論文97-34S-4-7

# 다중 관측열을 토대로한 HMM에 의한 음성 인식에 관한 연구

# (A Study on the Speech Recognition by HMM based on Multi-Observation Sequence)

鄭 義 鵬 *

(Jeoung Eui-Bung)

요 약

본 논문은 단독어 인식에 관한 연구로써, 다중관측열(multi-observation sequence)을 토대로 하는 HMM(hidden Markov model) 방법을 제안한다. 제안된 연구 방법은 각 단어를 몇 개의 구간(section)으로 나누어 MSVQ 코드북(codebook)을 만들고, 학습 데이타를 몇개의 구간으로 나누어, 거리값이 가까운 순으로 가중치를 주어서 각 구간별로 다중 관측열을 구한다. 그런 다음, 이 관측열을 구간별로 학습시키고, 인식실험때 확률값이 가장 높은 것을 인식된 단어로 한다. 전국의 146개 DDD 지역명을 인식대상단어로 선정하였고 특징 파라메타로 10차 LPC 캡스트럼(cepstrum) 계수를 사용하였다. 제안된 음성인식 실험방법 외에도 비교를 위해서 같은 조건상에서 같은 데이타 로 DP방법과 MSVQ 및 일반적인 HMM 인식 실험을 수행하였다. 실험 결과, 본 논문에서 제안한 다중관측열을 이용한 HMM이 DP 방법, MSVQ 방법 및 일반적인 HMM 모델보다 인식률과 인식 시간에서 우수하였다.

**Abstract**

The purpose of this paper is to propose the HMM(Hidden Markov Model) based on multi-observation sequence for the isolated word recognition. The proposed model generates the codebook of MSVQ by dividing each word into several sections followed by dividing training data into several sections. Then, we are to obtain the sequential value of multi-observation per each section by weighting the vectors of distance from lower values to higher ones. Thereafter, this the sequential value of multi-observation of each section is trained and this model is considered to recognize a word with high probability value while in recognition. 146 DDD area names are selected as the vocabularies for the target recognition, and 10 LPC cepstrum coefficients are used as the feature parameters. Besides the speech recognition experiments by way of the proposed model, for the comparison with it, the experiments by DP, MSVQ, and general HMM are made with the same data under the same condition.The experiment results have shown that HMM based on multi-observation sequence proposed in this paper is proved superior to any other methods such as the ones using DP, MSVQ and general HMM models in recognition rate and time.

## I. 서 론

본 논문은 중규모 어휘의 음성 다이얼링 (voice dialing) 시스템 개발을 위하여 146개의 DDD 지역명을 인식 대상으로한 한국어 단독어 인식에 관한 연구를 행한다. 그런데, 중규모 이하의 단어에 있어서는 음소나 음절 단위와 같이 부단어(subword) 단위로 음성 인식을 행하는 것보다 단어 단위로 음성 인식을 행하는 것이 더 실용적이라 할 수 있다. 이와 같이 단어 단위의 음성 인식을 하는데 있어서, 기존의 DP에 의한

* 正會員, 全北 産業大學校 情報通信工學科
(Dept. of Information & Telecommunication Eng., CHUNBUK SANUP University)

接受日字: 1996年9月23日, 수정완료일: 1997年4月10日

인식[1]은 인식 시간이 너무 길고 기억 용량도 크며, VQ[2-3]나 MSVQ[4-5]는 인식 시간 및 기억 용량은 적으나, 인식률이 떨어진다. 따라서, 다른 방법보다 인식률 및 계산 시간 등이 우수한 모델로 HMM 모델을 선정하여, 앞으로 다가올 인간과 기계와의 통신 시대에 대비하여 기계가 우리 인간의 자연어를 자연스럽게 인식할 수 있도록 하기 위한 일환으로 단독어 인식 시스템이 개발되고 있다. 그러나, 일반적인 HMM[6-9]의 경우도 모델을 작성하는데 많은 화자의 많은 학습용 데이타가 필요하며, 모델 학습시에 참여하지 않은 화자의 음성은 인식률이 현저하게 떨어질 수 있는데 본 연구에서는 관측열(observation sequence)을 구할 시에 다중 관측열(multi-observation sequence)를 이용하여 HMM 학습에 이용함으로 이런 단점을 보완하였다. 이와 함께 인식률을 향상시키고 또한 시간 단축을 하기 위해 MSVQ 코드북에 의해 다중 관측열을 구하고, 다중 관측열을 이용하여 다중구간(Multi-Section)에 의한 HMM 모델을 학습하며, 인식시에 구간마다 확률값을 비교하여 인식 대상 단어의 후보 수를 줄이는 HMM 모델을 제안한다.

기존의 HMM이 코드워드의 벡터값과 미지의 단어의 어떤 프레임의 벡터간의 거리값이 가장 작은 것 하나를 택하는데 반해, 제안된 다중 관측열에 의한 HMM 모델의 다중 관측열을 구하는 방법은 거리값이 짧은 것 몇개를 선택해서 거리값이 작은 순으로 적당한 가중치를 주는 방법이다. 따라서, 기존의 HMM에서는 학습시에 가장 거리값이 작은 것으로 선택되지 않아 확률값을 가지지 못하는 것이, 다중 관측열에 속하게 되면 어느 정도 확률값을 가질 수 있으므로, 인식시에 이것이 선택되며 확률값을 가지므로 인식률을 향상시킬 수 있다.

또한, HMM에서도 모델을 학습시키고 인식시키는데 있어서, VQ 음성 인식에 시간 정보를 포함시킨 MSVQ 음성 인식이 인식 시간을 줄여 주고, 인식률을 높여주는 것처럼, HMM에도 구간을 나누어 줌으로써 인식 시간을 줄여주고 인식률을 향상시킨 다중구간의 HMM을 이용해 인식 실험한다.

본 연구에서는 ZCR(zero crossing rate)과 에너지(energy)로 끝점 검출을 행했으며, 특징 파라메타는 LPC 캡스트럼 계수를 사용하였다.

## Ⅱ. 제안된 음성 인식 시스템

일반적인 이산 HMM의 경우는 각 열마다 하나의 심볼만 관측되는 것으로 생각하였는데 본 연구에서는 각 열마다 몇개의 다중 심볼이 적당한 가중치를 가지고 한 열을 구성하는 다중 관측열을 사용하였다. 그 이유는 HMM 학습시에 선택되지 않은 심볼은 학습 과정에서 확률값을 가질 수 없는 관계로 인식 실험시 학습에 참여하지 않는 심볼이 나올 경우에는 인식이 거의 불가능하기 때문이다. 따라서, 다중의 관측열을 둠으로써 같은 열에서 몇개의 관측될 심볼에 속하면 확률값을 가질 수 있으므로 인식시에 어느 정도의 확률값을 가져 인식될 가능성을 높여준다.

또한, 일반적인 HMM에 의한 단독어 인식에 있어서는 VQ 코드북을 작성한 후 이 코드북에 근거하여 관측열을 구하는 관계로 시간이 많이 걸리며, 시간의 변화에 대한 정보를 포함하지 못하여 인식률을 낮추는 경우가 종종 있는데 본 연구에서는 이를 극복하기 위해 MSVQ 코드북을 작성한 후 학습을 통해 모델을 만들 때에도 몇 개의 구간으로 나누어서 모델을 학습시켰으며, 인식시에도 몇개의 구간으로 나누어 인식을 시켰다. 이 때, 이 모델에 의해 인식을 행하면, 인식 시간이 거의 두배 정도 빨라지고, 인식률도 높아진다. 본 연구에서는 HMM 인식시 인식 시간을 줄이기 위해 각 구간마다 인식이 끝나면, 확률값을 비교하여 높은 순으로 인식 대상 중 몇개의 후보를 선택한다. 이것을 반복하여 최종 구간에서 하나의 단어만 선택하여 인식된 것으로 생각한다.

따라서, 본 연구에서는 다중 관측열을 이용한 다중구간의 HMM 음성 인식 시스템을 제안한다.

### 1. 기본적인 이론

본 연구에서 제안한 모델을 이용한 음성 인식을 하는데 있어서 음성 신호를 관측열로 변환하는데는 일반적인 HMM과 같이 먼저 음성의 전처리 과정, 특징 추출, 그리고 데이타 압축등을 거쳐서 이루어진다. 이런 분석이 끝난 후에 본 연구에서는 MSVQ 코드북을 작성하고 다중관측열 개념과 다중구간 개념을 이용해서 HMM 모델 및 인식 알고리즘을 세운다.

#### 1) MSVQ 이론

Burton[5]의 MSVQ 코드북을 이용한 음성 인식에 따르면 MSVQ 코드북 작성시에는 시간 정보를 이용

하는 관계로 단일 VQ를 이용한 음성 인식 방법보다 인식률 뿐만 아니라 인식 시간도 단축됨을 알 수 있다. 따라서, 본 연구에서는 HMM 모델에도 시간 감축 및 시간 정보를 이용하기 위해 MSVQ에 의한 코드북 작성에 따른 HMM 모델에 의한 인식 실험을 수행한다.

전체 단어에 대한 MSVQ 코드북은 그 단어를 몇개의 구간으로 나누고 각 구간 마다 집단화 알고리즘을 이용하여 작성한다. 즉, 2 MSVQ 코드북은 구간을 2개로 나누어 코드북을 만들었으며, 각 구간별 코드북은 각 특징 벡터마다 128개의 코드워드로 이루어지며, 전체 코드워드의 수는 256개로 한다.

또한, 4 MSVQ 코드북은 구간을 4개로 나누어 코드북을 만들었으며, 각 구간별 코드북은 각 특징 벡터마다 128개의 코드워드로 이루어지며, 전체 코드워드의 수는 512개로 한다.

2) 다중 관측열 나열

같은 음성에서도 발성 속도의 차이에 따라 음성의 길이가 일정하지 않다. 같은 사람이 같은 말을 하여도 그 때마다 그 길이가 바뀌어진다. 물론 똑같은 말이라도 발성하는 사람이 다르며 길이의 변동 역시 크다. 또한 발성 기관의 크기는 인간에 따라 달라 같은 형태로 하여 발성하여도 공진 주파수에 차이가 생긴다. 이것이 패턴상의 개인성이 되어 나타난다. 이것이, 음성 인식을 하는데 어려움을 준다.

그래서, 본 연구에서는 다중관측열과 다중구간 개념을 이용하여 이런 특성을 극복하려 한다. 어떤 발음을 했을 때 발음할 때마다 또는 화자에 따라 특성이 달라지지만 그 발음이 가지는 특성이 많이 벗어나지는 않을 거라는데 유의하여 다중관측열을 이용하였다.

즉, 각 프레임의 벡터와 VQ 코드북의 코드워드 중 거리값이 가장 짧은 것을 심볼로 선택하는 것이 일반적인 HMM에서 관측열을 구하는 방법이었다. 그런데, 본 연구에서는 일반적인 HMM에서 관측열을 구하는 것과는 달리 각 열마다 몇개의 유사한 특징을 가지는 심볼을 선택하는 방법을 제안한다.

이 방법은 일반적인 HMM이 학습할 때 관측열로 선택되지 않은 벡터는 확률값이 0으로 떨어지는데 반해 확률값이 0으로 떨어질 그런 벡터의 경우도 몇개의 다중 심볼 안에 들면 확률값을 가질 수 있으므로 인식시에도 심볼로 선택될 확률을 크게 가질 수 있다는 개념에서 비롯되었다.

실험에서 사용된 다중 관측열을 구하는데 사용한 가중치 법칙은 다음과 같다.

$$w_m = \frac{N-m+1}{\sum_{n=1}^{N} n} \quad (1)$$

여기서, m은 어떤 프레임의 벡터와 코드북의 각 코드워드 중 거리값이 작은 순으로 표시했을 때 몇 번째인지를 나타내고, N는 한 열에서 선택될 다중 심볼의 수를 의미하며, $w_m$는 전체 확률을 1로 했을 때 거리값이 m 번째인 심볼이 가질 확률값이다.

따라서, 식 (1)에 의해 구하면, 관측열의 다중 심볼수 N이 2인 경우는 $w_1 = 2/3$, $w_2 = 1/3$을 가지며, 3의 경우는 $w_1 = 3/6$, $w_2 = 2/6$, $w_3 = 1/6$로 구해져 전체 확률값이 1이 되는데 4이상의 경우도 윗 식 (1)에 의해서 전체 확률값 1이 되도록 구할 수 있다

2. 제안된 방법에 의한 음성 인식

HMM은 이중의 확률 처리로써 하나는 현재의 천이가 이루어질 천이 확률이고, 또 하나는 천이가 이루어졌을 때 유한개의 관측 대상으로부터 각 출력 심벌이 관측되는 조건부 확률을 규정하는 출력 확률 밀도 함수인데, 본 연구에서는 각 출력 심볼에 다중관측열 개념을 도입하여 가중치를 줌으로서 인식률의 향상은 물론 학습 데이타의 수를 줄일 뿐 아니라 또한 전 음성 부분을 몇개의 구간으로 나누어줌으로써 시간 정보를 주어 인식률을 높이는 한편 인식 시간을 줄이며, 더 나아가 각 구간마다 인식 대상 후보의 수를 줄여줌으로써 인식 시간 및 인식률을 증진시키는 방법에 대해서 제안한다.

1) 제안된 다중관측열을 이용한 HMM의 원리

본 연구에서 제안하고 있는 다중관측열 개념을 도입한 음성 인식 방법에서 사용되는 기호는 다음과 같다.

상태수 : N
전체 심볼수 : M
열의 관측될 심볼수 : S
상태 집합 : $Q = \{ q_1, q_2, ..., q_N \}$
심볼 집합 : $V = \{ v_1, v_2, ..., v_M \}$
관측열의 길이 : t = 1, 2, ..., T
t번째 관측 심볼열이 상태 $q_i$에 있고, t+1 번째 관측 심볼열이 상태 $q_j$를 선택할 확률

$$A = \{a_{ij}\}, \; a_{ij} = pr(q_j \text{ at } t+1 \mid q_i \text{ at } t) \quad (2)$$
$$(1 \le i,j \le N)$$

t번째 관측 심볼열이 $q_j$ 상태에서 다중 심볼 집합 $\{v_{k1}, v_{k2},..., v_{ks},..., v_{kS}\}$을 가지고 그 때의 각각의 심볼이 가질 가중치 집합을 $\{w_1, w_2,..., w_s,..., w_S\}$ 이라 하면

$$\sum_{s=1}^{S} w_s = 1 \tag{3}$$

이다. 따라서, t번째 관측 심볼열이 $q_j$ 상태에서 다중 심 볼 $v_k = \{v_{ks}\}$를 선택할 확률

$$\begin{aligned} \overset{S}{B}_{s=1} &= \{b_j(k)\},\ b_j(k) = \sum w_s b_{js}(k_s) \\ &= pr(v_k \text{ at } t \mid q_j \text{ at } t) \\ &= pr(\{v_{ks}\} \text{ at } t \mid q_j \text{ at } t) \\ &(1 \le j \le N),\ (1 \le k \le M),\ (1 \le s \le S) \end{aligned} \tag{4}$$

초기 상태에서 상태 $q_i$ 를 선택할 확률

$$\pi = \{\pi_i\},\ \pi_i = pr(q_i \text{ at } t = 1) \tag{5}$$

관측열 $O = O_1, O_2, ..., O_T$

관측 심볼의 구성 $O_t = \{ o_{t1}, o_{t2}, ..., o_{tS} \}$

이상의 정의를 이용한 모델은 $\lambda = (A,B,w,\pi)$로 표시할수 있는데 이 모델을 실제 응용하는데는 모델 작성시에 사용되는 알고리즘과 인식시에 사용되는 알고리즘으로 나누어 생각 할 수 있다.

1. 모델 학습 알고리즘

모델 학습에 앞서 생각 할 문제는 모델 파라메타 $\lambda = (A,B,w,\pi)$가 주어졌을때 관측열 $O = O_1, O_2, ..., O_T$의 계산 문제이다. 가장 단순하게 계산하는 방법으로는 가능한 모든 상태열에 대해여 A, B 행렬을 이용하여 확률을 계산하는 것이다. 상태열은 $I = i_1, i_2, ..., i_T$ 일경우

$$Pr(O_t \mid q_i \text{ at } t) = \sum_{s=1}^{S} w_s b_{is(t)}(O_{ts}) \tag{6}$$

$$Pr(O \mid I, \lambda) = b_{i(1)}(O_1) b_{i(2)}(O_2) ... b_{i(T)}(O_T) \tag{7}$$

$$= \left[\sum_{s=1}^{S} w_s b_{is(1)}(O_{1s})\right] ... \left[\sum_{s=1}^{S} w_s b_{is(T)}(O_{Ts})\right]$$

$$Pr(I \mid \lambda) = a_{i(1)}\ a_{i(1)i(2)}\ a_{i(2)i(3)} ... a_{i(T-1)i(T)} \tag{8}$$

이다. 따라서,

$$\begin{aligned} Pr(O \mid \lambda) &= \sum_{\text{all } I} Pr(O,I \mid \lambda) \\ &= \sum_{\text{all } I} Pr(O \mid I, \lambda) * P(I \mid \lambda) \\ &= \sum_{i_1,...,i_T} \pi_{i(1)} b_{i(1)}(O_1) a_{i(1)i(2)} b_{i(2)}(O_2) ... a_{i(T-1)i(T)} b_{i(T)}(O_T) \\ &= \sum_{i_1,.,i_T} \pi_{i(1)} \left[\sum_{s=1}^{S} w_s b_{is(1)}(O_{1s})\right] .. a_{i(T-1)i(T)} \left[\sum_{s=1}^{S} w_s b_{is(T)}(O_{Ts})\right] \end{aligned} \tag{9}$$

로부터 구할 수 있다.

좀더 능률적인 계산방법인 L. E. Baum에 의해 제안된 전향-후향 알고리즘[7]을 수정한 알고리즘을 간단히 살펴보자.

먼저 전향 변수 $\alpha_t(i)$을 살펴보면

$$\begin{aligned} \alpha_t(i) &= Pr(O_1, O_2,..., O_t, i_t = q_i \mid \lambda) \\ &(O_t = \{ o_{t1}, o_{t2}, ..., o_{tS} \}) \end{aligned} \tag{10}$$

로 정의 한다.

이는 주어진 모델 $\lambda$에 대해서 시간 t에서 관측열이 $O_1,O_2,...,O_t$이고 상태가 $q_i$일 확률이며 다음과 같은 절차에 의해서 구할 수 있다.

단계 1. 초기화

$$\begin{aligned} \alpha_1(i) &= \pi_i b_i(O_1) = \sum_{s=1}^{S} \pi_i\ w_s b_{is}(O_{1s}) \\ &, 1 \le i \le N \end{aligned} \tag{11}$$

단계 2. t = 1, 2,.....T-1 에 대해 반복$(1 \le i,j \le N)$

$$\alpha_{t+1}(j) = \sum_{s=1}^{S} \left[\sum_{I=1}^{N} \alpha_t(i) a_{ij}\right] w_s b_{js}(O_{(t+1)s}) \tag{12}$$

단계 3. 그러면

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_{T(i)} \tag{13}$$

같은 방법으로 후향 변수를 살펴 보면,

$$\begin{aligned} \beta_t(i) &= Pr(O_{t+1}, O_{t+2}, ..., O_{T-1}, O_T \mid i_t = q_i, \lambda) \\ &(O_t = \{ o_{t1}, o_{t2}, ..., o_{tS} \}) \end{aligned} \tag{14}$$

로 정의할 경우 다음 절차에 의해 구할 수 있다.

단계 1. 초기화 $\beta_T(i) = 1,\ 1 \le i \le N$ (15)

단계 2. t = T-1, T-2, ..., 1에 대해 반복$(1 \le i,j \le N)$

$$\beta_t(i) = \sum_{i=1}^{N} a_{ij}\ b_j(o_{t+1})\ \beta_{t+1}(j) \tag{16}$$

$$= \sum_{I=1}^{N} \sum_{s=1}^{S} a_{ij}\ w_s b_{js}(O_{(t+1)s})\ \beta_{t+1}(j)$$

이다. 일반적인 HMM에서 초기 파라메타들로 부터 $Pr(O|\lambda)$를 최대로 하는 $\lambda = (A,B,w,\pi)$를 재추정하는 것으로 Baum-Welch 재추정 알고리즘[7]을 이용하는데, Baum-Wetch의 재추정 알고리즘도 전향-후향 알고리즘과 마찬가지로 일반적인 알고리즘에서 수정하여 사용할 수 있다. 주어진 관측열과 모델 $\lambda$에 대해서 시간 t에서의 상태가 $q_i$이고 시간 t+1에서의 상태가 $q_j$일 확률 $\xi_t(i,j)$를 다음과 같이 정의한다.

$$\xi_t(i,j) = Pr(i_t = qi,\ i_{t+1} = q_j \mid O, \lambda) \qquad (17)$$

$$= \frac{\sum_{s=1}^{S} \alpha_t(i) a_{ij}\ w_s b_{js}(O_{(t+1)s})\ \beta_{t+1}(j)}{Pr(O|\lambda)}$$

$$(1 \le i,j \le N)$$

또한 주어진 관측열과 모델 $\lambda$에 대해서 시간 t에서 상태가 $q_i$일 확률은 다음과 같다.

$$\gamma_t(i) = Pr(i_t = q_i \mid O, \lambda) \qquad (18)$$

여기서 현재 모델 $\lambda = (A,B,w,\pi)$이라 하고, 다시 추정된 모델은 $\lambda = (\hat{A},\hat{B},\hat{w},\hat{\pi})$이라 하면 새로운 모델 파라메타는 아래의 식들과 같다.

$$\hat{\pi}_i = \gamma_1(i),\ 1 \le i,j \le N \qquad (19)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad (20)$$

$$= \frac{\sum_{t=1}^{T-1} \sum_{s=1}^{S} \alpha_t(i) a_{ij} w_s b_{js}(O_{(t+1)}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)}$$

$$(1 \le i,j \le N)$$

$$\hat{b}_j(k) = \frac{\sum_{t=1,O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} = \frac{\sum_{t=1,O_t=v_k}^{T} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \qquad (21)$$

$$(1 \le i,j \le N)$$

2. 인식 알고리즘

인식에 사용되는 알고리즘으로는 전향 알고리즘과 후향 알고리즘 외에 Viterbi 알고리즘[7]이 있다. Viterbi 알고리즘은 관측열 $O = O_1O_2,...,O_T$가 주어졌을때 최적의 상태열 $I= i_1,i_2, ...,i_T$를 구하는 알고리즘이다. 전향 및 후향 알고리즘은 앞에서 언급하였으므로 여기서는 Viterbi 알고리즘에 대해서 설명한다.

단계 1. 초기화

$$\delta_1(i) = \pi_i b_i(O_1) = \sum_{s=1}^{S} \pi_i\ w_s b_{1s}(O_{1s})$$

$$, 1 \le i \le N \qquad (22)$$

$$\Psi_1(i) = 0 \qquad (23)$$

단계 2. $2 \le t \le T$, $1 \le j \le N$ 에 대해 반복

$$\delta_t(j) = \max_{1\le i\le N} \sum_{s=1}^{S} [\delta_{t-1}(i)a_{ij}]\ w_s b_{js}(O_{ts}) \qquad (24)$$

$$\Psi_t(j) = \operatorname{argmax}_{1\le i\le N} [\delta_{t-1}(i)a_{ij}] \qquad (25)$$

단계 3. 종료

$$P^* = \max_{1\le i\le N} [\delta_T(i)] \qquad (26)$$

$$i^*_T = \operatorname{argmax}_{1\le i\le N} [\delta_T(i)] \qquad (27)$$

단계 4. 경로(상태열) 백트랙킹

$$t = T-1,\ T-2,\ ....,\ 1 \text{ 에 대해 반복}$$

$$i^*_t = \Psi_{t+1} i^*_{t+1} \qquad (28)$$

본 연구에서는 논문 (8)에 실험 결과에 의해 인식 알고리즘으로 전향 알고리즘을 사용하였다.

2) 제안된 모델 작성 및 인식 시스템의 구조.

다중 관측열을 이용한 HMM은 모델을 작성하기에 앞서 일반적인 HMM이 VQ 코드북에 의해 관측열을 구하는데 비해 제안된 모델은 MSVQ 코드북을 구한 뒤, 학습용 데이타들도 같은 수의 다중 구간으로 나누어 주어서 그 구간들끼리만 HMM 학습을 시키는 것이다. 이에 대한 순서도는 그림 1에 나타내었다.

또한 인식시에도 마찬가지로 MSVQ와 같은 수의 다중 구간으로 나누어 준 후 이를 바탕으로 인식을 시킨다. 이 때, 구간별로 확률값을 측정할 수 있는 관계로 구간마다 후보수를 줄여줌으로써 상당히 빠르게 인식할 수 있다. 이에 대한 순서도는 그림 2에 나타내었다.

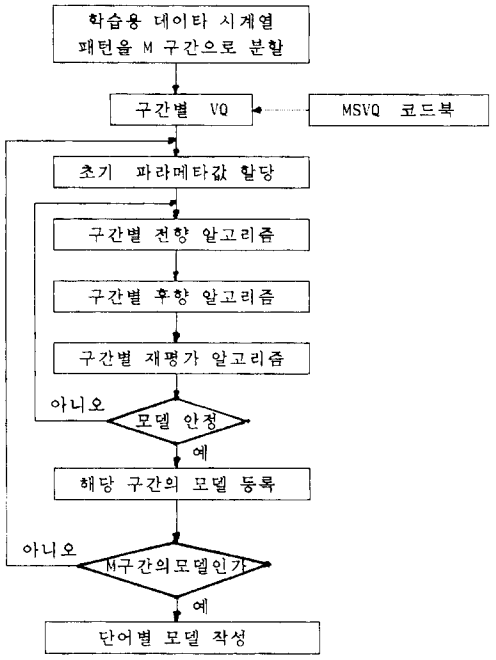



그림 1. 제안된 모델 작성 방법
Fig. 1. The method of proposed model generation.

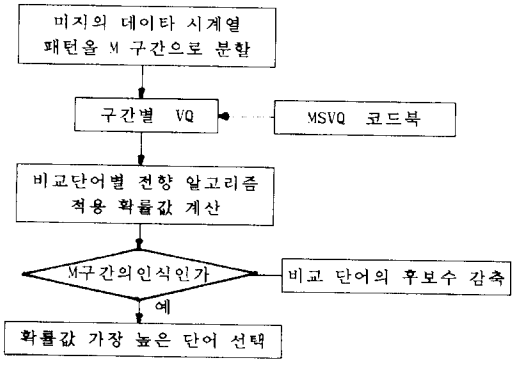



그림 2. 모델의 인식 방법
Fig. 2. The recognition method of model.

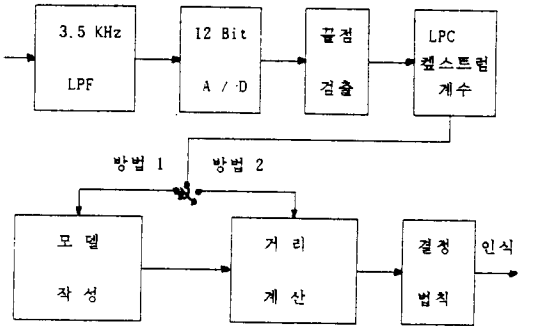
본 모델에서도 일반적인 HMM과 마찬가지로 세가지 문제점이 대두된다. 첫째는 모델의 파라메타를 어떻게 초기화할 것인가하는 점이고, 둘째는 T가 증가함에 따라 이들의 값이 지수함수적으로 감소하여 곧 언더플로우를 발생시킨다는 점이며, 세째는 학습 데이타의 양이 충분이 많지 않을 경우 학습 과정에서 다중 관측열에 의해 학습시키더라도 어떤 심볼이 모델의 어느 상태에는 나타나지 않는것으로 추정되었는데 실험 과정중에는 이 심볼이 나타나는 경우가 있다는 것이다. 이런 문제를 논문 (6)에 의해 해결하였으며, 본 연구에서는 천이 가능수가 2인 left-to-right 모델을 사용하였다.

## Ⅲ. 인식 실험 결과

본 연구는 인식 실험 대상어휘로 146개 DDD 지역명을 선정하였고, 8명 중 5명의 남성 화자에 의해 2번씩 발성한 것으로 모델을 작성하였으며, 나머지 3명이 2번씩 발음한 것으로 인식 실험을 수행하고 다른 인식 방법들과 비교하였다.

### 1. 인식 시스템 구성

본 연구에서 실험한 음성 인식 시스템은 그림 3와 같다.




방법 1 : 학습용 데이타
방법 2 : 시험용 데이타

그림 3. 인식 시스템의 구성도
Fig. 3. Block diagram of recognition system.

실험에 사용된 모든 DDD 지역명 데이타는 잡음이 섞인 환경하에서 릴 테이프에 의해 입력된 음성 신호로써, 샘플링 주파수를 8 KHz로 하였으며, 3.5 KHz의 저역 여파기를 통과한 후 12 Bit A/D 변환기를 거쳐 음성 데이타를 구하였다. 그리고, 시작점과 끝점을 검출한 후, LPC 계수보다 LPC 켑스트럼 계수를 특징 벡터로 사용하는 것이 더 좋다는 고찰[10]에 따라 특징 파라메타로는 LPC 켑스트럼을 사용하였으며, 이것을 이용하여 각 단어의 모델을 구하고, 이 모델을 기초로 인식 실험을 행했다.

### 2. 기존의 다른 방법에 의한 실험 결과

실험하는데 있어서, DP 패턴 매칭의 경우, 표준 패

턴을 선택하는 방법은 집단화 알고리즘을 사용하였고, 소구간 경로 제약 및 전체 경로 제약은 Sakoe와 Chiba의 방법[1]을 사용하였으며, MSVQ에 의한 음성 인식의 경우[5]는 구간 수를 8로 하여 실험하였으며, 각 구간 별 코드북의 크기는 4로 하여 총 32개의 코드워드를 택하여 실험하였다. 또한, 일반적인 HMM의 경우는 256개의 코드워드(심볼)을 주었으며, 상태수는 8로 하였다.

이 세가지 방법에 의한 인식 실험 결과는 표 1과 같으며, 이 중에서 HMM이 제일 인식률이 좋은 것으로 나타났다.

표 1. 기존의 방법에 의한 실험 결과
Table 1. Recognition result by classical method.

단위 : % (개)

| 인식률 / 화자 | DP 인식 방법 | MSVQ 인식 방법 | HMM 인식 방법 |
|---|---|---|---|
| 화 자 1 | 81.85 (239) | 75.34 (220) | 85.96 (251) |
| 화 자 2 | 79.45 (232) | 72.60 (212) | 80.82 (236) |
| 화 자 3 | 81.51 (238) | 77.05 (225) | 83.90 (245) |
| 전 체 | 80.94 (709) | 75.00 (657) | 83.56 (732) |

3. 다중관측열을 이용한 시스템에 의한 실험 결과

본 연구에서 제안한 방법에 의한 인식 실험에서는 먼저 다중 관측열에 의한 실험을 행한후 그것을 바탕으로 다중구간 개념을 도입하여 실험하였다.

이 때 state 수는 8로 하였고 codebook 크기는 256에 의해 실험하였다.

1) 다중관측열에 의한 HMM 실험 결과

다중 관측열의 가중치를 식 (1)과 같이 구하였으며 이 때 가중치를 가질 관측 심볼수 S는 2에서 8까지 선택하여 학습시켰고, 인식시에는 두가지 방법으로 실험하였다. 첫번째 방법은 인식시에도 학습시와 마찬가지로 다중 관측 심볼수을 선택하여 실험하였으며 그 결과는 표 2에 나타내었다. 두번째 방법은 인식시에 각 열마다 관측 심볼수를 1개만 선택하는 것으로 실험하였는데 그 결과는 표 3에 나타내었다. 이 때, 일반적인 HMM과 마찬가지로 상태수는 8로 하고 전체 심볼수는 256으로하여 실험하였다.

표 2와 표 3을 비교해 보면, 심볼수를 2, 3, 4, 8로 했을 경우는 인식시 다중 관측열을 선택했을 시가 인식률이 높지만 인식률이 좋은 심볼수가 5, 6, 7인 경우는 심볼수를 1개를 선택하여 인식 실험을 한 경우가 인식률이 더 좋음을 볼수 있다. 또한 심볼수가 다중이건 단일이건 모델 작성시 심볼수를 6를 택하는 경우가 인식률이 가장 좋다.

표 2. 다중 관측열을 이용한 HMM의 인식 결과(인식시 다중 심볼 선택시)
Table 2. Recognition result of HMM using multi-observation sequence.

단위 : % (개)

| 인식률 / 화자 | 관측될 심볼수 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 화자 1 | 89.04 (260) | 90.41 (264) | 90.41 (264) | 91.44 (267) | 91.78 (268) | 91.44 (267) | 91.44 (267) |
| 화자 2 | 85.27 (249) | 86.64 (253) | 86.99 (254) | 86.99 (254) | 87.67 (256) | 87.67 (256) | 86.99 (254) |
| 화자 3 | 87.33 (255) | 88.01 (257) | 88.70 (259) | 89.04 (260) | 89.04 (260) | 88.70 (259) | 88.70 (259) |
| 전체 | 87.21 (764) | 88.36 (774) | 88.70 (777) | 89.16 (781) | 89.50 (784) | 89.27 (782) | 89.04 (780) |

표 3. 다중 관측열을 이용한 HMM의 인식 결과(인식시 1개의 심볼 선택시)
Table 3. Recognition result of HMM using multi-observation sequence.

단위 : % (개)

| 인식률 / 화자 | 관측될 심볼수 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 화자 1 | 87.33 (255) | 90.41 (264) | 91.44 (267) | 91.78 (268) | 92.47 (270) | 91.78 (268) | 91.10 (266) |
| 화자 2 | 82.53 (241) | 84.93 (248) | 85.62 (250) | 86.99 (254) | 88.01 (257) | 88.36 (258) | 87.33 (255) |
| 화자 3 | 85.27 (249) | 87.33 (255) | 88.01 (257) | 89.04 (260) | 89.38 (261) | 89.38 (260) | 88.36 (258) |
| 전체 | 85.05 (745) | 87.55 (767) | 88.36 (774) | 89.27 (782) | 89.95 (788) | 89.73 (786) | 88.93 (779) |

모델링시에 다중 심볼을 선택할 때 너무 많은 심볼을 선택하면 오히려 어떤 심볼과 특징이 비슷하지 않는 심볼에 가중치를 주는 효과가 있어 오히려 인식률을 떨어뜨림을 알 수 있었고, 또한 모델링시에 다중 관측열을 주어 여러 가능한 심볼의 활률값을 주었으므로 인식시에 다중 관측열을 주는 것보다 주지 않는 것이 타당함을 보였다.

2) 제안된 방법에 의한 실험 결과

본 실험에서는 구간이 2인 경우, 3인 경우와 4인 경우를 적용하였는데 구간이 2인 경우는 구간별 심볼수

를 128개로 하였으며 상태수는 8로하여 실험하였고, 구간이 3인 경우는 구간별 심볼수를 100개로 하였으며 상태수는 6개로하여 실험하였고, 구간이 4인 경우는 구간별 심볼수를 64개로 하고 상태수를 4로하여 실험하였다.

인식 실험 결과는 표 4에 나타내었다. 표 2와 표 3 실험 비교에서 다중 심볼을 이용하는 것보다 단일 심볼을 이용해서 인식하는 것이 인식률이 더 좋으므로 본 실험에서는 단일 심볼을 이용하여 인식 실험을 수행했다.

표 4. 구간수의 변화에 의한 인식 결과
Table 4. Recognition result by number of sections.

단위 : % (개)

| 인식률 / 화자 | 구간수 | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 화 자 1 | 92.47 (270) | 93.49 (273) | 92.47 (270) | 91.44 (267) |
| 화 자 2 | 88.01 (257) | 89.04 (260) | 87.33 (255) | 85.27 (249) |
| 화 자 3 | 89.38 (261) | 90.75 (265) | 90.41 (264) | 89.04 (260) |
| 전 체 | 89.95 (788) | 91.10 (798) | 90.07 (789) | 88.58 (776) |

따라서, 구간수가 증가함에 따라 특징 벡터의 수가 줄어듬으로 이 특징 벡터가 그 구간의 모든 특징을 대표할 수 없는 관계로 인식률이 떨어지며, 구간수 2인 다중관측열을 이용한 HMM의 경우가 인식시에 가장 좋은 인식률을 나타낸다.

4. 종합적인 실험 결과

본 연구에서 제안한 다중 관측열과 다중구간 개념에 의한 HMM 모델을 이용한 음성 인식을 일반적인 HMM 및 다른 인식 방법과 비교하였다. 이 때 비교되는 것은 인식 방법에서 가장 좋은 인식률을 나타내는 것으로 기억 용량과 인식시 처리 속도등을 포함하여 표 5에 나타내었다.

여기에서 기억 용량 및 처리 속도는 Rabiner 등[9]에 의한 계산식을 사용하였다. 기억 용량 및 처리 속도 계산시에 평균 프레임수는 40으로 하였다.

종합적인 실험 결과에 의하면 제안된 방법이 기억 용량도 줄일 수 있고 처리 속도도 다른 방법에 비하여 빠르며, 가장 중요한 인식률도 더 좋게 나타남으로 이를 토대로해서 볼 때 본 연구에서 제안한 인식 시스템이 우수한 인식 시스템이라는 것을 알 수 있다.

표 5. 종합적인 실험 결과
Table 5. All-round experimental result.

| 구분 / 종류 | 기억 용량 | 처리 속도 | 인식률 (%) |
|---|---|---|---|
| DP | 64,240 | 1,284,800 multiply | 80.94 |
| MSVQ | 51,392 | 515,088 multiply | 75.00 |
| HMM | 311,168 | 206,080 multiply 93,440 log | 83.56 |
| 다중 관측열의 HMM | 311,168 | 206,080 multiply 93,440 log | 89.95 |
| 제안된 방법 | 311,168 | 106,240 multiply 49,920 log | 91.10 |

## Ⅳ. 결 론

HMM 모델을 이용한 음성 인식을 하는데 있어서 모델의 파라메타를 잘 학습시키는 것은 중요한 일이다. 따라서, 본 연구에서는 다중 관측열을 이용한 다중구간의 HMM 모델을 제안하고, 이 모델을 이용하여 단독어 인식 실험을 수행하였고 일반적으로 많이 사용되어 오고 있는 DP 방법와 MSVQ 방법에 의한 인식 실험 및 일반적인 HMM에 의한 인식 실험을 같은 조건하에서 실행하여 비교하였다.

기존의 대표적인 인식 방법인 DP 패턴 매칭 방법의 경우는 기억 용량이 크며, 계산 시간도 많이 걸리는 단점이 있고, MSVQ의 경우는 기억 용량이 적고 인식 시간이 적게 걸리나, 인식률이 낮은 단점이 있으며, 또한 일반적인 HMM의 경우는 모델을 만드는데 시간이 많이 걸릴 뿐만 아니라 학습하는데 많은 데이타가 필요하고, 화자 독립의 음성 인식에 사용하려면 많은 화자가 있어야 한다.

그런데, 본 연구에서처럼 열에서 하나의 심볼만 선택하여 학습한 후 인식시에 발생할 수 있는 오인식을, MSVQ 코드북과 각 구간별로 다중 관측열을 구하므로 줄여주었으며, 또한 모델을 구간별로 학습시키고 구간마다 인식 후보 단어수를 줄여줌으로서 인식 시간을 대폭 줄이고 인식률을 향상시켰다.

오인식된 단어들을 살펴 보면, 비슷한 단어들간의 애매한 발음(예를 들면, 대천을 대전에 가깝게, 성주를 상

주에 가깝게, 양양을 영양에 가깝게 등)과 잡음이 심한 경우로 D/A에 의해 들어 보아도 구별이 가지 않는 경우가 많았다.

따라서, 화자가 정확하게 발음해야 함은 물론 잡음에 잘 적응하고 보다 더 나은 인식률을 얻을 수 있는 인식 시스템의 개발이 필요하고, 한국어의 음운학적인, 그리고 음향학적인 특성에 맞는 알고리즘의 개발이 필요하다.

## 참 고 문 헌


[1] Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition",IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.

[2] R. M. Gray, " Vector Quantization", IEEE ASSP Magazine, Vol. 1, pp. 4-29, Apr. 1984

[3] F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition", IEEE Trans. on Acoustics, Speech, Signal Processing, Vol. ASSP-33, No. 4, Oct. 1985.

[4] D. K. Burton and J. E. Shore, " Speaker-Dependent Isolated Word Recognition using Speaker-Independent Vector Quantization Codebooks Augumented with Speaker-Specific Data", IEEE Trans. on Acoustics, Speech, and Signal Processing. ASSP-33, No. 2. pp 440-443. Apr. 1985.

[5] D. K. Burton, J. E. Shore and J. T. Buck, " Isolated-Word Speech Recognition using Multisection Vector Qantization Codebooks", IEEE Trans. on Acoustics, Speech, Signal Processing, Vol. ASSP-33, No. 4, Aug. 1985.

[6] L. R. Rabiner and B. H. Juang, " An Intorduction to Hidden Markov Models", IEEE ASSP Magazine, JAN. 1986.

[7] Kai-Fu Lee, "Automatic Speech Recognition", The Development of the SPHINX System", Kluwer Academic Publishers, 1989.

[8] T. O. Ann, Y. G. Byun and S. H. Kim, "Korean Speech Recognition using DHMM", Acoustical Society of Korea, Vol. 10. No. 1, pp 52-61, Feb. 1991.

[9] L. R. Rabiner, S. E. Levinson and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Model to Speaker-independent, Isolated Word Recognition", Bell System Technical Journal, Vol. 62, No. 4, Apr. 1983.

[10] Shikano, K. and Kohda, M., "On the LPC Distance Measures for Vowel Recognition in Continuous Utterance", Institute of Electrical and Communication Engineers of Japen, Trans. on D, J 63-D, May. 1980.

[11] 정의붕, "One Stage MSVQ/DP를 이용한 음성인식에 관한 연구", 한국음향학회, 제13권2호, 1994


## 저 자 소 개


**鄭 義 鵬(正會員)**

1984년 2월 원광대학교 전자공학과 졸업 (공학사). 1986년 8월 광운대학교 대학원 전자계산기공학과 졸업 (공학석사). 1992년 2월 건국대학교 대학원 전자공학과 졸업(공학박사). 1992년 9월-현재 전북산업대학교 정보통신공학과 조교수. 관심분야 : 디지털 신호처리 및 컴퓨터 음성인식, 컴퓨터 비젼