휘에 대해 신뢰성 있는 인식을 할 수 있는 단계에 있으며, 일부는 이미 상용화 되고 있다. 이러한 음성인식에 가장 널리 이용되고 있는 알고리듬으로는 DTW(Dynamic Time Warping)[11], 신경망에 의한 방법[6,8,11] 그리고 HMM(Hidden Markov Model)[2,3,4]을 이용한 방법을 들 수 있다.

음성을 시간적으로 매칭하는 DTW는 이미 저장되어 있는 기준패턴과 미지의 음성신호를 비선형으로 정합하여 가장 유사도(likelihood)가 높은 기준패턴의 음성을 인식하는 방법으로 기준 패턴의 준비가 쉽고, 높은 인식률을 얻을 수는 있지만 화자독립 인식이나 대규모 어휘인식으로의 확장이 어려운 단점이 있다. 신경망을 이용한 방법은 대단위 병렬성과 학습능력을 이용해서 음성을 인식하게 되며, 가장 최근에 시작된 방법이다. 신경망은 인간의 뇌세포에 해당하는 처리요소들이 신경에 대응하는 정보 채널을 통해 연결된 망으로 처리요소들이 병렬로 구성되어 있어서 데이터를 분산처리할 수 있고, 학습을 통해 음성의 특징을 찾아서 적응하는 능력이 있어 음성인식에 많이 이용되고 있다.

HMM은 Markov chain의 확률 함수를 이용하여

* 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: 1995년4월20일

음성을 통계적으로 분석, 학습하고, 이를 이용해서 인식하는 방법이다. 이 방법은 음성신호의 관측열과 모델의 상태열을 분리시킴으로써 보다 효율적으로 동적 프로그래밍을 수행하도록 개선한 인식 방법으로 대용량의 고속 음성인식 시스템에 적합한 방식이다. 이 방법에서는 각 기준 음성에 대한 HMM을 미리 저장한 후에 이들 모델로부터 입력음성의 관측확률(observation probability)을 구하여 가장 높은 확률을 가지는 모델의 음성으로 인식한다. 또한, HMM은 단어단위뿐만 아니라 음절, 반음절 또는 음소단위의 음성도 모델링이 가능하다. 그러나 HMM은 음성 세그먼트의 발생시간을 나타내는 상태의 지속시간 정보를 충실히 처리하지 못하는 문제점을 보이고 있다. 이에 이러한 문제점을 고려하여 인식률을 개선하려는 연구가 이루어져 왔는데[4,9,10], 이들 방법은 지속시간 정보를 효과적으로 이용할 수 있지만 지속시간의 확률 분포만으로는 상태천이를 강력히 억제할 수 없어 인식률이 저하되고 있다.

본 논문에서는 위와 같은 문제점을 고려하여 HMM의 상태천이를 제한하도록 하여 인식률을 향상시킬 수 있는 인식시스템을 구성하였다. 이 시스템에서는 MLP(MultiLayer Perceptron)를 확률추정기(probability estimator)로서 사용하고, 이의 출력으로 부터 준연속 HMM(SCHMM : Semi-Continuous HMM)를 구성하였다. 또 HMM의 각 상태에서 지속시간을 연속 확률함수로 근사화하고 최소 및 최대 지속시간의 경계치를 추정하여 HMM의 상태지속시간이 최소 및 최대 지속시간 범위에 속하지 않을 경우에 다른 상태로의 천이를 제한한다. 제안한 방법의 타당성을 입증하기 위하여 숫자와 도시명으로 구성된 21개의 어휘를 대상으로 HMM을 구성하여 화자독립 인식실험을 수행하였다. 또한, 실험을 통해 인식 음과 다른 음 사이에는 서로 다른 상태지속시간 확률분포를 이루고 있음을 확인하고, 각 HMM 상태의 최소 및 최대 지속시간을 설정하여 상태천이를 제한함으로써 인식률이 개선됨을 보였다.

II. MLP와 HMM을 이용한 음성인식

이 장에서는 MLP-FVQ와 인식 시스템에 대해 기술한다.

1. MLP 확률 추정기

MLP을 이용한 음성인식기에서 MLP는 VQ(Vector Quantizer)[8] 또는 확률추정기[7]로서 사용된다. MLP를 VQ로 사용하는 방법에서는 프레임 단위의 음성신호를 MLP의 입력층에 넣고 출력층에서 하나의 심볼을 얻는다. 각 프레임별로 구한 심볼열로 부터 이산 HMM(discrete HMM)이 구성된다. 이와 같은 방법은 인식시스템이 간단한 반면 훈련 데이타의 수가 많아야 하고 양자화 오류로 인해 인식률이 저하되는 단점을 가지고 있다. 한편, MLP를 확률추정기로 사용하는 방법은 MLP의 출력값을 관측확률의 추정값으로 보고 이로부터 HMM을 구성한다. 이 방법은 MLP의 학습이 어렵고, HMM의 모든 상태가 하나의 음소로 제한되는 단점이 있다. 이러한 단점을 감소시키는 방법으로 FVQ(Fuzzy VQ)[13]가 있으며, 입력벡터는 각 코드워드에 정합되는 정도를 나타내는 퍼지 score를 발생시킨다. 또한, MLP를 FVQ로 사용하는 MLP-FVQ[14]는 HMM의 최적화기법과 함께 HMM과 밀접하게 결합된다. MLP를 확률추정기로 사용하는 음성인식방법은 Bourlard와 Wellekens[15]의 연구에 기초한다. 이들은 분류인식을 하도록 학습된 MLP는 class 조건부 사후확률 추정기(class-conditional posterior probability estimator)가 됨을 증명하였다. 즉, 주어진 입력 X에 대한 MLP의 출력값은 class g_k 의 사후확률 $P(g_k|X)$의 추정값이 된다. 본 논문에서도 MLP를 FVQ로서 사용하였다.

2. 인식 시스템

본 논문에 사용된 인식 시스템의 기본적인 구조는 그림 1과 같다.

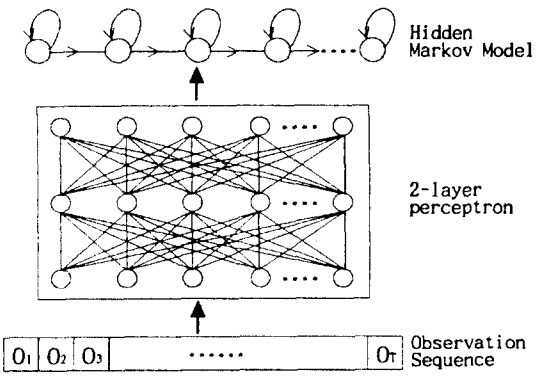


그림 1. 인식 시스템의 기본구조
Fig. 1. Basic structure of recognition system.

인식에 사용된 HMM은 좌-우(left-to-right)모델로서, 이의 Markov chain은 초기상태 확률벡터(initial state probability vector)

$U=(1,0,0,\cdots\cdots,0)^T$ 와 상태 천이행렬(state transition matrix) $A=[a_{ij}]$로 표현된다. 여기서 a_{ij}는 현재의 상태 i에서 다음의 상태 j로 천이할 확률로서

$$\sum_{j=1}^{N} a_{ij} = 1, \quad \text{for } 1 \le i \le N \tag{1}$$

의 조건을 만족하며, N은 상태의 수를 나타낸다. 시간 t에서 음성 프레임의 특징벡터를 O_t로, 전체 프레임별 특징벡터열을 $O=(O_1, O_2, \cdots, O_T)$ (T는 프레임 수)로 각각 표기하기로 하자. MLP는 특징벡터열 O를 입력으로 하여 음소단위의 분류인식을 하도록 학습된 다층 신경망이다. 즉, M개의 음소를 분류인식하는 경우, 각 음소에 대응하는 MLP의 출력노드 M개는 각각 자기 음소에 대한 출력노드의 값을 1로, 다른 음소에 대해서는 0으로 학습된다. MLP의 출력노드는 각 음소 g_k에 대응되며, 입력 O_t에 대한 각 출력노드의 값 $f_k(O_t)$는 음소 gk의 사후확률 $P(g_k|O_t)$의 추정값이 된다.

각 상태 j에서 O_t의 관측확률밀도 함수 $b_j(O_t)$는 다음식으로 표현할 수 있다.

$$b_j(O_t) = \sum_{j=1}^{M} c_{jk} f_k(O_t), \quad \text{for } 1 \le j \le N \tag{2}$$

여기서 M은 MLP의 출력노드 수 이다. 식(2)는 기존의 HMM에서 관측확률밀도 함수 bj(·)를 MLP 출력값에 가중치 c_{jk}를 곱하여 구하므로 HMM은 준연속(semi-continuous) HMM이나 여기서는 HMM으로 표기하기로 한다. 단, c_{jk}는

$$\sum_{k=1}^{M} c_{jk} = 1, \quad \text{for } 1 \le j \le N \tag{3}$$

을 만족해야 한다. c_{jk}의 재추정식(reestimation formula)은

$$\hat{C}_{jk} = \frac{\sum_{t=1}^{T} f(O, s_t = j, h_t = k|\lambda / f(O|\lambda)}{\sum_{t=1}^{T} f(O, s_t = j|\lambda) / f(O|\lambda)} \tag{4}$$

이다[16]. 여기서 λ는 HMM이고, $\hat{C}_{jk}$는 재추정된 값이며, s_t와 h_t는 시간 t에서 O_t가 발생된 상태 및 MLP 출력 노드를 각각 나타낸다. 이와 같은 구조의 인식시스템은 MLP의 변별력을 HMM에 부여하여 인식률을 개선하고, 또한, 하나의 MLP를 모든 HMM이 공유할 수 있으므로 시스템이 간단해지는 장점이 있다.

3. 지속시간(duration) 정보

HMM은 근본적으로 상태의 지속시간 정보를 충실히 표현하지 못하는 단점을 가지고 있다. 즉, HMM의 상태 j로 천이한 후 지속시간이 τ일 확률 $D_j(\tau)$는

$$D_j(\tau) = a_{jj}^{\tau-1}(1-a_{jj}) \tag{5}$$

이므로 HMM이 표현하는 지속시간의 확률 분포는 기하분포(geometric distribution)가 됨을 알 수 있다. 그러나 실제의 음성에 대한 상태지속시간의 분포는 Gamma분포, Poisson분포 또는 음이항(negative binomial) 분포에 가까운 것으로 알려져 있다.[12] 이와 같은 HMM의 단점을 보완하기 위한 방식으로 상태의 지속시간을 고려하여 인식률을 개선하는 연구가 발표되었다.[4,9,10] 본 논문에서는 상태지속시간의 최소 및 최대치를 구하는 식을 유도하고, 인식실험을 통하여 타당성을 확인하였다.

Ⅲ. 제한적 상태지속시간을 갖는 HMM을 이용한 음성인식

1. 제한적 상태지속시간을 갖는 HMM

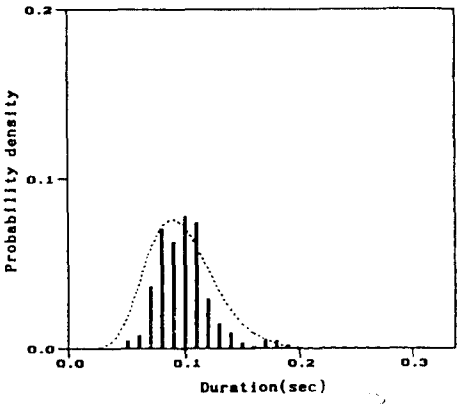
본 논문에서 제안한 상태지속시간을 갖는 HMM은 기존 HMM에서 상태의 지속시간이 미리 설정된 지속시간의 구간에서만 상태천이가 생길수 있도록 제한하는 모델이다. 상태별 지속시간의 최소 및 최대치를 지정해 두고, 인식시 최소치보다 작을 때는 다른 상태로 천이할 수 없으나 최대치 이상일 때는 무조건 다른 상태로 천이하도록 한다.

HMM에서의 상태의 지속시간은 음성 세그먼트의 발성시간을 나타내는 중요한 정보이다. 앞 절에서 언급한 몇몇 연구에서는 지속시간 정보를 효과적으로 이용할 수 있지만 지속시간의 확률분포만으로는 상태의 천이를 강력히 억제할 수 없기 때문에 유사 단어간의 인식률이 저하된다. 예를 들면, HMM "칠"에 음성 "일"이 입력된 경우, "일 "의 앞부분은 HMM "칠"의 첫번째 상태에서의 관측확률이 매우 작으므로 곧바로 두번째 상태로 천이한다. 이 경우, 다음 상태에서의 관측확률이 더 높을때, 첫번째 상태에서 지속시간이 매우 짧더라도 HMM은 상태천이를 허용하게 된다. 따라서 음성 "일 "은 HMM "칠"의 나머지 상태에서 모두 발생한 것으로 계산되어 전체적인 관측확률은 음성 "칠"이 입력된 경우의 값과 비슷하게 된다. 결과적으로 "일"과 "칠"이 서로 오인식될 수 있다. 이를 고려하여 본 논문에서는 각 상태의 지속시간을 연속 확률함수로 근사화하고, 이로부터 최소 및 최대 지속시간을 추정한 다음, 인식단계에서는 특정한 상태에서의 지속시간이 최소 및 최대 지속시간 범위에 들지 않으면 다른 상태로 천이할 수

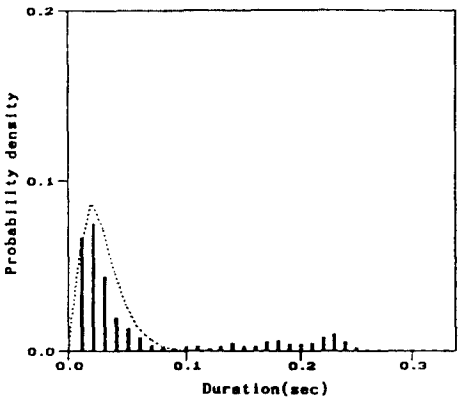
없도록 제한하였다.

2. 지속시간의 최소 및 최대치의 추정

그림 2는 5상태 좌-우 구조의 HMM "칠"에 500개의 음성 "칠"을 입력시켜 상태 1과 3의 지속시간 분포를 그린 것이며, 전체적인 분포는 대략 감마분포(그림 2의 점선)에 가까움을 알 수 있다.



(a)



(b)

그림 2. HMM '칠'의 상태 1과 3에서 측정된 지속시간의 히스토그램
(a) 상태 1 (b) 상태 3

Fig. 2. Measured durational probability distribution of state 1 and state 3 of HMM '7'.
(a) stste 1. (b) stste 3.

최소 및 최대 지속시간의 추정방법은 다음과 같다. 훈련에 사용되는 전체 단어수를 L이라 할때, ℓ번째 단어 $O^{(\ell)}$을 HMM λ에 입력하고 Viterbi 알고리듬을 이용하여 경로역추적(path backtraking)으로 음성을 상태별로 분류한다. 이때 상태 j의 지속시간을 τ_j로, 또한 j번째 상태의 관측벡터열(음성 세그먼트)을 $O^{(\ell)}{}_j = \{O^{(\ell)}{}_{t=1}, O^{(\ell)}{}_{t+2}, \cdots, O^{(\ell)}{}_{t+\tau_j}\}$로 각각 표기하기로 하자. HMM 상태 j에서 O_j에 대한 관측확률의 기하평균은

$$p^{(\ell)}{}_j = \sqrt[\tau_j]{\Pi_{d=1}^{\tau_j} b_j(O^{(\ell)}{}_{t+d})} \tag{6}$$

로 표현된다. 또, L개의 단어에 대해 상태 j에서 지속시간이 τ_j인 관측벡터열이 발생할 확률의 평균을 $p_0(\tau_j)$로, 그렇지 않을 확률을 $p_1(\tau_j)$로 각각 표현하면

$$p_0(\tau_j) = \sum_{\ell=1}^{L} p^{(\ell)}{}_j / L \tag{7a}$$

$$p_1(\tau_j) = 1 - p_0(\tau_j) \tag{7b}$$

이 된다. 상태지속시간의 최소치를 설정하기 위해서 식(7a)와 식(7b)와 같이 상태 j에서 지속시간이 τ_j 이하일 확률 $P_0\min(\tau_j)$와 지속시간이 τ_j+1 이상일 확률에 가중치 α를 곱한 값 $P_1\min(\tau_j)$를 정의한다.

$$P_{0min}(\tau_j) = \sum_{t=1}^{\tau_j} p_0(t) \tag{8a}$$

$$P_1\min(\tau_j) = \sum_{t=\tau_j+1}^{\infty} p_1(t) \cdot \alpha \tag{8b}$$

상태지속시간의 최소치 $\tau_{\min}$은 그림 3에서 보인 것과 같이 $P_0\min(\tau_j)$와 $P_1\min(\tau_j)$의 교차점으로 정한다. 또한, 상태지속시간의 최대치를 설정하기 위해 식(9a)와 식(9b)와 같이 지속시간이 τ_j+1 이상일 확률 $P_0\max(\tau_j)$와 지속시간이 τ_j 이하일 확률에 가중치 β를 곱한 값 $P_1\max(\tau_j)$를 정의한다.

$$P_0\max(\tau_j) = \sum_{t=\tau_j+1}^{\infty} p_0(t) \tag{9a}$$

$$P_0\max(\tau_j) = \sum_{t=1}^{\tau_j} p_1(t) \cdot \beta \tag{9b}$$

상태지속시간의 최대치 τmax는 그림 3에서 나타낸것처럼 P0max(τj)와 P1max(τj)의 교차점으로 정한다. 여기서 α와 β는 상수로서 인식실험을 통해서 결정하게 된다. $\tau_{\min}$과 $\tau_{\max}$는 결정된 α, β값을 이용

하여 자동적으로 계산된다.

이 방법을 이용하여 τ_{min}과 τ_{max}의 설정 예를 그림 3에서 보이고 있으며,그림은 5상태 HMM "칠"에 숫자음과 지명을 포함한 21개 단어의 음성 1000개를 입력하여 α는 0.06이고, β는 0.02일때 상태 4에서 지속시간의 확률분포에 의해 설정된 τ_{min}과 τ_{max}를 보이고 있다.

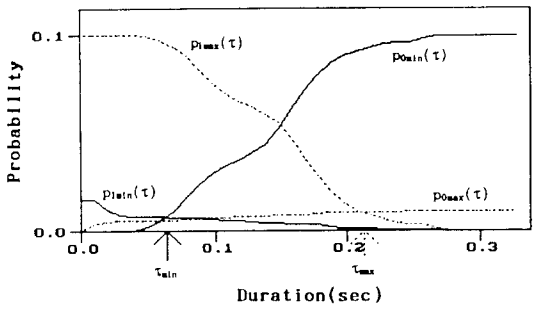


그림 3. HMM '칠'의 상태 4에서 측정된 최소 및 최대 지속시간

Fig. 3. Measured minimum and maximum duration of state 4 of HMM '7'.

한편, 최대 및 최소 지속시간과 병행하여 지속시간의 확률분포를 1)Gauss분포, 2)Gamma분포로 근사화하여 인식시에 이용하였다. 사용된 Gauss분포 함수는

$$D_j(\tau_j)=(2\pi\sigma_j^2)^{-1/2}\exp[-(\tau_j-m_j)^2/2\sigma_j^2] \qquad (10)$$

로서, mj 와 σ_j^2은 각각 상태 j에서 점유시간의 평균과 분산이며, Gamma분포함수는

$$D_j(\tau_j)=\Gamma(\nu_j)^{-1}\eta_j^{\nu_j}\tau_j^{\nu_j-1}\exp(-\eta_j\tau_j) \qquad (11)$$

이고, 여기서 $\Gamma(\cdot)$는 감마함수를 나타내며, ν_j 및 η_j는 감마 분포함수를 정의하는 파라메타이다.

3. 인식 알고리듬

제안한 모델의 음성인식 알고리듬은 다음과 같다.

[1] Initialization:

$$\delta_1(1) = \log[b_1(O_1)]$$

$$d_1(1) = 1$$

for $t=2,3,\cdots,N$

$$\delta_1(j) = -\infty$$

$$d_1(j) = 0$$

[2] Recursion:

for t=2,3,$\cdots$,N

$$\delta_t(1) = \delta_{t-1}(1) + \log[b_1(O_t)]$$

$$d_t(1) = d_{t-1}(1)+1$$

for j=2,3,$\cdots$,N

if$(\delta_{t-1}(j-1) + \log[D_{j-1}d_{t-1}(j-1)] > \delta_{t-1}(j))$

and $(d_t(j) > \tau_{min}(j))$ and $(d_t(j) < \tau_{max}(j))$

$$\delta_t(j) = \delta_{t-1}(j-1) + \log[D_{j-1}d_{t-1}(j-1)]$$

$$+ \log[b_1(O_t)]$$

$$d_t(j) = 1$$

else

$$\delta_t(j) = \delta_{t-1}(j) + \log[b_1(O_t)]$$

$$d_t(j) = d_{t-1}(j)+1$$

[3] Otimal Observation probability P*:

$$\log P^* = \delta_t(N) + \log[D_N d_T(N)]$$

$$d_t(1) = d_{t-1}(1)+1$$

Ⅳ. 실험 및 결과

1. 음성 데이타

인식실험에 사용된 음성데이타는 표 1과 같이 숫자음 11개와 도시명

10개의 21개 어휘로 구성되어 있으며, 남성화자 6인이 15회씩 발음한 6명x15회x21단어=1890개의 데이타중 3명x15회x21단어=945개의 데이타를

HMM의 학습데이타로 이용하였고, 나머지 3명x15회x21단어=945개의 데이타를 인식실험에 사용하였다.

표 1. 음성 데이타

Table 1. Speech data.

숫자음	"공" "영" "일" "이" "삼" "사" "오" "육" "칠" "팔" "구"
지명	"서울" "부산" "대구" "대전" "인천" "광주" "강릉" "청주" "전주" "제주"

전체 인식시스템은 그림 4와 같이 구성되며, 먼저 음성신호를 차단주파수가 5.2kHz인 저역통과 필터를 통해 12kHz로 양자화한 신호를 $1-0.97z^{-1}$인 디지틀 필터로 고주파 성분을 강조하도록 전처리하였다. 그리고 음성을 20msec 길이로 프레임을 분할하고 이웃 프레임간의 겹침구간을 10msec로 하여 해밍윈도우 함수를 이용, 음성의 특징파라메타를 추출하였다. 이때 음성의 특징 파라메타로는 16차 LPC 켑스트럼을 사용하였다.

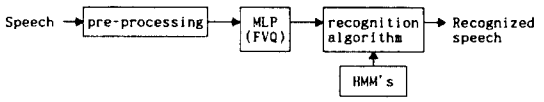


그림 4. 음성인식 시스템의 블럭도
Fig. 4. Block diagram of overall speech recognition system.

MLP의 구조는 그림 5와 같이 구성하여 인식 대상인 고립단어음 21가지에 포함된 기본 음소 18종류 1080개의 음소데이타를 유사 음소간의 분별력이 높이기 위해 모음그룹, 비음그룹 및 무성음그룹의 3개 그룹으로 나누고, 각 그룹별로 신경망을 학습하였다. 이와 같이 유사한 음소를 같은 그룹으로 분류하면 신경망은 같은 그룹에 속한 음소간의 차이만 구분하도록 학습되므로 분별력이 증가된다. MLP의 입력은 현재의 프레임과 전/후 프레임의 특징파라메타 48개의 켑스트럼으로서, 이로부터 은닉층 1 및 2를 거쳐 출력층의 값이 계산된다. 그림에서 맨 우측 신경망은 그룹분류 신경망으로 각 프레임의 음성을 그룹으로 분류하는 역할을 한다.

즉, 모음,비음 및 무성음은 음성적인 특징이 현저히 다르므로 그룹분류 신경망은 입력 음성을 3가지중 하나의 그룹으로 분류한다. 이때 그룹분류망의 은닉층 2의 값은 자기 음소그룹에 대해서는 1로, 다음 음소에 대해서는 0으로 각각 학습한다. 출력층의 값은 이들 3개 그룹의 출력과 그룹분류망의 출력으로 부터 계산된다. 출력층에서는 최종적인 출력값을 계산하는 것으로, 여기에 연결된 가중치는 각 그룹 및 그룹 신경망의 출력값으로 부터 학습되며, 모든 신경망은 오차 역전파 알고리듬[8]을 이용하여 학습했다.

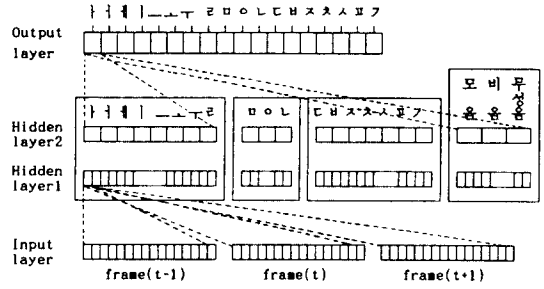


그림 5. MLP의 구조
Fig. 5. Structure of MLP.

2. 실험결과

그림 6과 그림 7은 각각 식(10)과 식(11)을 이용하여 3명×15회×21단어=945개 단어를 인식했을때 3,5,7상태 HMM모델에서 α와 β값의 변화에 따른 오인식률의 실험결과를 보이고 있다.

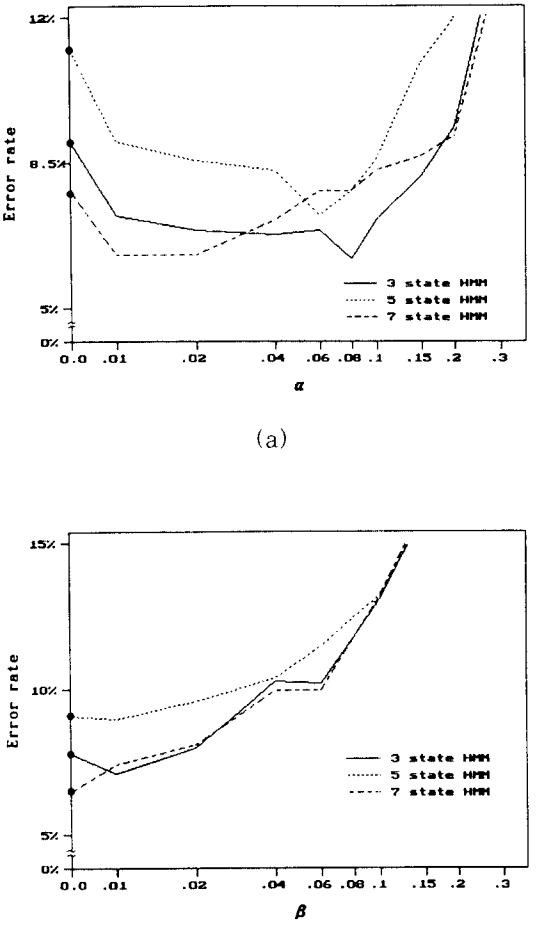
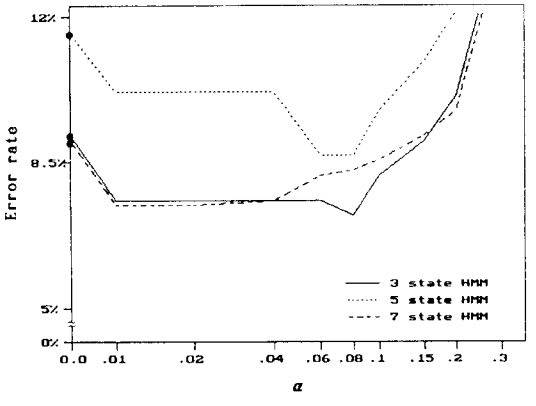


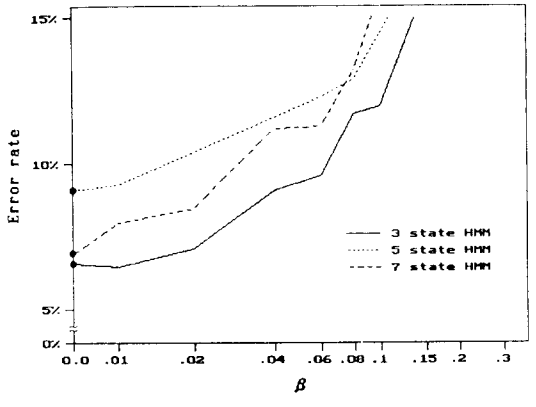
그림 6. HMM/Gauss분포에서 α와 β의 변화에 의한 오인식률
(a) α값에 따른 오인식률
(b) β값에 따른 오인식률
Fig. 6. Error rate as α and β values in HMM/Gauss distribution.
(a) Error rate as α values,
(b) Error rate as β values.

그림 6과 7에 나타난 것처럼 식(8)의 α값에 따라 설정된 상태의 최소 지속시간 τ_{min}을 조정함으로써 오인식률이 Gauss분포에서나 Gamma분포에서 α가 0.0일때의 오인식률보다 크게 감소되었으며, α의 범위는 0.01에서 0.08사이가 됨을 알 수 있다. 그림에서

α=0.0 및 β=0.0 에서의 오인식률은 상태의 지속시간을 제한하지 않은 경우의 결과를 나타낸다. 최대치 τ_{max}도 동일한 방법으로 설정했을때, 3상태 HMM에서는 β가 0.01일때 약간의 인식률의 개선이 있었지만 전체적으로 인식률을 크게 개선시키는 효과가 없음을 보이고 있다. 음성의 길이는 화자에 따라 다양한 차이를 보이고 있으며, 특히, 자음보다는 모음의 변화가 크므로 τ_{max}를 설정하여 지속시간을 제한하는 것은 인식률에 큰 영향을 미치지 못하는 것으로 생각된다.



(a)



(b)

그림 7. HMM/Gamma분포에서 α와 β의 변화에 의한 오인식률
(a) α값에 따른 오인식률 (b) β값에 따른 오인식률

Fig. 7. Error rate as α and β values in HMM/Gamma distribution.
(a) Error rate as α values, (b) Error rate as β values.

따라서 본 방법에서 상태지속시간의 τmin의 설정으로는 인식률의 향상을 기대할 수 있었지만 최대치 τmax는 인식률에 거의 영향을 미치지 않음을 실험을 통해 알 수 있었다.

표 2. 인식률(%)의 비교
Table 2. Comparisons of recognition rate(%).

구 분	상태	3	4	5	6	7	평균
CHMM	숫자음	87.47	87.71	86.46	86.67	87.68	87.20
	도시명	82.22	86.89	84.22	87.78	97.78	85.78
HMM/SD	숫자음	91.92	88.28	90.30	92.32	93.74	91.31
	도시명	89.56	91.78	88.67	91.44	91.11	90.31
HMM/CS T/Gauss	숫자음	95.97	93.94	94.55	94.55	95.56	94.91
	도시명	92.44	93.78	91.56	92.22	92.22	92.44
HMM/CS T/Gamma	숫자음	96.16	91.92	93.13	94.55	94.95	94.06
	도시명	91.33	93.11	91.11	92.44	91.56	91.91

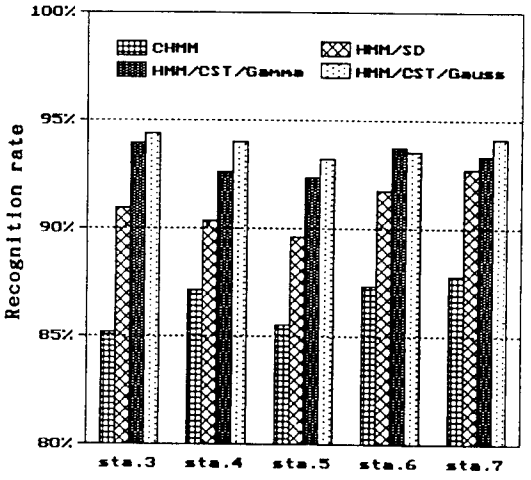


그림 8. 상태별 전체 인식률의 비교
Fig. 8. Comparisons of total recognition rate for each state.

표 2는 숫자음과 도시명의 인식률을 나타낸 것으로서, CHMM은 기존의 연속 HMM을, HMM/SD는 상태지속시간을 고려한 HMM을 나타낸다. 또, HMM/CST는 τ_{min}과 τ_{max}를 이용한 HMM이다. 그림 8은 표 2에서 보인 전체 인식률을 상태별로 나타낸 것이다. 그림에서 보는 바와 같이 HMM/SD에 비해서, HMM/CST/Gauss는 3상태 HMM에서 3.50%, 4상태 HMM에서 3.81%, 5상태 HMM에서 3.70%, 6상태 HMM에서 2.01%, 7상태 HMM에서 1.38%의 인

식률이 향상되었으며, HMM/CST/Gamma는 2.97%, 2.44%, 2.75%, 2.12%, 0.64%의 인식률이 향상되었음을 보였다. 실험결과에서 나타난 것처럼 제안한 방법의 HMM은 상태수가 7인 HMM에서 인식률을 HMM/SD보다 거의 향상시키지 못하고 있지만 상태수가 5이하인 HMM에서는 오인식률을 약 1/3정도 감소시켜 인식률을 개선시킴을 알 수 있다.

V. 결 론

본 논문에서는 HMM의 상태천이를 제한하여 인식률을 향상시킬 수 있는 인식시스템을 구성하였다. 이 시스템에서는 HMM의 각 상태에서 지속시간을 연속 확률함수로 근사화하여 최소 및 최대 지속시간의 경계치를 추정하고, 상태의 최소, 최대 지속시간 동안에는 다른 상태로의 천이를 제한하였다. 제안한 모델의 성능을 확인하기 위해 숫자와 도시명으로 구성된 21개의 단어를 대상으로 화자독립 고립 단어 인식실험을 수행한 결과 기존의 모델에 비해 오인식률을 1/3정도 감소시킬 수 있었다. 상태의 수가 5이하일때 HMM/CST/Gauss인 경우의 오인식률은 6.3%이고, HMM/CST/Gamma인 경우는 7.2%로서 기존 모델의 9.9%에 비해 인식률이 각각 3.6%, 2.7%씩 개선되었으며, 상태의 수가 6과 7일때는 각각 6.3%이고, 6.6%로서 기존 모델의 8.0%에 비해 인식률이 각각 1.7%, 1.4%씩 개선되었다. 제안한 모델을 이용한 인식 알고리듬은 기존의 인식 알고리듬에 상태지속시간만 고려하므로 계산량을 증가시키지 않으면서도 인식률을 개선할 수 있다.

참 고 문 헌

[1] H.Sakoe, S.Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on ASSP*, vol. ASSP-26, p.43-49, Feb.1978.

[2] L.R.Rabiner, S.E.Levinson, and M.M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *B.S.T.J.*, vol.62, pp.1075-1105, Apr.1983.

[3] L.R.Rabiner, B.H.Juang, S.E.Levinson, and M.M.Sondhi, "Recognition of isloated digits using hidden Markov model with continuous mixture densities," *AT &T Tech J.*, vol.64, pp.1211-1234, July-Aug.1985.

[4] B.H.Juang, "Maximum-likelihood estimation for mixture multivariate stochastic obsevations of Markov chains," *AT&T Tech J.*, vol.64, pp. 1235-1249, July-Aug.1985.

[5] R.P.Lippmann, "Review of neural networks for speech recognition," *Readings in Speech Recognition*, Morgan Kufmann, Pub., 1990.

[6] A.Waibel, T.Hanazawa, G.Hinton, K. Shikano and K.J.Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans.* on ASSP, vol. ASSP-37, pp.328-339, Mar.1989.

[7] S.Renals, N.Morgan, H.Bourlard, M. Coheal and H.Franco, "Connectionist Probability Estimators in HMM Speech Recognition," *IEEE Trans.* on speech and audio processing, vol.2, no.1, part II, pp.161-174, Jan.1994.

[8] P.L.Cerf, W.MA and D.V.Compernolle, "Multilayer Perceptrons as Labelers for Hidden Markov Models," *IEEE Trans.* on speech and audio processing, vol.2, no.1, part II, pp.185-193, Jan.1994.

[9] B.H.Juang and L.R.Rabiner, "Mixture Autoregressive Hidden Morkov Models for Speech Signals," *IEEE Trans.* on ASSP, vol.33, no.6, pp.1404-1413, Dec.1985.

[10] H.Y.Gu, C.Y.Tseng and L.S.Lee, "Isolated-Utterance Speech Recognition Using Hidden Morkov Models with Bounded State Durations," *IEEE Trans.* on signal processing, vol.39, no.8, pp.1743-1752, Aug.1991.

[11] J.Tebelskis and A.Wailbel, "Performance Through Consistency: MS-TDNN's for Large Vocabulry Continuous Speech Recognition," *Advanced in Neural Information Processing Systems* V, pp.696-703, Dec.1992.

[12] T.H.Crystal and A.S.House, "Charact-

erization and Modeling of Speech-Segment Duration," *Proc. ICASSP'86*, pp.2791-2794, Tokyo, Japan, 1992.

[13] H.P.Tseng, M.Sabin, and E.Lee, "Fuzzy vector quantization applied to hidden Markov modeling," *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.* on Speech Audio Processing, vol.2, no.1, part 2, pp.185-193, Jan.1994.

[15] H.Bourland and C.J.Wellekens, "Links between Markov models and multi-layer perceptrons," *Advanced in Neural Information Processing Systems*, vol.1, pp.502-510, 1989.

[16] 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.

저 자 소 개

李 基 熙(正會員) 第 30卷 B編 第 5號 參照
현재 대유공업전문대학 사무자동화과 전임강사

林 寅 七(正會員) 第 30卷 B編 第 2號 參照
현재 한양대학교 전자공학과 교수