세그먼트 차원압축을 이용한 HMM의 음절인식

Syllable Recognition of HMM using Segment Dimension Compression

김 주 성*, 이 양 우**, 허 강 인*, 안 점 영***
(Joo Sung Kim*, Yang Woo Lee**, Kang In Hur*, Jum Young Ahn***)

※이 연구는 1995학년도 동의대학교 자체 학술 연구 조성비에 의하여 연구되었음.

요 약

본 논문은 단음절 전구간에 대해 4프레임폭과 7프레임폭을 결합하여 만든 40차원의 세그먼트를 K-L전개와 신경망으로 각각 10, 14, 20차원으로 압축하여 연속분포 HMM의 음성인식 특징파라미터로 사용하였다. 그리고 이산지속시간, 회귀계수 그리고 혼합분포를 특징파라미터로 추가한 경우와 비교검토하였다.

단음절 100개에 대한 인식실험결과 연속분포 HMM의 인식률 85.19%에 비해 회귀계수를 부가한 경우 1.4%, 혼합분포를 이용한 경우 2.36%, 이산 지속시간제어를 한 경우 2.78%의 인식률이 향상되었다. 그리고 K-L전개에 의한 압축파라미터만 이용한 경우는 멜켑스트럼＋회귀계수의 경우보다 인식률이 낮았으나, K-L전개에 의한 압축파라미터에 멜켑스트럼과 회귀계수를 부가한 경우는 동등한 결과를 얻을 수 있었다. 신경망에 의한 압축파라미터를 이용한 경우에는 비선형 변환인 시그모이드 함수를 사용하므로 음성의 동적변화가 잘 반영되어 K-L전개 및 다른 방법에 비해 향상된 인식결과를 얻을 수 있었다.

ABSTRACT

In this paper, a 40 dimensional segment vector with 4 frame and 7 frame width in every monosyllable interval was compressed into a 10, 14, 20 dimensional vector using K-L expansion and neural networks, and these was used to speech recognition feature parameter for CHMM. And we also compared them with CHMM added as feature parameter to the discrete duration time, the regression coefficients and the mixture distribution.

In recognition test at 100 monosyllable, recognition rates of CHMM＋△MCEP, CHMM＋MIX and CHMM＋DD respectively improve 1.4%, 2.36% and 2.78% over 85.19% of CHMM. And those using vector compressed by K-L expansion are less than MCEP＋△MCEP but those using K-L＋MCEP, K-L＋△MCEP are almost same. Neural networks reflect more the speech dynamic variety than K-L expansion because they use the sigmoid function for the non-linear transform. Recognition rates using vector compressed by neural networks are higher than those using of K-L expansion and other methods.

Ⅰ. 서 론

최근 컴퓨터 및 정보통신 기술의 급속한 발전과 보급이 활발해 짐에 따라 음성에 의한 인간-기계의 인터페이스에 대한 기대가 높아지게 되었고, 향후 온라인 시스템, 대화시스템과 자동통역 시스템의 구현을 위해서는 무엇보다도 음성인식 기술의 연구가 선행 되어져야 할 것이다. 현재의 음성 인식은 DP매칭, HMM 및 신경망에 의한 연구가 진행되고 있다.[3][8]

기존의 DP매칭법은 시계열 패턴의 시간축상에서의 비선형 신축을 허용하는 패턴 조합 알고리즘이다. 이것은 단어(또는 음운, 음절)의 표준적 특징을 가지고 있는 시계열 패턴을 표준패턴으로 하고 입력된 시계열 패턴을 비선형으로 신축해 가면서 조회하는 방법이다. 이 방법은 시계열 패턴의 시간적 구조의 변동을 잘 처리할 수가 있지만 화자의 개인차 등에 따라 발생하는 스펙트럼의 변동에 대해서는 어려운 점이 많다. 이러한 단점을 보완할 수 있는 방법으로서 확률적 인식 방법인 HMM이 있다.

HMM은 확률모델을 이용하기 때문에 개인차나 조음결합에 의한 음성 패턴의 변동을 반영하기 쉽고, 이론적인 전개가 용이하며 언어처리도 동일한 확률모델로서 표

*동아대학교 전자공학과
**동의대학교 전기공학과
***동의대학교 전자공학과
접수일자: 1996년 2월 7일

현할 수 있는 장점이 있다.[1][2] 그러나 모델구조 결정에 있어서 시행착오에 대한 의존성이 높고, 학습시 많은 데이터가 필요하며, 음성의 과도적 정보를 경시하는 경향이 있으며 단순 Markov과정으로 가정하기 때문에 패턴의 시간적 상관에 대한 표현력이 부족하다는 단점이 있다.[6][7]

일반적으로 음성특징량의 동적 변화를 반영하기 위한 방법으로 상태수를 늘이는 방법, 특징량의 시간축 방향의 회귀계수를 파라미터로 이용하는 방법, 이산 지속시간제어를 하는 방법, 상태에 대응하는 음성구간에서 여러개의 단일 정규분포를 조합하여 하나의 혼합분포로 근사화하는 방법과 세그먼트 통계량을 이용하는 방법 등이 있다.

본 논문에서는 세그먼트 통계량을 이용하기 위하여 멜 캡스트럼에 대해 여러개의 프레임을 결합한 고정길이의 세그먼트를 구성하였다. 그러나 구성한 세그먼트는 40차원이므로 많은 학습 시간이 소요되고 특히 학습데이터가 불충분하면 모델의 파라미터를 추정하기 어렵기 때문에 파라미터의 차원을 압축하여 추정할 파라미터의 차원수를 감소시켰다. 차원압축은 주성분 분석인 K-L(Karhunen Loeve) 전개와 비선형변환의 시그모이드 함수를 사용하는 신경망으로 40차원을 10, 14, 20차원으로 압축하여 연속분포 HMM의 음성인식 특징 파라미터로 사용하였다. 그리고 이산지속시간, 회귀계수 그리고 혼합분포를 특징 파라미터로 부가한 연속분포 HMM의 경우와 비교 검토하였다.

Ⅱ. HMM에 의한 음성 인식

2.1 연속출력 확률분포 HMM

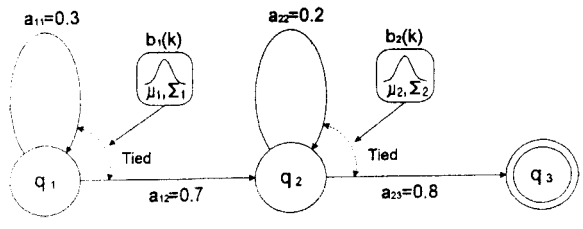


그림 1. 연속출력 확률분포 HMM의 예

Left-to-right형 HMM은 그림 1과 같은 유한 오토마타로 정의된다. HMM을 이용하여 음성인식을 할 때에는 미리 필요한 표준패턴 수 만큼의 모델을 준비해 놓고 미지의 입력 패턴에 대해서 출력확률을 최대로 하는 표준패턴을 인식결과로 하는 것이다. 연속출력분포의 경우 상태 i에서 j로의 천이확률을 a_{ij} 및 천이경로에서 심볼 k의 출력확률을 b_{ijk}를 학습데이터에서 구하기 위한 Baum-Welch 알고리듬은 다음과 같다.[4][7][10]

상태수를 N, T를 심벌 계열의 길이, 전향확률을 $\alpha(i, t)(i=1, 2, \cdots, N; t=1, 2, \cdots, T)$이라 하고 후향확률을 $\beta(j, t)(j=1, 2, \cdots, N; t=T, T-1, \cdots, 0)$이라 할 때 모델 M의 심벌 계열 $o=o_1\, o_2 \cdots o_T$를 출력하는 확률 $P(o|M)$을 이용해서 상태 i에서 상태 j로의 천이가 시각 t에서 발생할 확률을

$$\gamma_t(i, j) = \frac{\alpha(i, t-1)\, a_{ij}\, b_{ij}(o_t, \mu_{ij}, \sum_{ij})\, \beta(j, t)}{P(o|M)} \tag{1}$$

로 정의하면, 천이확률의 추정식은

$$a_{ij} = \frac{\sum_t \gamma_t(i, j)}{\sum_t \sum_j \gamma_t(i, j)} \tag{2}$$

와 같고, 출력벡터 o_t가 n차원의 정규분포에 따른다고 가정할 수 있는 경우 출력확률밀도 함수는

$$b_{ij}(o_t, \mu_{ij}, \sum_{ij}) = \frac{\exp\{-(o_t-\mu_{ij})^t \sum_{ij}^{-1}(o_t-\mu_{ij})/2\}}{(2\pi)^{n/2}\, |\sum_{ij}|^{1/2}} \tag{3}$$

로 주어진다. 여기서, μ_{ij}는 출력벡터의 평균치, $\sum_{ij}$는 공분산행렬, t는 전치, -1은 역행렬을 나타낸다. 여기서 μ_{ij}, $\sum_{ij}$의 추정식은 다음 식으로 주어진다.

$$\mu_{ij} = \frac{\sum_t \gamma_t(i, j)\, o_t}{\sum_t \gamma_t(i, j)} \tag{4}$$

$$\sum_{ij} = \frac{\sum_t \gamma_t(i, j)(o_t-\mu_{ij})(o_t-\mu_{ij})^t}{\sum_t \gamma_t(i, j)} \tag{5}$$

2.2 이산 지속시간제어 HMM

Viterbi 알고리듬을 그대로 사용하면 종래의 HMM에서는 상태 i에 n시간으로 멈출확률은 $a_{ii}^{t-1}\cdot(1-a_{ii})$이 되고 n의 증가와 함께 지수 함수적으로 감소하며 과도구간과 정상구간의 시간구조를 충분히 표현할 수 없다. 실제 각 상태에서의 지속 시간은 음성 세그먼트의 발성시간을 나타내는 중요한 정보이므로 인식시에 이를 고려하는 것이 바람직하다. HMM에서 상태 지속시간은 음성 세그먼트의 길이를 나타내며 일반적으로 음성 세그먼트의 길이는 감마 분포나 포아송 분포에 가까운 것으로 알려져 있다.

상태 지속 시간제어를 통계적으로 실시하기 위해 a_{ii}의 자기 루프 천이를 제거하고 대신에 상태 i가 t시간 지속될 지속 시간의 확률 $d_i(t)$을 구하고 이것을 새로운 파라미터로 추가한다. 이때 음성 지속시간 확률밀도 분포를 구하는 방법에 따라 이산분포 지속시간제어 모델과 감마 분포나 포아송 분포를 이용하는 연속분포 지속시간제어 모델을 구성할 수 있다. 훈련 샘플 수가 비교적 충분하다고 가정하고 계산시간이 빠른 장점을 이용하기 위해 이

산분포 지속시간제어 모델만을 고려한다.

단 지속시간 확률은

$$\sum_{\tau} d_i(\tau) = 1$$

의 조건을 만족해야 하고 이 파라미터를 도입하면 연속출력 확률분포 HMM의 경우 Baum-Welch의 재추정 알고리듬은 다음과 같이 변화된다.[6][7][10]

$$\alpha(i, j) = \sum_{j} \sum_{\tau \leq t} \alpha(j, t-\tau)\, a_{ij} d_i(\tau) \prod_{n=1}^{\tau} b_{ji}(o_{t+1-n}) \quad (6)$$

$$\beta(i, j) = \sum_{j} \sum_{\tau \leq T-t} a_{ij} d_j(\tau) \prod_{n=1}^{\tau} b_{ij}(o_{t+n}) \beta(j, t+\tau) \quad (7)$$

여기서

$$\gamma_t(i, j, \tau) = \frac{\alpha(i, t-\tau)\, a_{ij} d_j(\tau) \prod_{n=1}^{\tau} b_{ij}(o_{t+1-n}) \beta(j, t)}{P(o|M)} \quad (8)$$

로 하면 천이확률 a_{ij}와 정규분포의 파라미터 μ_{ij}, Σ_{ij}의 추정식은 각각 다음 식으로 주어진다.

$$a_{ij} = \frac{\sum_{t} \sum_{\tau \leq t} \gamma_t(i, j, \tau)}{\sum_{t} \sum_{j} \sum_{\tau \leq t} \gamma_t(i, j, \tau)} \quad (9)$$

$$\mu_{ij} = \frac{\sum_{t} \sum_{\tau \leq t} \gamma_t(i, j, \tau) \sum_{n=1}^{\tau} o_{t+1-n}}{\sum_{t} \sum_{\tau \leq t} \gamma_t(i, j, \tau)} \quad (10)$$

$$\Sigma_{ij} = \frac{\sum_{t} \sum_{\tau \leq t} \gamma_t(i, j, \tau) \sum_{n=1}^{\tau} (o_{t+1-n} - \mu_{ij})(o_{t+1-n} - \mu_{ij})^t}{\sum_{t} \sum_{\tau \leq t} \gamma_t(i, j, \tau)\, \tau} \quad (11)$$

또 지속시간확률 $d_i(\tau)$의 추정치는

$$d_i(\tau) = \frac{\sum_{t} \sum_{t} \gamma_t(i, j, \tau)}{\sum_{t} \sum_{j} \sum_{\tau \leq t} \gamma_t(i, j, \tau)} \quad (12)$$

로 된다. 식(12)에 의한 추정만으로 학습횟수가 진행됨에 따라서 지속시간 확률분포의 차가 너무 크게 나타나기 때문에 다음 식(13)과 같은 가중치 평균을 이용해서 확률분포의 스므딩을 실시한다.

$$d_j(\tau) = \begin{cases} \{2d_j(\tau) + d_j(\tau+1)\}/3 & \text{if } \tau = 1 \\ \{d_j(\tau-1) + 2d_j(\tau)\}/3 & \text{if } \tau = \Delta T \\ \{d_j(\tau-1) + 2d_j(\tau) + d_j(\tau+1)\}/4 & else \end{cases} \quad (13)$$

2.3 동적특징 파라미터를 이용한 HMM

연속출력 확률분포 HMM에서는 음성의 정상구간의 HMM을 1개의 상태에 대응시킨다. 음성의 특징은 스펙트럼 변화에 많은 정보가 포함되어 있으므로 스펙트럼의 동적 변화도를 고려해야 한다는 것은 잘 알려진 사실이지만 연속출력 확률분포 HMM에서는 정적인 스펙트럼의 불안정성이나 시간의 흐름을 확률적으로 모델링하고 있을 뿐 실제로 동적인 파라미터는 고려되지 않고 있다.

따라서 음성인식의 특징 파라미터로서 회귀계수에 의한 스펙트럼의 동적 특징량의 이용을 고려하였다. 음성의 정적 특징량(켑스트럼 계수) 벡터의 각각의 차원에 대해서 식(15)로 계산되는 시간축 방향의 선형 회귀계수를 파라미터로 추가한다. 학습·인식에 있어서 2개의 특징 파라미터인 정적 특징량과 회귀계수의 각각에 대해서 식(3)으로 출력확률을 계산해 놓고 식(14)와 같이 이들의 곱을 벡터에 대한 출력확률로 한다. 여기서 정적 특징량과 회귀 계수의 상관은 고려하지 않았다.[6][7][10]

$$b_{ij}(o_t) = b_{ij}^{cep}(o_t) \cdot b_{ij}^{reg}(o_t) \quad (14)$$

시계열 $x(t)$에 있어서 시간 t를 중심으로 한 $2n+1$폭의 선형회귀계수 $\Delta x(t)$는 다음 식으로 계산된다.

$$\Delta x(t) = \frac{\sum_{i=-n}^{n} i \cdot x(t+i)}{\sum_{i=-n}^{n} i^2} \quad (15)$$

여기서 t는 프레임, n은 회귀계수의 폭, i는 회귀 계수의 차수를 나타낸다.

2.4 혼합연속출력 확률분포 HMM

연속출력 확률분포 HMM은 음성구간을 단일정규분포로 근사화해서 파라미터를 추출하였다. 그러나 어떤 상태에 대응하는 음성구간은 한개의 정규분포로 근사할 수 없는 경우가 있다. 그래서 여러개의 단일정규분포를 조합하여 하나의 복잡한 분포로 근사화하기 때문에 학습패턴에 가장 근사된 표현을 할 수 있다.[5][10]

따라서 혼합연속출력 확률분포로 HMM 모델을 구성하면 상태 i에서 j로의 천이에서 출력벡터 o_t의 출력확률 $b_{ij}(o_t)$는 다음과 같이 M개의 연속분포의 가중치합에 의해서 표현되는 것으로 한다.

$$b_{ij}(o_t) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(o_t) \quad (16)$$

단

$$\sum_{m=1}^{M} \lambda_{ijm} = 1, \qquad \int b_{ijm}(o)\, do = 1 \quad (17)$$

여기서 λ_{ijm}은 분기확률로 m번째의 출력확률 밀도분포의 출현확률이고, b_{ijm}은 m번째의 출력 확률밀도분포를 나타낸다. $b_{ijm}(o)$로써 정규분포를 가정한 경우

$$b_{ijm} = b_{ij}(o_t, \mu_{ijm}, \Sigma_{ijm}) \quad (18)$$

로 된다. 천이확률의 추정식은 식(2)와 같고

$$\gamma_t(i, j, m) = \frac{\alpha(i, t-1)\, a_{ij} b_{ijm}(o_t)\, \beta(j, t)}{P(o|M)} \quad (19)$$

로 정의하면 λ_{ijm}, μ_{ijm}, $\sum_{ijm}$의 추정식은 다음 식으로 주어진다.

$$\lambda_{ijm} = \sum_t \gamma_t(i, j, m) / \sum_t \sum_m \gamma_t(i, j, m) \quad (20)$$

$$\mu_{ijm} = \sum_t \gamma_t(i, j, m)\, o_t / \sum_t \gamma_t(i, j, m) \quad (21)$$

$$\sum_{ijm} = \frac{\sum_t \gamma_t(i, j, m)(o_t - \mu_{ijm})(o_t - \mu_{ijm})^t}{\sum_t \gamma_t(i, j, m)} \quad (22)$$

회귀계수까지 고려하면 출력확률은 다음과 같이 된다.

$$b_{ij}(o_t) = \sum_{m=1}^{M} \lambda_{ijm}^{mel}\, b_{ijm}^{mel}(o_t^{mel}) \sum_{m=1}^{M} \lambda_{ijm}^{rgc}\, b_{ijm}^{rgc}(o_t^{rgc}) \quad (23)$$

단,

o_t^{mel} : 멜 캡스트럼 계수
o_t^{rgc} : 회귀계수
λ_{ijm}^{mel}, b_{ijm}^{mel} : 멜 캡스트럼 계수의 분기확률과 출력확률
λ_{ijm}^{rgc}, b_{ijm}^{rgc} : 회귀계수의 분기확률과 출력확률

Ⅲ. 파라미터 압축에 의한 음성 인식

3.1 차원 압축법

세그먼트는 그림 2와 같이 (a)의 4 프레임폭(인접한 4 프레임을 결합)과 (b)의 7 프레임폭(한 프레임씩 건너서 프레임을 결합)으로 설정하였다. 이 때 1 세그먼트는 40차원이다. 이 세그먼트를 1 프레임씩 이동시키면서 전음절 구간에 대하여 K-L전개와 신경망으로 차원압축한다.

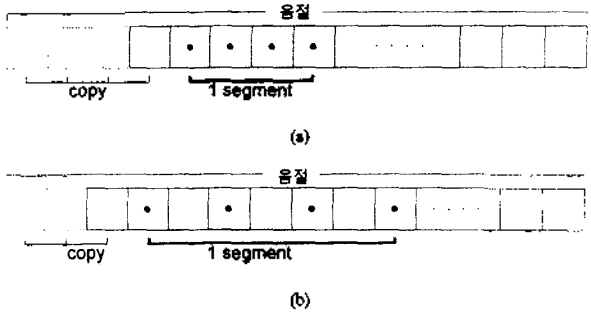


그림 2 세그먼트 구성방법 (a)4 프레임폭, (b)7 프레임폭

3.1.1 K-L전개에 의한 차원압축

몇개의 프레임을 결합한 세그먼트를 1개의 벡터로 해서 음성 특징량으로 취급하면 추정해야 할 벡터의 차원수가 증가한다. 그러므로 세그먼트를 표현하는 벡터에서 분산이 적은 1차 결합성분은 소거하고 큰 분산만을 갖는 성분만을 취하기 위해서 다음과 같은 순서로 K-L전개를 한다.[8][11][12]

① 샘플에 의한 공분산 행렬 $A = [a_{lm}]$의 추정.

$$a_{lm} = \sum_{i=1}^{I} x_i^l\, x_i^m \quad (24)$$

② 고유치 $\{\lambda_j\}$와 고유벡터 $\{\Phi_j\}$를 계산한다.

$$A\Phi_j = \lambda_j \Phi_j \quad (25)$$

③ $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p$가 되도록 고유치와 그것에 대응하는 고유벡터를 정렬한다.

④ 압축후의 파라미터를 계산한다.

변환행렬을 $B = [\Phi_1 \Phi_2 \cdots \Phi_p]^T$라 하면

$$y_i = Bx_i \quad (26)$$

여기서

N : 파라미터의 차원
x_i : 샘플($i = 1, \cdots, I$)
$x_i = [x_i^1, x_i^2, \cdots, x_i^N]^T$, 단 기호 T는 전치를 나타낸다.
x_i^k : 백터 x_i의 k번째 요소
x_M : 샘플의 평균 백터
z_i : $x_i - x_M (z_i = (z_i^1, z_i^2, \cdots, z_i^N))$
p : 차원 압축 후의 특징파라미터의 차원
y_i : 차원 압축 후의 특징파라미터
$y_i = [y_i^1, y_i^2, \cdots, y_i^p]^T$

로 한다.

3.1.2 신경망에 의한 차원 압축

중간층의 unit수가 입력층과 출력층 보다도 적은 MLP를 이용하여 입력층의 입력신호와 동일한 신호를 출력층의 교사신호로 부여하면 차원압축을 할 수 있다. 일반적으로 3층의 신경망은 K-L전개보다도 압축율이 떨어지므로 그림 3과 같은 5층의 신경망을 이용한다.[8][11][12]

신경망에서는 비선형변환인 시그모이드함수를 이용하기 때문에 선형변환인 K-L전개에 비해 보다 효율적인 차원압축을 기대할 수 있다. 여기서는 제 1층과 제 3층의 unit에 대해서는 시그모이드 함수로 비선형으로 하였고, 입력층과 제 2층 그리고 출력층에 대해서는 선형으로 하였다.

3.2 차원 압축 오차

3.2.1 K-L전개에 의한 압축오차

K-L전개에 의한 데이터 압축정도는 학습데이터에 대해서 공분산행렬의 고유치의 누적기여율로 평가했다. 누적기여율Θ_i는 다음과 같이 정의한다.

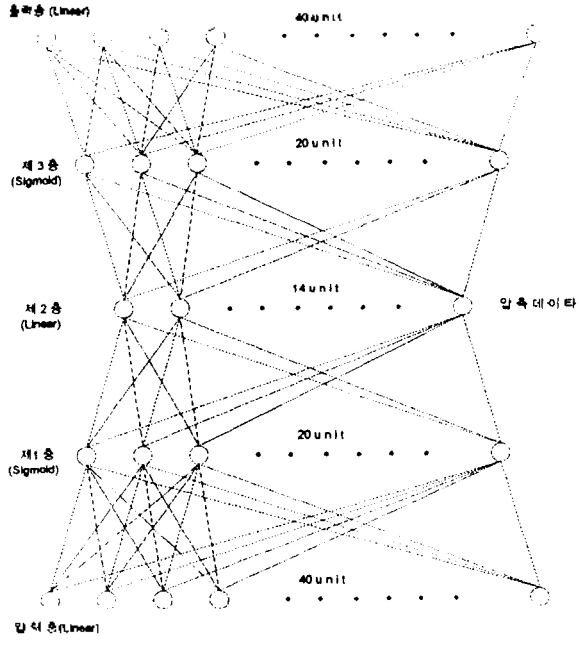


그림 3. 차원압축용 5층 신경망 구조(14차원으로 압축하는 경우)

$$\Theta_i = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{40} \lambda_k} \quad (\lambda_k \geq \lambda_{k+1}) \tag{27}$$

그리고, K-L전개에 의한 세그먼트내의 스펙트럼의 동적 특징량이 잘 보존되고 있는지를 알아보기 위해서 K-L 전개로 압축된 데이터를 원래의 세그먼트(4 프레임)로 복원해서 프레임간의 자승오차와 인접하는 2프레임간의 평균 자승거리를 비교하였다.

여기서 복원오차가 인접 2프레임간의 평균거리보다 작으면 복원된 벡터계열이 원래의 인접 프레임보다 동적특징량을 잘 보존하고 있다고 할 수 있다. 또 회귀계수를 파라미터에 부가한 경우, 세그먼트내의 프레임은 직선으로 근사화할 수 있으므로 이 1차 직선과 실제 프레임과의 거리(오차)를 구하였다. 프레임 i에서 제 n차원의 최소자승 오차직선(선형회귀계수)의 기울기와 오차직선의 절편 및 직선으로 근사화시킨 경우의 평균 자승오차를 구하는 식을 각각 식(28), (29), (30)에 나타내었다.

$$a_i^n = \frac{\sum_{k=-w}^{w} k \cdot y_{i+k}^n}{\sum_{k=-w}^{w} k^2} \tag{28}$$

$$b_i^n = \frac{\sum_{k=-w}^{w} y_{i+k}^n}{2w+1} \tag{29}$$

$$\text{평균오차} = \frac{1}{2w+1} \sum_{k=-w}^{w} \sum_{n} (y_{i+k}^n - k \cdot a_i^n - b_i^n)^2 \tag{30}$$

단 y_i^n은 프레임 i의 제 n차원 요소이고, $2w+1$은 회귀계수의 계산시간 프레임폭(본 실험에서는 $w=5$)이 된다.

표 2는 인접 2 프레임간 평균거리, 최소자승 직선에 근사화시킨 경우의 오차, K-L전개에 의한 압축오차를 나타내었다. 자승오차 직선을 근사화시킨 경우는 인접 2 프레임간의 평균거리보다 작았다. 그리고 프레임폭이 다르므로 단순하게 비교할 수 없지만 K-L전개에 따른 압축오차(복원오차)는 음절 데이터에서 인접 2 프레임간 평균거리보다 작으므로 양호한 스펙트럼의 동적특징량을 보존하고 있다고 할 수 있다. 그러고 4 프레임폭과 7 프레임폭의 세그먼트에서는 누적기여율과 비례해서 압축오차도 감소하였다.

표 1. 누적기여율(20차원까지 나타내었다)

(a) 4 프레임 폭

i	θ_i	i	θ_i
1	0.4251	11	0.9616
2	0.6992	12	0.9651
3	0.7974	13	0.9682
4	0.8420	14	0.9711
5	0.8814	15	0.9736
6	0.9042	16	0.9759
7	0.9218	17	0.9781
8	0.9354	18	0.9800
9	0.9466	19	0.9820
10	0.9556	20	0.9837

(b) 7 프레임 폭

i	θ_i	i	θ_i
1	0.4065	11	0.9453
2	0.6762	12	0.9509
3	0.7725	13	0.9561
4	0.8163	14	0.9600
5	0.8546	15	0.9632
6	0.8774	16	0.9663
7	0.8987	17	0.9694
8	0.9149	18	0.9721
9	0.9275	19	0.9746
10	0.9367	20	0.9769

표 2. K-L전개에 의한 압축오차

방	법	음절데이타
인접 프레임간 평균거리		0.1050
회귀계수(11프레임폭)		0.0351
K-L (4프레임폭)	10차원	0.0211
	14차원	0.0156
	20차원	0.0096
K-L (7프레임폭)	10차원	0.0257
	14차원	0.0185
	20차원	0.0102

3.2.2 신경망에 의한 압축오차

사용된 신경망은 각층의 unit수가 40×15×10×15×40, 40×20×14×20×40, 40×30×20×30×40인 3종류의 5층의 구조를 이용하여 4 프레임폭과 7 프레임폭의 세그먼트(40차원)를 각각 10차원, 14차원, 20차원으로 압축한다. 비선형변환을 행하는 시그모이드함수는 제 1층과 제 3층의 unit에서만 사용하고, 그외 층의 unit에서는 선형으로 한다.

신경망의 학습은 역전파 알고리듬을 사용하였으며, 학습후 신경망와 입력층에 학습데이타 및 평가데이타를 주고, 그때의 제 2층의 출력을 이용하여 HMM의 학습과 평가실험을 수행한다.

표 3에는 압축된 데이타를 복원하였을때의 자승오차를

나타내었다. 신경망에 의한 압축오차는 10차원, 14차원에서 K-L전개보다 크게 나타났는데, 이것은 신경망의 학습횟수를 2000회로 제한하였기 때문이다. 신경망의 학습횟수를 높이면 압축오차는 더욱 줄어들 것이다. 20차원의 경우 신경망의 압축오차는 K-L전개보다 감소하였다.

표 3. 신경망에 의한 압축오차

방	법	음절데이타
4프레임폭	10차원	0.0491
	14차원	0.0409
	20차원	0.0013
7프레임폭	10차원	0.0577
	14차원	0.0569

Ⅳ. 인식실험 및 고찰

4.1 음성 DB 및 분석 방법

음성 데이터는 표 4와 같이 20대 남성화자 5명이 5회 발성한 한국어 단음절 100개를 이용하였고, 분석조건은 표 5와 같다. 음성 DB는 신문사설과 초등교과서 중에서 빈도수가 높은 음절을 조사하여 그 중 100개를 사용하였다. 본 실험에서는 5명의 화자가 5회 발성한 음절중에서 3회분은 학습용, 나머지 2회분은 평가용으로 사용하였다.

표 4. 음성 DB

가	간	갈	감	거	게	고	고	기
나	난	날	남	네	노	누	니	
다	단	달	담	데	도	동	두	다
라	란	람	로	롱	루	리		
마	만	말	에	모	무	미		
바	반	발	보	봉	부	비		
사	산	살	삼	세	소	수	시	
아	안	알	암	에	우	이		
자	잔	잘	잠	체	조	종	주	지
차	찬	찰	참	체	초	추	치	
키	킬							
타	탄	탈	탐	토	투	티		
파	판	팔	포	피				
하	한	할	함	후	히			

표 5. 음성데이터의 분석조건

A/D 데이터	10 KHz, 12 bit
고역강조	1 차 차분
프레임 간격	5 ms
분석창 길이	10 ms
특징 파라메터	LPC Cepstrum(14차) → LPC Melcepstrum(10차)
회귀계수	10 차, 45 ms 폭

4.2 파라미터에 대한 인식결과

각 실험은 다음의 방법에 대해서 상태수를 4~7로 늘이면서 Baum-Welch 알고리듬을 이용하여 학습하였다. 모델의 초기치 추정은 학습용 단음절 데이터를 균등하게 4개의 구간으로 나누어서 각 구간에 할당된 벡터로 부터 평균벡터와 공분산행렬의 초기치를 구하였다. 최대 반복 학습의 횟수는 10회로 하였다.

각 방법에 대한 음절 인식률은 표 6과 같다.

①CHMM

연속출력 확률분포로 하여 멜켑스트럼을 이용.

②CHMM + DD

연속출력 확률분포로 하고 이산 지속시간제어를 한 경우.

③CHMM + △MCEP

파라미터로서 멜켑스트럼과 동적특징량으로 선형회귀계수를 부가한 경우(본 실험에서 선형회귀계수의 폭은 11 프레임폭)

④CHMM + MIX

출력확률분포를 여러개의 확률분포로 이용한 경우 (본 실험에서 혼합수는 2)

표 6. 파라미터에 따른 인식률 [단위 %]

상태수	①CHMM	②CHMM+DD	③CHMM+△MCEP	④CHMM+MIX
4	82.85	85.11	84.72	87.39
5	85.59	88.80	87.16	88.12
6	87.73	89.35	88.29	88.59
7	87.40	91.40	89.01	88.93
평균	85.89	88.66	87.29	88.25

인식실험에서 연속분포 HMM보다 동적 특징량으로 회귀계수를 부여한 경우 1.4%, 혼합분포를 사용한 경우 2.36%, 이산지속시간 제어를 한 경우 2.78%의 인식률이 향상되었다. 또한 상태수 증가에 따라서도 인식률이 향상되었으며 상태수 4와 5사이에서 가장 큰 폭으로 인식률이 향상되었다. 그러나 연속분포 HMM에서 상태수가 7일때 오히려 인식률이 상태수 6보다 감소되는 현상이 나타났다. 상태수를 증가하면 학습 및 인식시간이 많이 소요되므로 시행착오적인 방법으로 최적인 HMM을 구성해야 한다.

4.3 세그먼트 차원압축에 의한 인식결과

모델은 5상태 4출력분포의 연속출력 분포형으로 하였다. 학습용 알고리듬, 모델의 초기치 추정 및 최대반복 횟수는 앞절의 4.2의 경우와 같다. 또 스펙트럼의 동적 특징량에 있어서 회귀계수를 이용한 경우 11 프레임폭에 대해서 구하였다.

각 방법에 대한 평균 음절인식률은 표 7과 그림 4에 나타내었다.

①MCEP

특징파라미터로서 프레임마다 멜켑스트럼을 이용한 경우.

②MCEP + △MCEP

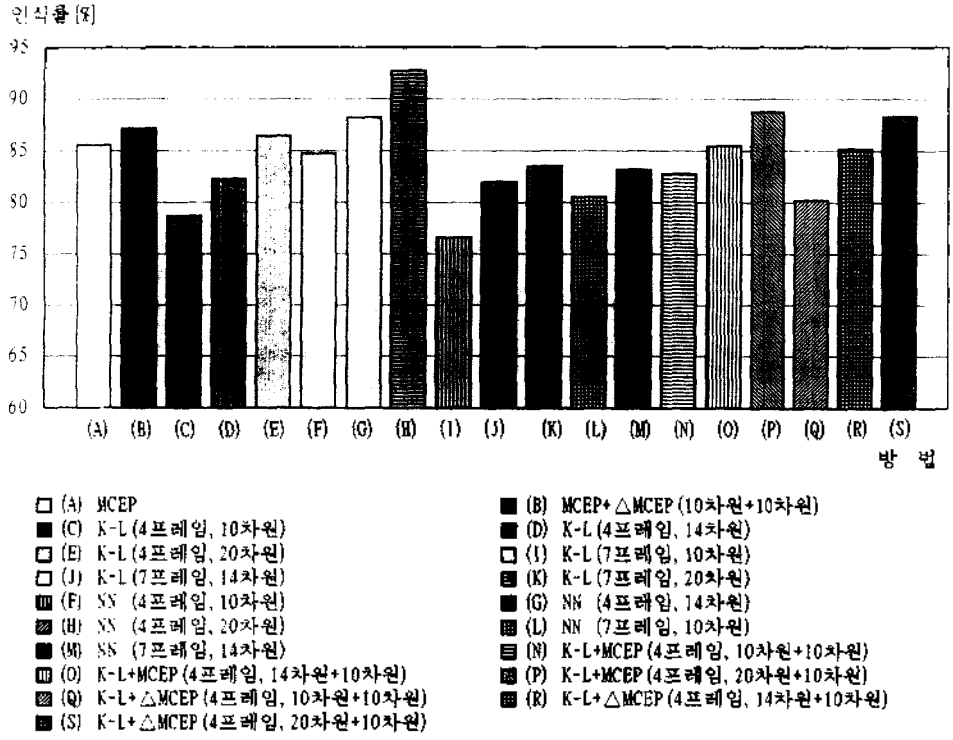


그림 4. 평균인식률의 비교

①에 동적특징량으로서 회귀계수를 부가한 경우.

③K-L (4프레임폭 세그먼트)

1세그먼트 40차원을 1프레임씩 이동하여 K-L전개로 10, 14, 20차원으로 압축한 것을 이용한 경우.

④K-L (7 프레임폭 세그먼트)

1프레임 건너서 4프레임을 결합하여 1세그먼트(40차원)로 하고 1프레임씩 이동하여 K-L전개로 10, 14, 20차원으로 압축한 것을 이용한 경우

⑤NN (4프레임폭 세그먼트)

1세그먼트 40차원을 1프레임씩 이동하여 신경망으로 10, 14, 20차원으로 압축한 것을 이용한 경우.

⑥NN (7프레임폭 세그먼트)

1프레임 건너서 4프레임을 결합하여 1세그먼트(40차원)로 하고 1프레임씩 이동하여 신경망으로 10, 14, 20차원으로 압축한 것을 이용한 경우.

⑦K-L (4프레임폭 세그먼트)＋MCEP

③의 압축백터에 멜켑스트럼을 부가한 경우.

⑧K-L (4프레임폭 세그먼트)＋△MCEP

③의 압축백터에 회귀계수를 부가한 경우.

실험결과 K-L전개로 10차원, 14차원 압축한 실험 ③의 경우는 인식률이 각각 실험 ①, ②의 경우(표 6의 상태수 5일 경우 참조)보다는 인식률이 저조하였다. 그러나 20차원으로 압축한 경우에는 실험 ①의 경우보다는 0.87% 향상되었지만 실험 ②의 경우보다는 인식률이 저조하였다.

실험 ④에서는 실험 ①, ②, ③의 경우보다 인식률이 저조하였는데 이것은 7프레임폭 세그먼트의 경우 1프레임씩 건너서 4프레임을 결합하여 1 세그먼트로 취하기 때문에 프레임간의 상관 관계가 떨어지기 때문이라고 생각된다.

실험 ⑤에서는 10차원으로 압축한 경우 실험 ③의 14차원 보다도 2.46% 향상되었지만, 20차원보다는 저조하

표 7. 세그먼트 차원압축에 의한 인식률 [단위 %]

방 법	세그멘트폭	차 원 수	평균인식율(%)
①MCEP	·	10차원	85.59
②MCEP＋△MCEP	·	10＋10차원	87.16
③④K-L	4프레임폭	10차원	78.72
		14차원	82.27
		20차원	86.46
	7프레임폭	10차원	76.63
		14차원	82.02
		20차원	83.52
⑤⑥NN	4프레임폭	10차원	84.73
		14차원	88.16
		20차원	92.70
	7프레임폭	10차원	80.52
		14차원	83.14
⑦K-L＋MCEP	4프레임폭	10＋10차원	82.79
		14＋10차원	85.48
		20＋10차원	88.75
⑧K-L＋△MCEP	4프레임폭	10＋10차원	80.20
		14＋10차원	85.18
		20＋10차원	88.29

였다. 그러나 14차원, 20차원의 경우는 실험 ①, ②, ③, ④의 경우보다 인식률이 향상되었다.

실험 ⑥에서는 실험 ①, ②, ③, ④에 비해서는 인식률이 향상되었다. 그러나 ⑤에 비해서는 인식률이 낮았는데 이것은 실험 ④와 상황과 동일한 이유이다.

K-L전개에 의한 실험 ③, ④의 인식률을 향상시키기 위하여 실험 ⑦, ⑧과 같이 K-L전개에 의한 압축벡터에 MCEP, △MCEP를 각각 부가한 결과 인식률이 향상되었다.

실험 ⑦에서는 실험 ①의 경우 보다 10차원의 경우 인식율이 2.8% 떨어졌지만, 14차원 압축인 경우와는 거의 비슷하였고, 20차원 압축인 경우에는 실험 ①, ②의 경우보다 각각 3.16%, 1.59% 향상되었다.

실험 ⑧에서는 실험 ①의 경우보다 10차원, 14차원 압축인 경우에 각각 5.39%, 0.41%로 인식률이 떨어졌으나 20차원 압축인 경우에는 2.7% 향상되었다. 또, 20차원 압축인 경우는 실험 ⑦의 20차원의 경우보다 인식률이 48% 떨어졌다.

그리고 신경망으로 압축하면 10차원, 14차원에서 압축오차가 K-L전개보다 크지만 비선형 변환에 의한 압축이므로 음성의 동적 특징량이 잘 반영되어 K-L전개 및 다른 방법들에 비해서 인식률이 향상되었다고 생각된다. 일반적으로 모음이 같은 음절이 다른 음절로의 오인식이 많았는데 이것은 구간 길이가 짧은 자음(초성 또는 종성)이 긴 구간을 갖는 모음의 영향을 받아 자음이 무시되기 때문이라고 생각된다.

V. 결 론

기본 HMM은 구조 결정에 있어서 시행착오에 대한 의존성이 높고, 학습시 많은 데이터가 필요하며, 음성의 과도적 정보를 경시하는 경향이 있고, 단순 Markov과정으로 가정하기 때문에 패턴의 시간적 상관에 대한 표현력이 부족하다.

이를 보완하기 위한 방법으로 첫번째는 이산 지속시간 제어를 한 경우, 동적특징량으로 회귀계수를 부가한 경우, 혼합분포를 사용한 경우 등을 각각 연속분포 HMM의 파라미터로 부가하여 상태수를 늘이면서 인식실험을 수행하였으며 그 결과를 연속분포 HMM(MCEP 사용)과 비교하였다.

실험결과 연속분포 HMM의 인식률 85.59%보다 동적특징량으로 회귀계수를 부여한 경우 1.4%, 혼합분포를 사용한 경우 2.36%, 이산 지속시간제어를 한 경우 2.78%로 각각 향상되었다. 일반적으로 상태수가 증가하면 인식률은 향상되지만 연속분포 HMM의 경우 상태수가 7일때 상태수 6보다 오히려 인식률이 감소되었다. 이처럼 상태수가 증가할 경우 학습 및 인식 시간이 많이 소요되면서 인식률이 오히려 감소되는 경우도 있으므로 시행착오적인 방법으로 최적인 HMM을 구성해야 한다.

두번째로는 음성의 동적 변화량을 반영하기 위하여 전 음절 구간에 대해서 인접한 4 프레임을 결합한 4 프레임폭 세그먼트와 1 프레임씩 건너서 4 프레임을 결합한 7 프레임폭 세그먼트를 구성하여 40차원의 특징파라미터를 K-L전개(선형변환)와 신경망(비선형변환)으로 각각 10, 14, 20차원으로 압축하였다. 학습 및 인식실험은 MCEP만을 이용한 경우, MCEP+△MCEP한 경우, K-L전개에 의한 압축벡터를 이용한 경우, 신경망에 의한 압축벡터를 이용한 경우, K-L전개에 의한 압축벡터에 MCEP과 △MCEP를 각각 파라미터로 부가한 경우에 대해서 비교 실험을 수행하였다.

7프레임폭 세그먼트를 K-L전개 및 신경망으로 압축한 경우는 4프레임폭 세그먼트에 비해 프레임간의 상관 관계가 떨어지기 때문에 인식률이 저하되었다. 이것을 보완하기 위해 K-L전개에 의한 압축벡터에 MCEP과 △MCEP를 각각 파라미터로 부가한 결과 향상된 인식률을 얻을 수 있었다.

그리고 신경망으로 압축한 벡터만을 이용한 경우 K-L전개에 비해 10, 14차원에서 압축오차가 크게 나타났지만, 신경망의 학습횟수를 증가시키면 압축성능이 더욱 향상될 수 있을 것이다. 또한 신경망은 비선형 변환인 시그모이드 함수를 사용하여 압축을 수행하므로 선형변환인 K-L전개보다 음성의 동적변화를 잘 반영하고 인식실험에서도 K-L전개 및 다른 방법에 비해 향상된 인식결과를 얻을 수 있었다.

향후 단어나 연속음성에서 세그먼트 차원압축으로 차원을 감소하여 이를 특징파라미터로 사용한다면 음성의 동적변화량이 잘 반영되어 음성인식의 향상을 기대할 수 있을 것이다.

참 고 문 헌

1. L. R. Rabiner, "On the application of vector quantization and hidden Markov models to speaker independent, isolated word recognition," Bell Syst. Tech. J., 62, pp. 1075-1105, 1983.
2. L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," AT&T Tech. J, Vol. 64, pp. 1211-1234, July-Aug. 1985.
3. A. waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang : "Phoneme Recognition : Neural Network vs. Hidden Markov Models," IEEE, S3.3, 1988.
4. K-F.Lee and H-W.Hon, "Large vocabulary speaker-independent continuous speech recognition using HMM," Proc. ICASSP, pp. 123-126, 1988.
5. E. Frangoulis, "Vector quantization of the continuous distributions of a HMM speech recognizer based on mixtures of continuous distributions," Proc. ICASSP-

90, pp. 9-12, 1990.
6. 中川聖一, "確率モデルによる音声認識," 電子情報通信學會編, 1989.
7. 中川聖一, "連續出力分布型HMMによる日本音韻認識," 音響學會論文誌, Vol. 46, pp. 486-496, 1990.
8. 船橋賢一, "3層ニューラルネットワークによる恒等寫像の近似的實現についての理論的考察," 信學論, J73 A, 1, pp. 139-145, 1990.
9. 上坂吉則 : パターン認識と學習のアルゴリズム, 文一統合出版, 1994.
10. 박창호, 허강인, "음성 인식을 위한 HMM의 파라미터 확장에 관한 연구," 제11회 음성통신 및 신호처리 워크샵 논문집, Vol. 11, No. 1, pp. 152-156, 1994.
11. 박창호, 이영재, 허강인, "세크멘트 통계량을 이용한 HMM의 한국어 음절 인식," 제12회 음성통신 및 신호처리 워크샵 논문집, Vol. 12, No. 1, pp. 175-178, 1995.
12. 김주성, 박창호, 허강인, 안점영, "신경망의 차원압축 능력을 이용한 음절 인식," 제8회 신호처리합동학술대회 논문집, Vol. 8, No. 1, pp. 48-51, 1995.

▲金 柱 聲(Joo Sung Kim) 1967년 11월 23일생

1990년 2월 : 동의대학교 전자공학과(공학사)
1993년 2월 : 동아대학교 대학원 전자공학과(공학석사)
1994년 3월~현재 : 동아대학교 대학원 전자공학과 박사과정
1994년 3월~현재 : 동의대학교 공과대학 전자공학과 강사
※주관심분야 : 음성인식·합성, 신경회로망

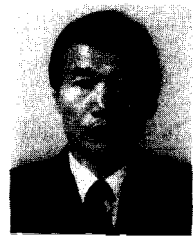
▲李 良 雨(Yang Woo Lee) 1948년 10월 14일생

1974년 2월 : 부산대학교 전기공학과(공학사)
1978년 2월 : 부산대학교 대학원 전기공학과(공학석사)
1988년 2월 : 부산대학교 대학원 전기공학과(공학박사)
1987년 3월~현재 : 동의대학교 공과대학 전기공학과 부교수

▲許 康 仁(Kang In Hur) 1955년 2월 20일생

1980년 2월 : 동아대학교 전자공학과(공학사)
1982년 2월 : 동아대학교 대학원 전자공학과(공학석사)
1990년 8월 : 경희대학교 대학원 전자공학과(공학박사)
1984년 9월~현재 : 동아대학교 공과대학 전자공학과 교수
1988년 9월~1989년 8월 : 일본 筑波大學 객원연구원
1992년 9월~1993년 8월 : 일본 豊橋大學 객원연구원
※주관심분야 : 디지탈신호처리, 음성인식·합성, 신경회로망

▲安 点 榮(Jeom Young Ahn) 1942년 7월 13일생



1964년 2월 : 한국항공대학교 항공전자공학과(공학사)
1979년 2월 : 동아대학교 대학원 전자공학과(공학석사)
1986년 8월 : 동아대학교 대학원 전자공학과(공학박사)
1987년 3월~현재 : 동의대학교 공과대학 전자공학과 교수