Journal of the Industrial Technology Research Institute.
Korea Maritime University. Vol. 15. 1997.

DHMM(Discrete Hidden Markov Model)의 출력 분포의 유사도(Similarity)를 이용한 음성 인식 시스템

A Study on Digit Recognition System Based on Similarity of Output Distribution from Multiple DHMMs

Gang - Ju You, Ok - Keun Shin***

Abstrat

We choose, in general DHMM(Discrete Hidden Markov Model) speech recognition method, as the recognized word, the index of the word modelwhose output probability is maximum among those of all the words in the vocabulary. In this case, the decisions are made by comparing the similarities of the input feature vector with respect to those of reference features, which might result in the misrecognition when there exists more than two reference features similar to each other.

In this paper, we propose a method of comparing, for a given utterance, the distribution of the output probabilities from all the word models. In fact, we adopt four DHMMs whose feature vectors are LPC, CEP, Weighted CEP and MEL, and linearly combine all the output distributions of these four DHMMs. This method is based on the assumption that every word has a unique output probability distribution for a given DHMM with a specific feature vector, and that the output probability distribution is not the same for DHMMs employing different feature vectors. We need two training data sets : the first for DHMM training, andthe second for the generation of reference distribution pattern. To test the effectiveness of the proposed method, the four DHMMs and a vocabulary consisting of 13 isolated digits are generated, and the reference distribution pattern corresponding to the four feature vectors of each word is estimated. Experimental result shows an improvement of recognition rate by 6.3%.

* 한국해양대학교 대학원
** 한국해양대학교 컴퓨터 공학과 조교수

Ⅰ. 서 론

음성 인식기를 구현하려는 연구는 최근 몇 년 동안 많은 발전을 하여 현재 수십에서 수백단어의 어휘에 대해 신뢰성 있는 인식을 할 수 있는 단계에 있으며, 이미 상용화된 것들도 많이 있다. 이러한 음성 인식기들에 널리 사용되고 있는 알고리즘으로는 DTW(Dynamic Time Warping)[12], 신경망(Neural Network)[13]에 의한 방법, HMM(Hidden Markov Model)[6, 7, 8][14, 15]을 이용한 방법 등을 들 수 있다.

HMM은 화자의 개인차에 따른 음성 패턴의 변동을 통계적으로 처리한 다음, 그 통계량을 확률적인 형태의 모델에 적용하여 음성을 인식하는 방법이다. 이 방법은 확률모델을 사용하기 때문에 개인차나, 조음결합등의 영향으로 나타나는 음성 패턴의 변동을 보다 정확하게 반영할 수 있다. 그러나 모델의 구조를 결정할 때 시행착오나 경험에 의하는 경우가 많고, 학습시에는 다량의 샘플데이터와 계산능력이 필요하다. HMM의 이러한 결점에 대한 대책과 인식률을 향상시키기 위하여 다양한 음성 특징 벡터를 병용하는 방법등이 연구되고 있다.

본 논문에서는 제한된 개수의 숫자음 인식을 위하여 주어진 특정 발화의 특징 벡터를 모든 인식 대상 단어의 모델에 인가하여 얻은 출력값의 분포를 유사도 분포 패턴(Similarity Distribution Pattern)이라 할때, 특정 발화를 모든 인식 대상 단어의 모델에 인가하여 얻은 유사도 분포 패턴과 각 인식 대상 단어에 대한 기준 유사도 분포 패턴의 거리를 이용한 인식기를 구현하였다. 이 인식기는 각 단어의 특징 벡터마다 모든 인식 대상 단어의 모델에 대한 고유의 유사도 분포 패턴을 가지고 있다는 점을 이용하여, 4개의 특징 벡터를 바탕으로 하는 4개의 DHMM을 구현한 다음, 특정 발화의 각 특징 벡터에 대한 모든 인식 대상 단어 모델에서의 유사도 분포 패턴과 각 단어의 특징 벡터에 해당하는 기준 패턴의 거리를 조합하여 최종 결정을 내리는 방법이다. 제안된 방법을 시험하기 위해 13개의 숫자를 단어로 하는 DHMM을 구성하였다. 각 단어의 기준 패턴은 각 단어의 특징 벡터마다 30개의 발화를 모든 인식 대상 단어의 모델에 입력 하여 얻은 출력값의 기하 평균으로부터 추정하였으며, 제안된 방법에 의해서 구현된 인식기가 기존의 DHMM보다 인식률이 나아질 수 있다는 것을 확인하였다.

Ⅱ. 음성의 분석 및 벡터 양자화

2.1 음성의 분석

음성의 분석이란 샘플링된 일련의 음성 신호로부터 각 구간의 스펙트럼 특징을 표현하는 특징 벡터(Feature Vector)를 추출하는 과정으로, 이를 이용하면 음성의 정보를 압축할 수 있으므로 음성의 분석은 음성합성이나 음성인식 분야에 있어서 필수적인 과정이다. 특히 음성인식 분야에 많이 사용되는 특징 벡터로는 선형예측 계수(LPC Coefficient)[9, 10, 11], 켑스트럼 계수(Cepstrum Coefficient)[10, 11],

가중 켑스트럼 계수(Weighted Cepstrum Coefficient)[10, 11], 멜 켑스트럼 계수(Mel Cepstrum Coefficient)[10, 11]등이 있다. 그림 2.1은 이들을 추출하는 과정을 도시한 블록 선도이다.

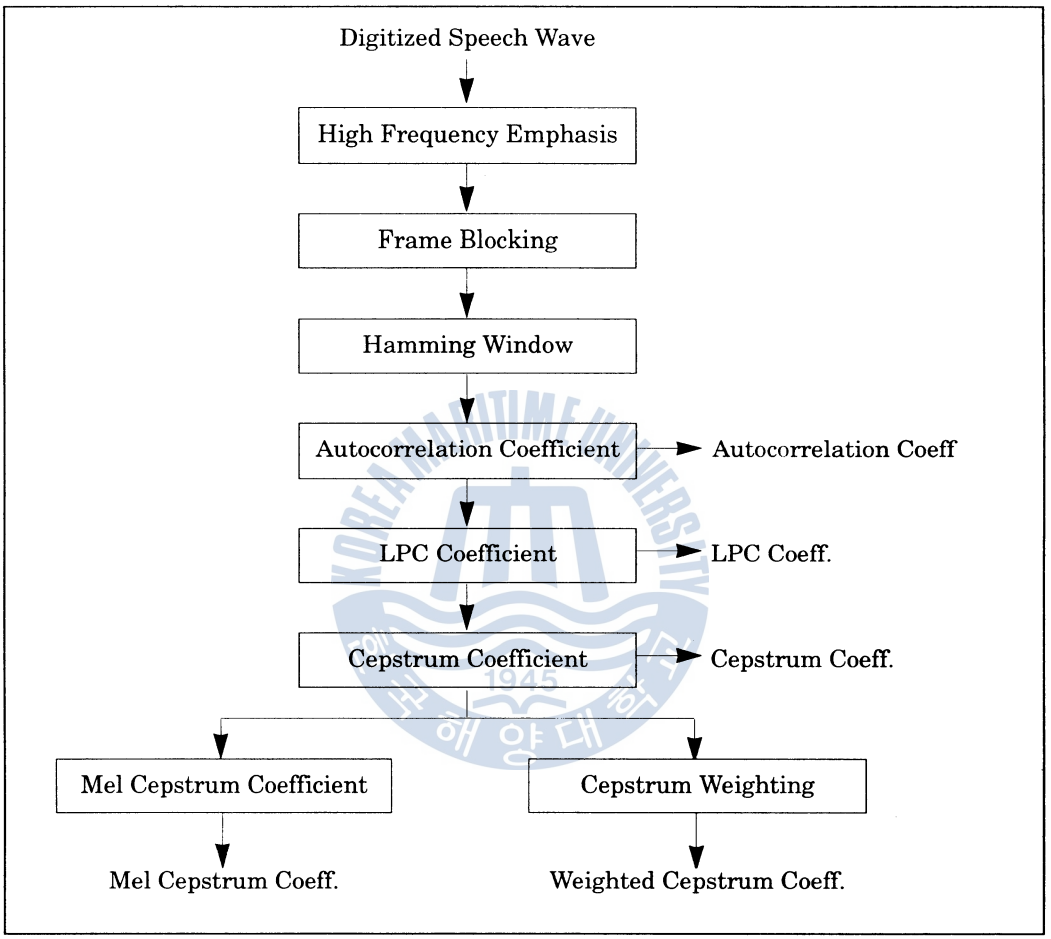


그림 2. 1 특징 백터의 추출

2. 2 벡터 양자화(Vector Quantization)

벡터 양자화[1, 2, 3, 4, 5][16]는 선형예측 계수나 켑스트럼 계수 등으로 만들어진 양자화 테이블(Codebook)과 입력되는 음성신호로 부터 추출한 계수와의 거리가 가장 가까운 양자화 테이블의 심블을 얻는 것을 말한다. 양자화 테이블의 생성에는 K-means 알고리즘이나 이진 트리 알고리즘이 주로 사용되며, 본 논문에서는 양자화 테이블을 생성하기 위해서 이진 트리 알고리즘을 사용하였다. 그림 2.2는 벡터 양자화의 개념도이다.

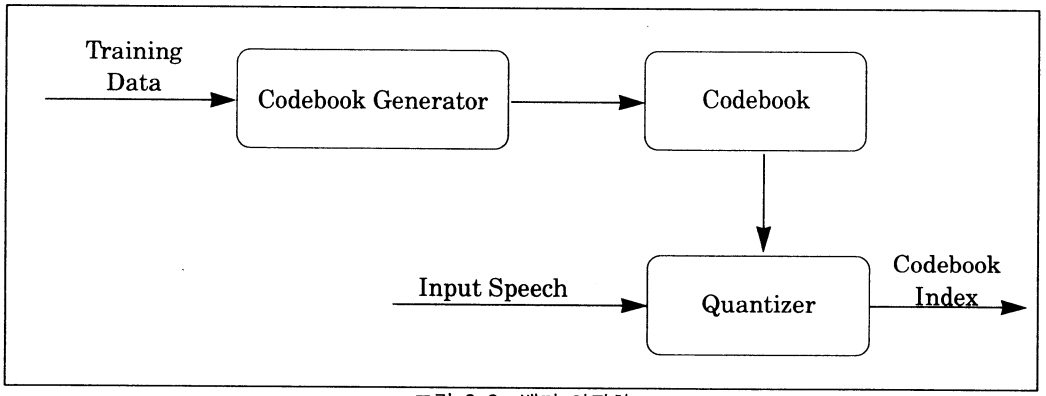


그림 2.2 백터 양자화

Ⅲ. HMM

3.1 HMM

마르코프 모델에서 단지 시간에 관계하는 출력만 알 수 있을 때 이를 HMM이라 하며, 음성인식에서 이것을 이용할 때에는 음성이 발성구조의 시간적 변화에 의해서 발생된 신호이므로 프로세스가 한쪽 방향으로만 천이하도록 제한시킨 Left-to-Right 모델을 주로 사용한다. 이때 마르코프 모델은 초기상태의 확률 벡터 Π와 상태 천이 행렬 $A=\{a_{ij}\}$로 표현된다.

$$\Pi=\{\pi_i\},\ \pi_i=\begin{pmatrix}1 & i=1\\ 0 & otherwise\end{pmatrix},\quad 1\le i\le N \tag{3.1}$$

π_i는 i라는 상태에서의 확률이다. a_{ij}는 현재 상태 i 에서 다음의 상태 j 로 천이할 확률이며, 식(3.2)는 i라는 상태에서 모든 j의 상태로 천이할 확률의 합이 1이라는 것을 나타낸다.

$$\sum_{j=1}^{N} a_{ij}=1,\qquad 1\le i\le N \tag{3.2}$$

여기서, N은 모든 상태의 개수를 나타 낸다. 그리고 시간 t 에서의 음성 신호의 프레임에 대한 특징 벡터를 O_t 라 하면, 음성신호의 전체 프레임에 대한 특징 벡터열은

$$O=\{O_1,\ldots,O_T\} \tag{3.3}$$

로 표현된다. 여기서 T 는 음성신호의 전체 프레임 개수이다. 어떤 상태 i 에서 출력 O_t 가 나올 관측 확률은 $b_i(O_t=k)$로 표현되며, 식 (3.4)의 조건을 만족 한다.

$$\sum_{j=1}^{M} b_i(O_j)=1,\qquad 1\le i\le N \tag{3.4}$$

$$B=\{b_i(O_j)\},\qquad 1\le i\le N,\ 1\le j\le M \tag{3.5}$$

식 (3.5)는 i라는 상태에서 j가 나올 관측확률로 구성된 행렬 B를 나타낸것이다. 여기서 M 은 양자

화 테이블의 크기이다.

HMM은 Π, A, B 로 구성되며, $\lambda(\Pi, A, B)$로 표현된다. 그리고 HMM모델 λ의 A와 B는 학습 데이터로부터 Baum-Welch 알고리즘에 의해서 구해질 수 있다. 상태수를 N으로 하고, 심블의 개수를 T로 했을 때 전향 확률 $\alpha_t(i)$는

$$\begin{aligned} a_t(i) &= P(O_1,\ldots,O_t, q_t=i \mid \lambda) \\ &= \Big[\sum_{j=1}^{N} a_{t-1}(j)\, a_{j,i}\Big] b_i(O_t) \end{aligned} \tag{3.6}$$

이고, 후향 확률 $\beta_t(i)$는

$$\begin{aligned} \beta_t(i) &= P(O_{t+1}, O_{t+2},\ldots, O_T \mid q_t=i, \lambda) \\ &= \sum_{j=1}^{N} a_{i,j}\; b_j(O_{t+1})\, \beta_{t+1}(j) \end{aligned} \tag{3.7}$$

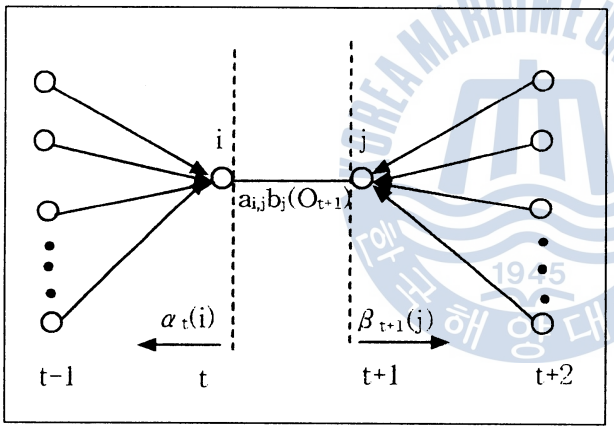


그림 3. 1 전향 확률 및 후향 확률

이다. 여기서 $\alpha_t(i)$는 $O_1, O_2,\ldots, O_t$를 시간에 따라 생성하고 t라는 시간에 i라는 상태에 도달하는 확률이고, $\beta_t(i)$는 $t+1$시간에 j라는 상태에서 시작해서 시간의 흐름에 따른 상태천이에 의해서 $O_{t+1},\ldots, O_T$를 생성하는 확률이며, 이들을 그림 3.1에 표시했다.

모델 λ에서 심볼열 $O=\{O_1,\ldots, O_T\}$가 출력될 확률 $P(O\mid\lambda)$를 이용해서, t시간에는 i라는 상태에 있고 $t+1$시간에는 j 상태에 있을 확률 $\xi_t(i,j)$를

$$\begin{aligned} \xi_t(i,j) &= \frac{P(q_t=i, q_{t+1}=j, O\mid\lambda)}{P(O\mid\lambda)} \\ &= \frac{\alpha_t(i)\, a_{i,j}\, b_j(O_{t+1})\, \beta_{t+1}(j)}{P(O\mid\lambda)\}} \end{aligned} \tag{3.8}$$

로 정의하고, t시간에 i라는 상태에 있을 확률 $\gamma_t(i)$를

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{3.9}$$

이라 하면, 상태 천이확률 $a_{i,j}$ 와 관측확률 $b_j(O_t=k)$의 재추정식은 다음과 같이 된다.

$$\overline{a_{i,j}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \xi_t(i,j)} \tag{3.10}$$

$$b_j(O_t=k)=\frac{\sum_{t=1}^{T}\gamma_t(j)\,\phi(O_t,k)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{3.11}$$

여기서 $\phi(O_t,k)$는 다음과 같이 정의한다.

$$\phi(O_t,k)=\begin{pmatrix}1, & O_t=k\\ 0, & Otherwise\end{pmatrix} \tag{3.12}$$

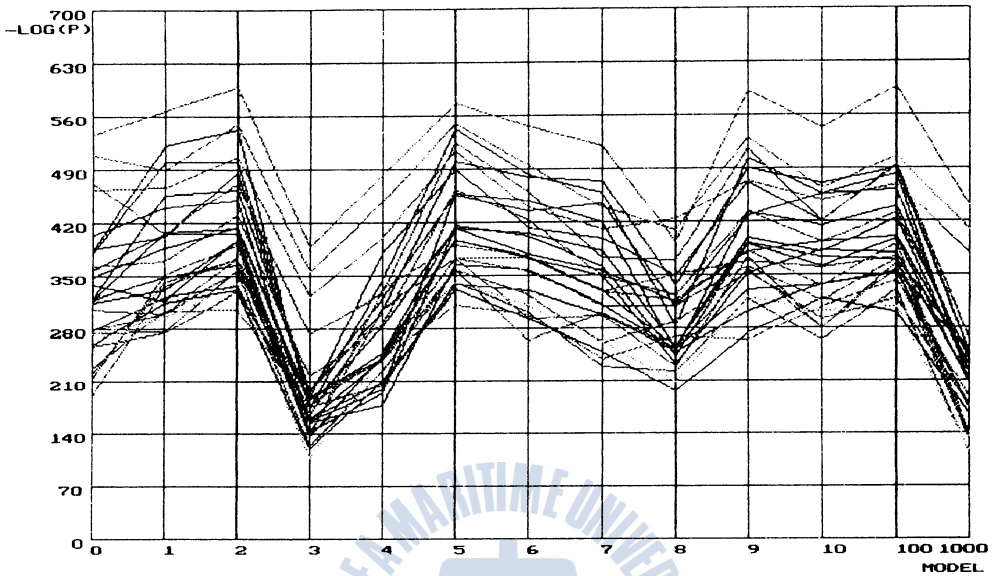
이러한 방법으로 재추정식을 반복해서 수행하면 $P(O|\lambda)$가 최대값을 가지는 a_{ij} 와 $b_j(O_t=k)$에 도달한다.

Ⅳ. DHMM의 출력 패턴과 기준 패턴과의 유사도를 이용한 음성 인식 시스템

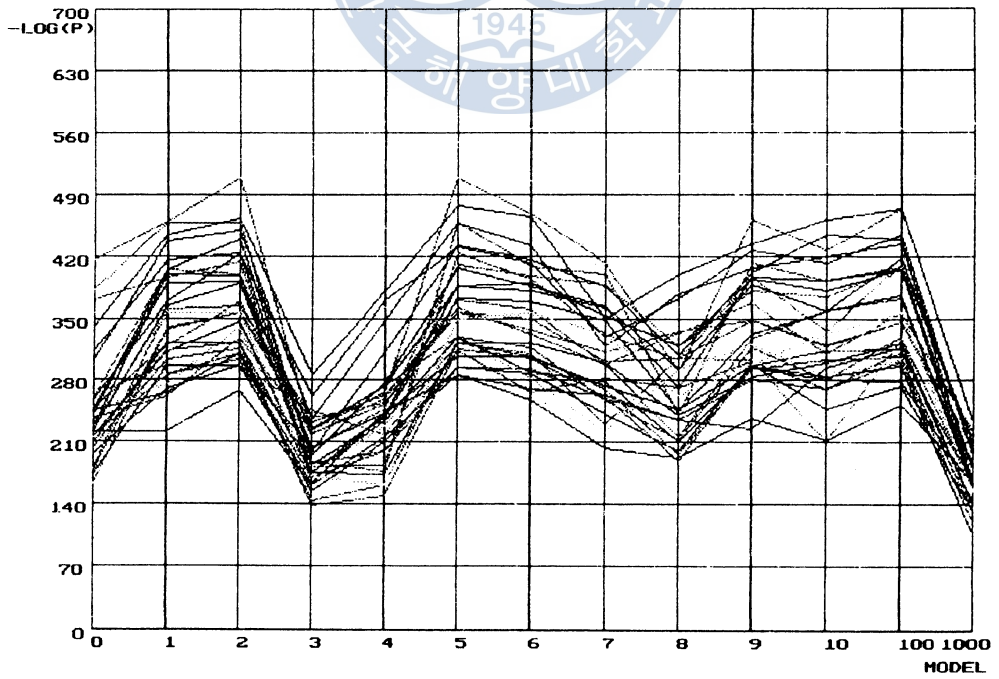
DHMM을 이용한 일반적인 음성인식 방법은 주어진 미지의 발화를 모든 인식 대상 단어들의 모델에 입력하여 얻은 출력값 중에서 최대값을 가지는 단어의 모델을 인식 단어로 한다. 이 경우 유사 단어간의 상호 오인으로 인하여 전체적인 인식률이 저하될 수 있다.

예를 들면, 그림 4.1의 (a)는 30개의 발화 "삼"의 멜 켑스트럼 계수를 모든 인식 대상 단어의 DHMM 모델에 인가하여 얻은 출력값이며, 그림 4.1의 (b)는 30개의 "천"이라는 발화의 멜 켑스트럼 계수를 모든 인식 대상 단어의 모델에 입력하여 얻은 출력값이다. 이 그림들의 횡축은 인식 대상 단어인 13개의 모델("영"~"십", "백", "천")를 나타내며, 종축은 "삼"과 "천"의 모든 인식 대상 단어의 모델에 인가하여 얻은 출력값을 음의 대수 스케일(Negative Log Scale)로 나타낸 것이다. 그림 4.1의 (a)에서는 발화 "삼"에 대한 각 모델에서의 출력값 중에서 "삼"과 "천"의 모델에 해당하는 값이 유사함을 볼 수 있다. 따라서 발화 "삼"은 "천"으로 오인될 가능성이 높다는 것을 알 수 있다. 마찬가지로 그림 4.1의 (b)에서는 발화 "천"에 대한 각 모델에서의 출력값 중에서 "천"과 "삼"의 모델에 해당하는 값이 유사하다는 것과 발화 "천"을 "삼"으로 오인할 가능성이 높다는 것을 알 수 있다.

본고에서는 그림 4.1과 같은 분포, 즉 미지의 발화 X의 특징 벡터를 모든 인식 대상 단어의 모델에 인가하여 얻어지는 유사도의 분포를 유사도 분포 패턴이라 하고, 이를 이용하여 인식률을 개선할 수 있는 방법을 제안 한다. 예를 들면, 그림 4.1 (a)의 "삼"의 유사도 분포 패턴과 그림 4.1 (b)의 "천"의 유사도 분포 패턴이 서로 다르다는 것을 알 수 있다. 따라서 "삼"과 "천"에 대한 각각의 유사도 분포 패턴의 기준 패턴을 만들어서, 이 기준 패턴과 미지의 발화에 대한 유사도 분포 패턴의 거리를 이용하면 "삼"과 "천"의 변별력이 개선 될 수 있을 것이다. 제안한 방법은 각 단어마다 고유의 유사도 분포 패턴을 가지고 있다는 점과 동일한 단어라도 특징 벡터의 종류에 따라 유사도 분포 패턴이 다르다는 점을 이용해서, 각 단어의 각 특징 벡터에 대하여 기준 패턴을 만든 다음, 이 기준 패턴과 미지의 발화 X에 대한 각 특징 벡터의 유사도 분포 패턴 사이의 거리를 조합하여 최종 결정을 내리는 것이다.



(a) 발화 "삼"에 대한 모든 인식 대상 단어의 모델에서의 출력



(b) 발화 "천"에 대한 모든 인식 대상 단어의 모델에서의 출력

그림 4.1 "천"과 "삼"의 발화에 대한 모든 인식 대상 단어의 모델에서의 출력

4.1 기준 패턴의 추정과 인식 알고리즘

본 논문에서 제안하는 음성 인식 시스템의 학습 과정은 두 개의 단계로 이루어진다. 첫 번째 단계는 선형예측 계수, 켑스트럼 계수, 가중 켑스트럼 계수, 그리고 멜 켑스트럼 계수 각각의 특징벡터에 해당하는 각 단어의 DHMM을 학습하는 과정이다. 두 번째 단계는 각 단어의 각 특징 벡터에 해당하는 기준 패턴을 결정하는 과정으로 각 단어의 모든 인식 대상 단어의 모델에 대한 유사도 분포 패턴을 잘 표현하도록 기준 패턴을 구할 필요가 있다.

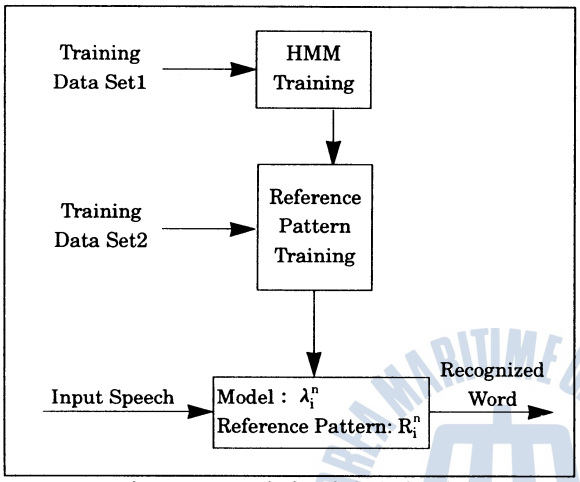


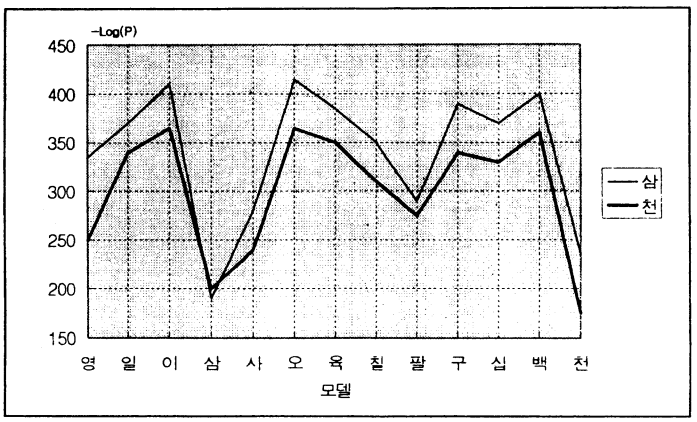
그림 4.2 DHMM의 학습과정 및 인식과정

인식의 과정에서는 인식하고자 하는 미지의 발화의 특징 벡터들을 이미 만들어져 있는 모든 음성 모델에 인가해서 얻은 유사도 분포 패턴과 각 단어의 기준 패턴의 거리를 각 특징 벡터에 대해서 구한 다음, 이들을 조합하여 최종 결정을 내리게 된다.

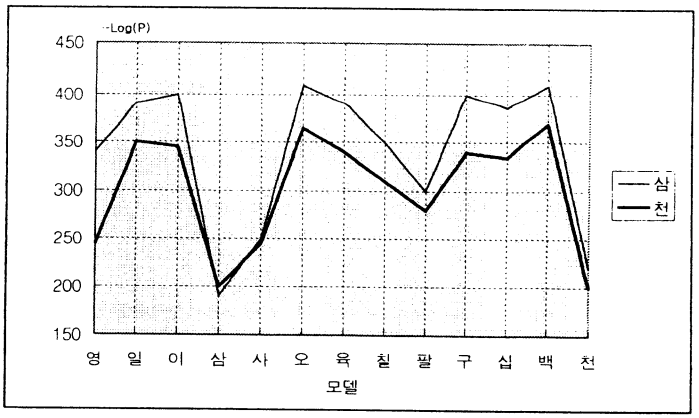
그림 4.2는 위에서 설명한 학습 과정과 인식 과정을 그린 것이다.

4.1.1 기준 패턴의 추정

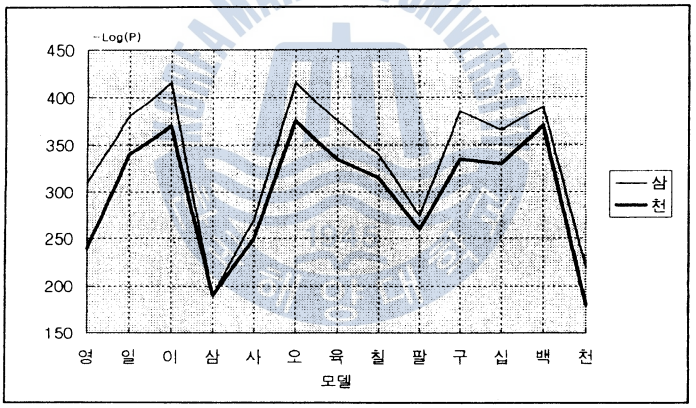
그림 4.3은 단어 "삼"과 "천"의 멜 켑스트럼 계수, 가중 켑스트럼 계수, 켑스트럼 계수 그리고 선형예측 계수에 대한 기준 패턴을 그린 것이다. 이것은 단어 "삼"의 발화 30개와 단어 "천"의 발화 30개를 각 특징 벡터에 해당하는 모든 인식 대상 단어의 모델에 인가하여 얻은 출력값의 기하 평균을 나타낸 것이다. 이 그림에서는 단어 "삼"과 "천"의 기준 패턴이 각 특징 벡터마다 다르다는 것을 알 수 있다.



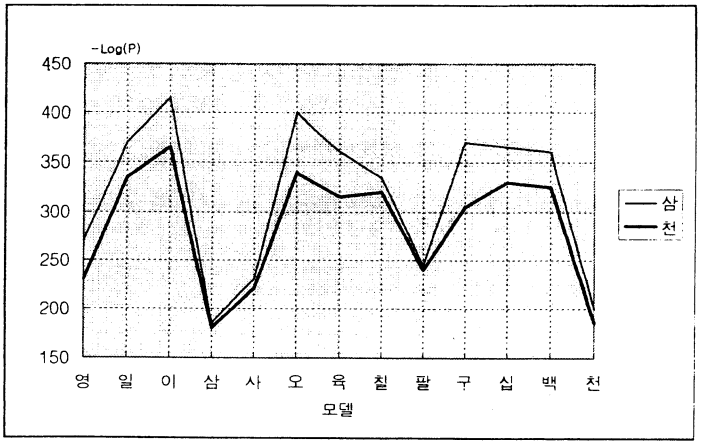
(a) 멜 켑스트럼 계수에 대한 "삼"과 "천"의 기준 패턴



(b) 가중 켑스트럼 계수에 대한 "삼"과 "천"의 기준 패턴



(c) 켑스트럼 계수에 대한 "삼"과 "천"의 기준 패턴



(d) 선형예측 계수에 대한 "삼"과 "천"의 기준 패턴

그림 4.3 "삼"과 "천"에 대한 기준 패턴

아래에 각 단어의 기준 패턴을 구하는 일반적인 방법을 서술한다. 이후의 각 식에서 사용되는 첨자, DHMM모델(λ) 그리고 발화(O)를 다음과 같이 정의한다.

i : 단어의 인덱스 ($i=1,\cdots,Z$)

n : 특징 벡터의 인덱스($lpc, wcep, cep, mel$)

j : 동일단어의 발화에 대한 인덱스($j=1,\cdots,K$)

λ_i^n 을 특징 벡터 n에 대한 단어 i의 DHMM모델이라 정의 할 때 λ^n은 주어진 특징 벡터 n에 대한 모든 단어들의 모델의 집합이다.

$$\lambda^n = \{\lambda_1^n,\ldots,\lambda_i^n,\ldots,\lambda_Z^n\} \tag{4.1}$$

또 모든 특징 벡터들의 집합 λ는 식 (4.2)와 같이 정의한다.

$$\lambda=\{\lambda^{lpc},\lambda^{mel},\lambda^{cep},\lambda^{wcep}\} \tag{4.2}$$

$O_i^{n,j}$ 를 모델 n 에 인가되는 단어 i 의 j 번째 발화라고 하면 O_i^n은 모델 n에 인가되는 단어 i의 모든 발화의 집합이다.

$$O_i^n=\{O_i^{n,1},\ldots,O_i^{n,j},\ldots,O_i^{n,K}\} \tag{4.3}$$

또 단어 i의 모든 발화의 집합 O_i는 식 (4.4)과 같으며,

$$O_i=\{O_i^{lpc},O_i^{mel},O_i^{cep},O_i^{wcep}\} \tag{4.4}$$

모든 발화의 집합 O는 식 (4.5)과 같다.

$$O=\{O_1,\ldots,O_i,\ldots,O_Z\} \tag{4.5}$$

다음은 각 단어의 기준 패턴을 추출하는 과정으로 이미 만들어져 있는 DHMM모델에 기준 패턴을 만들기 위한 학습 데이터를 입력시켜 얻어지는 출력(확률)을 위주로 설명한다. 먼저 O_i^n를 모델 λ_i^n에 인가했을 때 모델의 출력값에 대한 기하 평균 $P^n(O_i^n)$ 은

$$P^n(O_i^n)=P(O_i^n|\lambda_i^n)=\frac{1}{K}\sum_{j=1}^{K}\log(P(O_i^{n,j}|\lambda_i^n)) \tag{4.6}$$

이며, 이는 DHMM 모델이 학습될때의 단어와 동일한 단어가 모델에 인가되었을 때 모델에서 출력되는 확률값의 기하 평균이다. O_i^n를 단어 i가 아닌 다른 단어 h로 학습된 모델 λ_h^n에 인가했을 때 모델의 출력값에 대한 기하 평균 $Q_h^n(O_i^n|\lambda_n^n)$ 은

$$Q_h^n(O_i^n|\lambda_h^n)=\frac{1}{K}\sum_{j=1}^{K}\log(P(O_i^{n,j}|\lambda_i^n)),\ h\neq i \tag{4.7}$$

이다.

식 (4.6)과 (4.7)의 합 H_i^n은

$$H_i^n = P^n(O_i^n) + \sum_{\substack{h=1 \\ h \neq i}}^{Z} Q_h^n(O_i^n \mid \lambda_h^n) \tag{4.8}$$

이 된다. 식 (4.6)과 (4.7)을 식 (4.8)에 의해서 정규화(Normalization)하면

$$Y_i^{n,i} = \frac{P^n(O_i^n)}{H_i^n} \tag{4.9}$$

$$Y_i^{n,i} = \frac{Q_h^n(O_i^n \mid \lambda_h^n)}{H_i^n}, \quad h = 1, \ldots, Z(h \neq i) \tag{4.10}$$

이 되며, 이것은 단어 i의 n번째 특징 벡터에 대한 기준 패턴이 된다.

4.1.2 인식 알고리즘

제안한 방법의 인식 알고리즘은 아래와 같다. 미지의 입력 데이터 X에 대한 모든 인식 대상 단어의 DHMM 모델에서 출력되는 값의 합 RT^n은

$$RT^n = \sum_{h=1}^{Z} \log(P^n(X^n \mid \lambda_h^n)), \quad \text{for each } n = 1, \cdots, M \tag{4.11}$$

이다. 여기서 X^n은 발화 X의 n 번째 특징 벡터이다. 식 (4.11)로 각 모델의 출력치를 정규화 한 값 R_h^n은

$$R_h^n = \frac{P^n(X^n \mid \lambda_h^n)}{RT^n}, \quad h = 1, \ldots, Z \tag{4.12}$$

이다. 모든 기준 패턴과 입력 데이터의 유사도 분포 패턴과의 거리 D_h를

$$D_h = \sum_{n=1}^{M} \sum_{i=1}^{Z} (R_i^n - Y_h^{n,i})^2, \quad h = 1, \ldots, Z \tag{4.13}$$

으로 정의 하면, 미지의 발화 X는 X의 유사도 분포 패턴과 기준 패턴의 거리들 중에서 최소가 되는 단어로 인식한다. 즉 V^*를 인식된 단어의 인덱스라 하면

$$V^* = \underset{1 \leq h \leq Z}{\arg MIN} \; D_h \tag{4.14}$$

이 된다.

Ⅴ. 실험 및 결과

5. 1. 음성 데이터

인식 실험에 사용된 데이터는 "영", "일", …, "십", "백", "천"과 같이 13개의 숫자로 구성되어 있으며, 남성화자 25명이 3번씩 발음한 975개의 음성데이터와 여성화자 25명이 3번씩 발음한 975개의 음성 데이터 중에서 남성화자 10명과 여성화자 10명의 데이터(단어별 60개의 발화)로 크기가 512인 양자화 테이블과 DHMM모델을 훈련시켰다. 그리고 DHMM모델의 훈련에 참여하지 않은 데이터 중에서 남성화자 5명과 여성화자 5명의 데이터(단어별 각 30개의 발화)를 가지고 기준 패턴의 학습에 사용하였으며, 나머지 남성화자 10명과 여성화자 10명의 데이터(단어별 60개의 발화)를 가지고 인식 실험에 이용하였다.

11Khz로 샘플링된 모든 음성신호를 30msec의 길이로 프레임을 분할하고, 프레임과 프레임 사이의 겹침구간을 20msec로 하였다. 그리고 각 프레임들을 1-0.97 Z^{-1}의 디지털필터로 고주파 성분을 강조한 다음, 해밍 윈도우를 적용했다. 또한 각 프레임에 대해서 자기상관 계수와 Levinson - Durbim 알고리즘을 이용하여 12차의 선형예측 계수를 구한 후에 이를 이용하여 켑스트럼 계수를 구하고, 이들로부터 가중 켑스트럼 계수와 멜 켑스트럼 계수를 구하였다. 이들을 이용하여 각 특징 벡터에 해당하는 양자화 테이블을 만들고, 모든 음성 데이터를 양자화 하였다.

5. 2 인식 실험및 결과

본 논문에서 제안한 방법의 인식률 증가 여부를 알아보기 위해 DHMM의 학습이나 기준 패턴의 학습에 참여하지 않은 데이터를 가지고 인식 실험을 하였다.

그림 5.1은 기준 패턴과의 유사도를 고려하지 않은 기존의 DHMM에서의 숫자음 인식 결과를 그

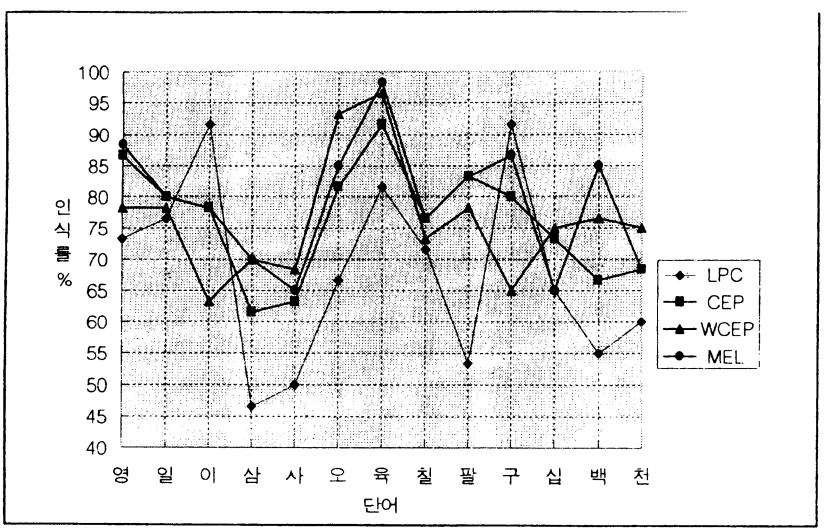


그림 5. 1 HMM의 인식 결과

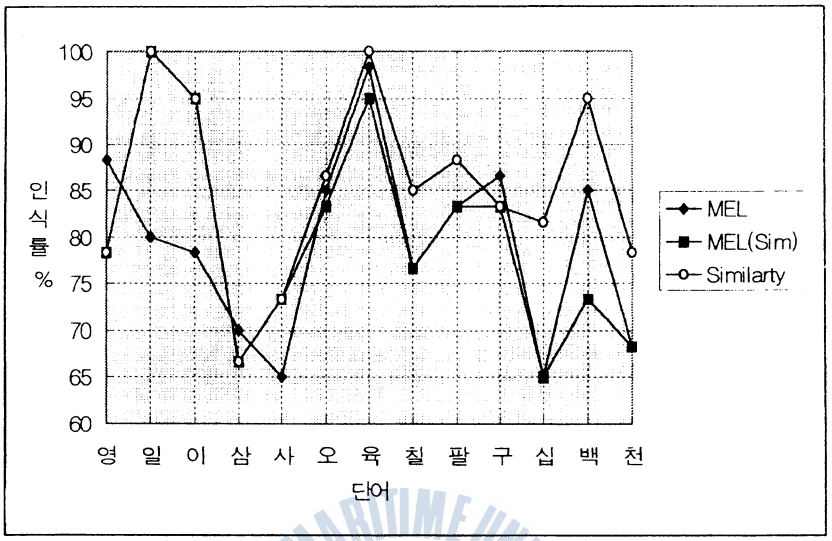


그림 5.2 멜 켑스트럼 계수의 DHMM과 유사도를 이용한 인식시스템의 인식 결과

린것이다.

이 그림에서는 특징 벡터의 종류에 따라 각 단어의 인식률이 다르다는 것과 전체 평균 인식률이 가장 좋은 특징 벡터는 멜 켑스트럼 계수라는 것을 보여준다. 숫자 음성별 오인식 상태를 분석한 결과 "삼"과 "사", "칠"과 "십", "삼"과 "천"을 상호 오인하는 경우가 주류를 이룬다. 이는 스펙트럼 상태가 서로 유사하기 때문인 것으로 생각되며, 여러 논문에서 지적된 바와 같다.[17, 18]

그림 5.2는 DHMM의 유사도 분포 패턴과 기준 패턴의 유사도를 이용한 음성 인식 시스템의 인식 결과와 멜 켑스트럼 계수의 DHMM 모델의 인식 결과를 나타낸 것이다. 이 그림에서 "MEL"은 일반적인 DHMM의 방법을 이용한 인식 시스템을, "MEL(Sim)"은 멜 켑스트럼 계수의 DHMM 모델에서의 유사도 분포 패턴과 기준 패턴의 유사도를 이용한 인식 시스템을, "Similarity"는 멜 켑스트럼 계수, 가중 켑스트럼 계수, 켑스트럼 계수, 그리고 선형예측 계수에 대한 각각의 DHMM 모델에서의 유사도 분포 패턴과 각 특징 벡터에 해당하는 기준 패턴의 유사도를 이용한 인식 시스템을 나타낸다. "MEL(Sim)"의 경우는 기존의 DHMM의 방법보다 인식률이 약간 증가 했지만, "Similarity"의 경우는 기존의 DHMM의 방법보다 인식률이 현저히 개선되었다는 것을 알 수 있다. "Similarity"의 경우 대체적으로 오인식 상태가 개선 되었으나 "일", "삼", 그리고 "구"는 오히려 인식률이 저하 되었다.

그림 5.3은 각 특징 벡터의 DHMM과 유사도를 이용한 인식 시스템의 전체적인 인식률을 나타낸 것이다. 이 그림에서는 인식률이 가장 높은 특징 벡터는 멜 켑스트럼 계수이고, 인식률이 가장 낮은 특징 벡터는 선형예측 계수라는 것을 알 수 있다. 그리고 멜 켑스트럼 계수의 경우에만 기준 패턴과의 유사도를 이용한 시스템은 멜 켑스트럼 계수의 DHMM 모델 보다 0.9%의 인식률 향상을 보이고, 모든 특징 벡터에 대해서 기준 패턴과의 유사도를 이용한 인식 시스템은 멜 켑스트럼 계수의 DHMM 모델 보다 6.3%의 인식률이 향상됨을 보인다.

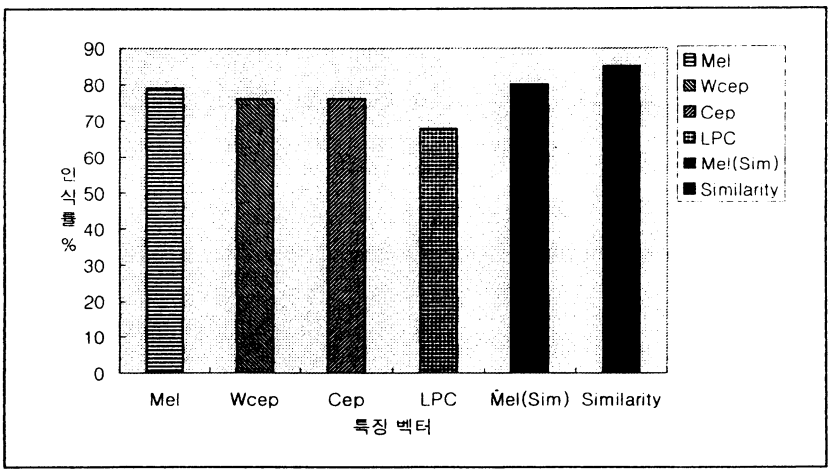


그림 5.3 각 특징 벡터에 대한 DHMM 모델의 인식률과 유사도를 이용한 음성 인식기의 인식률

Ⅵ. 결 론

본 논문에서는 DHMM의 유사도 분포 패턴과 각 단어의 기준 패턴의 유사도를 이용하는 음성 인식 시스템을 구현 하였다. 이 방법은 각 단어의 특징 벡터 마다 모든 인식 대상 단어의 모델에 대한 유사도 분포 패턴이 다르다는 점을 기초로 하여, 각 인식 대상 단어의 각 특징 벡터에 대해서 기준 패턴을 만든 다음, 이 기준 패턴과 모든 인식 대상 단어의 모델에 대한 특정 발화의 유사도 분포 패턴의 거리를 이용해서 인식률을 개선하는 것이다. 제안한 방법의 타당성을 검증하기 위하여 13개의 숫자 음성을 인식대상 단어로 하는 DHMM을 선형예측 계수, 켑스트럼 계수, 가중 켑스트럼 계수, 멜 켑스트럼 계수 각각의 특징 벡터에 대하여 구성하였다. 각 단어의 각 특징 벡터에 해당하는 기준 패턴은 각 단어의 특징 벡터마다 30개의 발화를 모든 인식 대상 단어의 모델에 인가하여 얻은 출력값의 기하 평균으로부터 추정하였다. 그리고 남성화자 10명과 여성화자 10명이 각각 세 번씩 이야기한 13개의 숫자 음성으로 인식 실험을 하였다. 그결과 제안한 방법은 가장 인식률이 높은 멜 켑스트럼 계수의 DHMM 보다 약 6.3%정도의 인식률이 향상 되었으며, 인식률 향상에 효과가 있음을 알 수 있었다. 그러나 특징 벡터의 종류가 많아지면 모델의 학습 시간이나 인식 시간이 길어지고 계산량이 증가되는 단점이 있다. 제안한 방법의 인식률을 더 향상 시키위해서는 음성의 시작점과 끝점을 좀 더 효과적으로 찾는 알고리즘과 벡터 양자화에서 발생하는 오차와 정보의 손실을 줄일 수 있는 양자화 테이블생성 알고리즘이 필요하리라 생각된다.

Reference

[1] Lawrence Rabiner and Biing Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, pp 122 - 133, 1993.

[2] John Makhoul, Salim Roucos and Herrert Gish, "Vector Quantization in Speech Coding", Proc. IEEE, VOL. 73, NO. 11, pp.1551 - 1588, NOVEMBER, 1985.
[3] Alex Waibel and Kai Fu Lee, Reading in Speech Recognition, Morgan Kaufmann, pp.75 - 100, 1990.
[4] Yoseph Linde, Andres Buzo and Robert M. Gray, "An Algorithm for Quantizer Design", IEEE Trans. Communications , VOl. COM - 28, NO. 1, pp.84 - 95, JANUARY , 1980.
[5] Lawrence Rabiner and Biing Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, pp. 242 - 320, 1993.
[6] Lawrence Rabiner, "A Tutorial on Hidden Markov Models and SelectedApplications in Speech Recognition" Proc. IEEE, VOL.77, NO.2, FEBRUARY, 1989.
[7] 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" Bell System Technical Journal, Vol.62, NO.4, APRIL, 1983.
[8] Lawrence Rabiner and Biing Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, pp 321 - 386, 1993.
[9] John Makhol, "Linear Prediction : A Tutorial Review" Proc. IEEE, Vol.63, NO.4, APRIL , 1975.
[10] Joseph W.Picone, "Signal Modeling Techniques in Speech Recognition", Proc. IEEE , Vol.81, NO.9, SEPTEMBER, 1993.
[11] Lawrence Rabiner and Biing Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, pp. 97 - 121, 1993.
[12] Lawrence Rabiner and Biing Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, pp. 200 - 240, 1993.
[13] Ravi P. Ramachandran and Richard J. Mammone, Modern Methods of Speech Recognition, Kluwer Academic, pp 159 - 183, 1995.
[14] S. E. Levinson, L. R. Rabiner and M. M. SONDHI, "An Introduction to the Application of the Theory of Probabilistic Function of a Markov Process to Automatic Speech Recognition", Bell System Technical Journal, Vol. 62, No. 4, APRIL 1983.
[15] S. E. Levinson, L.R.Rabiner and M. M. SONDHI, "An introduction to the Application of the Theory of Probabilistic Function of a Markov Proc. IEEE, Vol. 73, No. 11, NOVEMEBER 1985.
[16] Ravi P. Ramachandran and Richard J. Mammone, Modern Methods of Speech Recognition, Kluwer Academic, pp 23 - 50, 1995.
[17] Alex Waibel and Kai Fu Lee, Reading in Speech Recognition, Morgan Kaufmann, pp.340 - 345, 1990.
[18] 李 基熙, 林 寅七, "제한적 상태지속시간을 갖는 HMM을 이용 한 고립단어 인식", 電子工學會論文誌 第 32 券 B編 第 5 號,1995, 5.