음성인식을 위한 HMM의 파라메터 확장에 관한 연구

°박 창 호* 허 강 인*

* 동아대학교 전자공학과

A Study on the Extension of HMM Parameters for Speech Recognition

°Chang-Ho Park* Kang-In Hur*

* Dept. of Electronic Eng., Dong-A Univ.

요 약

본 논문에서는 연속출력 확률분포 HMM 모델의 단점을 보완하기 위해 (1) 지속시간 확률분포를 갖는 HMM, (2) 동적특징 파라메터를 부여한 HMM, (3) 혼합연속출력 확률분포 HMM을 구성하여 한국어 단음절(100개)에 대한 인식실험을 하였다. 실험 결과 화자 종속에서는 연속출력 확률분포 HMM보다 지속시간 확률분포를 갖는 HMM의 경우 0.70%, 동적특징 파라메터를 부여한 HMM의 경우 1.06%, 혼합연속출력 확률분포 HMM의 경우 1.64%의 인식률이 향상되었다. 화자 독립에서는 연속출력 확률분포 HMM보다 동적특징 파라메터를 부여한 HMM의 경우 1.4%, 혼합연속출력 확률분포HMM의 경우 2.36%, 지속시간 확률분포를 갖는 HMM의 경우 2.78%의 인식률이 향상되었다.

또한 상태수 증가에 따라서도 인식률이 향상되었으나 상태수를 증가할 경우 학습 및 인식 시간이 많이 소요되고 인식률이 오히려 감소될 수도 있으므로 시행착오적인 방법으로 최적인 HMM 모델을 구성해야 한다.

I. 서 론

종래의 음성인식의 주류를 이루었던 DP매칭은 입력된 시계열 패턴을 비선형으로 압축하면서 표준패턴과 조화하는 방법이다. 이 방법은 음성패턴의 시간축상의 변동을 잘 흡수할 수 있지만, 연속 음성인식이나 불특정화자 음성인식에서 인식시간이 길어지고 화자의 개인차에 기인하는 스펙트럼 그 자체 변동에 대해서는 처리하기 어렵다. 이러한 단점을 보안할 수 있는 방법으로서 확률적인 인식 방법으로 알려진 HMM(Hidden Markov Model)이 음성인식의 주류중 하나가 되고 있다.

이러한 HMM의 음성인식에서는 확률모델을 이용하기 때문에 개인차나 조음결합에 의한 음성패턴의 변동을 반영하기 쉽고, 이론적인 전개가 용이하며 언어처리도 같은 확률모델로서 표현할 수 있다는 장점이 있다. 그러나 모델구조 결정에 있어서 시행착오에 대한 의존성이 높고, 학습시 많은 데이타가 필요하며, 음성의 과도적 정보를 경시하는 경향이 있고, 단순 Markov과정으로 가정하기 때문에 패턴의 시간적 상관에 대한 표현력이 부족하다는 단점이 있다.

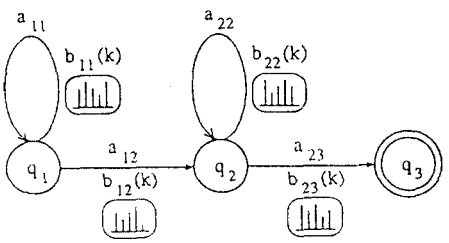
이러한 연속출력 확률분포 HMM 모델의 단점을 보완하기 위하여 구성한 HMM 모델은 다음과 같다.

(1) 지속시간 확률분포를 갖는 HMM
(2) 동적특징 파라메터를 부여한 HMM
(3) 혼합연속출력 확률분포HMM

이들 HMM 모델로서 한국어 단음절(100개)에 대한 인식실험을 하여 파라메타 및 상태수에 따라 비교 실험하였다.

II. 기본 HMM

2.1 이산출력 확률분포 HMM



<그림 1> 이산출력 확률분포 HMM의 예

left-to-right형 HMM은 그림 1과 같은 확률 유한 오토마타로 정의된다. HMM을 이용하여 음성인식을 행할 때에는 미리 필요한 표준패턴 수 만큼의 모델을 준비해 놓고 미지의 입력패턴에 대해서 그 출력확률을 최대로 하는 표준패턴을 인식결과로 하는 것이다. 이산 출력분포 HMM의 경우, 상태 i에서 j로의 천이확률 a_{ij}및 그 천이경로에서 심볼 k의 출력확률 b_{jk}를 학습 데이타에서 구하기 위한 Baum-Welch 알고리즘은 다음과 같다.

상태수를 N, T를 심볼 계열이라 하고 전향확률을 $\alpha(i,t)(i=1,2,\cdots,N : t=1,2,\cdots,T)$. 후향확률을 $\beta(j,t)(j=1,2,\cdots,N : t=T,T-1,\cdots,0)$ 및 모델 M이 심볼 계열 $o = o_1 o_2 \cdots o_T$ 를 출력하는 확률 $P(o \mid M)$이라하고 상태 i에서 상태 j로의 천이가 시각 t에서 발생할 확률을

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

로 정의하면 천이확률 a_{ij}와 출력확률 b_{ijk}의 재추정치는 각각

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

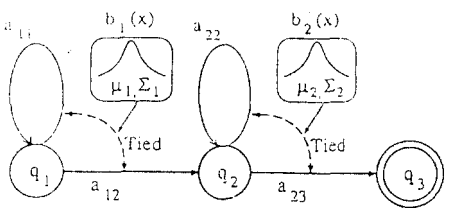
$$b_{ijk} = \sum_{t:o_t=k} \gamma_t(i,j) / \sum_t \gamma_t(i,j) \quad (3)$$

로 된다. 단 $\alpha(i,t)$는 $o_1, o_2, \cdots, o_t$를 생성해서 상태 i에 이르는 확률이고 $\beta(i,t)$는 상태 j에서 시작되는 상태 천이에 의해서 $o_{t+1}, o_{t+2}, \cdots, o_T$가 생성되는 확률이다. 이러한 파라메터 재추정을 반복함으로써 부분적으로 최적치에 수렴한다.

이산 HMM은 입력 파라메터 벡터를 미리 유한개의 심볼로 변환시키지 않고 보통 벡터 양자화에 의해서 실현된다. 그러나 양자화 오차나 벡터 요소간의 상관 정보가 손실되어 인식율이 다소 저하된다.

2.2 연속출력 확률분포 HMM

HMM의 출력확률분포가 이산분포인 경우 음성의 특징 파라메터는 벡터 양자화에 의해 코드북 벡터 번호로 부호화하여 심볼 계열로 된다. 그러나 벡터 양자화에 의한 양자화 오차를 피할 수 없으므로 본 실험에서는 그림 2와 같이 출력확률 분포를 연속분포로 하였다.



<그림 2> 연속출력 확률분포 HMM의 예

출력벡터 o_t가 n차원의 정규분포로 가정하는 경우 출력확률 밀도함수는

$$b_{ij}(o_t, \mu_{ij}, \Sigma_{ij}) = \frac{\exp\{-(o_t-\mu_{ij})^t \Sigma_{ij}^{-1}(o_t-\mu_{ij})/2\}}{(2\pi)^{n/2} |\Sigma_{ij}|^{1/2}} \quad (4)$$

로 주어진다. 여기서, μ_{ij}는 출력 벡터의 평균치, Σ_{ij}는 공분산 행렬, t는 전치, -1은 역행렬을 나타낸다. 이때

$$\gamma_t(i,j) = \frac{\alpha(i,t-1)a_{ij}b_{ij}(o_t, \mu_{ij}, \Sigma_{ij})\beta(j,t)}{P(o \mid M)} \quad (5)$$

로 정의하면, 천이확률의 추정식은 식(2)와 같고, μ_{ij}, Σ_{ij}의 추정식은 다음식으로 주어진다.

$$\mu_{ij} = \sum_t \gamma_t(i,j)o_t / \sum_t \gamma_t(i,j) \quad (6)$$

$$\Sigma_{ij} = \frac{\sum_t \gamma_t(i,j)(o_t-\mu_{ij})(o_t-\mu_{ij})^t}{\sum_t \gamma_t(i,j)} \quad (7)$$

Ⅲ. HMM의 파라메터 확장

3.1 지속시간 확률분포를 갖는 HMM

Viterbi 알고리즘을 그대로 사용하면 종래의 HMM에서는 상태 i에 n시간으로 멈출확률은 $a_{ii}^{n-1} \cdot (1-a_{ii})$로 되고 n의 증가에 따라 지수 함수적으로 감소하며 과도구간과 정상구간의 시간 구조를 충분히 표현할 수가 없다. 실제로 각 상태에서의 지속시간은 음성 세그멘트의 발성시간을 나타내는 중요한 정보이므로 인식시에 고려하는 것이 바람직하다. 지속시간은 감마분포로 근사화하는 것이 좋으며 지속시간 제어는 통계적으로 실시하는 것이 바람직하다. 그래서 a_{ii}로 되는 자기루프는 제거하고 대신에 상태 i로 멈춘 시간장 τ에서 그 지속시간의 확률 $d_i(\tau)$를 구하고 이것을 새로운 파라메터로 해서 부가하는 것으로 한다. 단

$$\sum_\tau d_i(\tau) = 1 \quad (8)$$

로 되는 구속조건을 만족하는 것으로 한다. 이 파라메터를 도입하면 연속출력 확률분포 HMM의 경우 Baum-Welch 알고리즘의 재추정 알고리즘은 다음과 같이 된다.

$$\alpha(i,t) = \sum_j \sum_{\tau \le t} \alpha(j,t-\tau)a_{ij}d_i(\tau) \prod_{n=1}^{\tau} b_{ij}(o_{t+1-n}) \quad (9)$$

$$\beta(i,t) = \sum_j \sum_{\tau \le T-t} a_{ij}d_i(\tau) \prod_{n=1}^{\tau} b_{ij}(o_{t+1-n})\beta(j,t+\tau) \quad (10)$$

여기서

$$\gamma_t(i,j,\tau) = \frac{\alpha(i,t-\tau)a_{ij}d_j(\tau)\prod_{n=1}^{\tau} b_{ij}(o_{t+1-n})\beta(j,t)}{P(o \mid M)} \quad (11)$$

로 되고 천이 확률 a_{ij}와 정규분포의 파라메터 μ_{ij}, Σ_{ij} 추정식은 각각 다음과 같이 된다.

$$a_{ij} = \frac{\sum_t \sum_{\tau \le t} \gamma_t(i,j,\tau)}{\sum_t \sum_j \sum_{\tau \le t} \gamma_t(i,j,\tau)} \quad (12)$$

$$\mu_{ij} = \frac{\sum_t \sum_{\tau \geq t} \gamma_t(i, j, \tau) \sum_{n=1}^{\tau} o_{t+1-n}}{\sum_t \sum_{\tau \geq t} \gamma_t(i, j, \tau)\tau} \tag{13}$$

$$\Sigma_{ij} = \frac{\sum_t \sum_{\tau \geq t} \gamma_t(i, j, \tau) \sum_{n=1}^{\tau} (o_{t+1-n} - \mu_{ij})(o_{t-n} - \mu_{ij})^t}{\sum_t \sum_{\tau \geq t} \gamma_t(i, j, \tau)\tau} \tag{14}$$

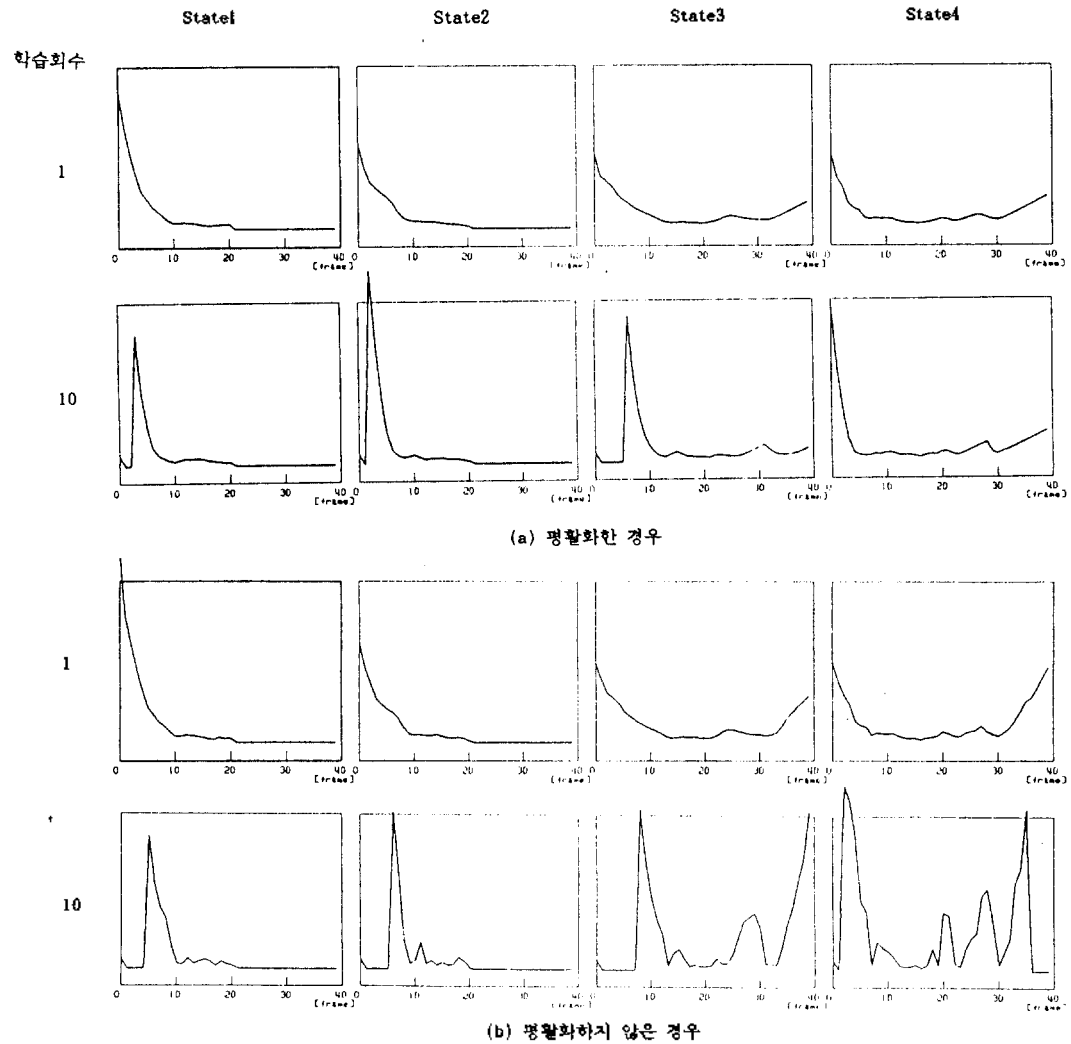
또 지속시간 확률 $d_j(\tau)$의 추정치는

$$d_j(\tau) = \frac{\sum_t \sum_i \gamma_t(i, j, \tau)}{\sum_t \sum_i \sum_{\tau \geq t} \gamma_t(i, j, \tau)} \tag{15}$$

로 된다. 여기서 초기치로써 $d_j(\tau)=0$로 하면 추정치도 $d_j(\tau)=0$로 되므로 각 상태의 지속시간의 허용범위를 유지할 수 있다. 식(15)에 의한 추정만으로 학습회수가 진행됨에 따라서 지속시간 확률분포의 요철이 급격하게 되기 때문에 다음식으로 이용해서 확률분포의 평활화를 실시한다.

$$d_j(\tau) = \begin{cases} \{2d_j(\tau)+d_j(\tau+1)\}/3 & \text{if } \tau=1 \\ \{d_j(\tau-1)+2d_j(\tau)\}/3 & \text{if } \tau=\Delta T \\ \{d_j(\tau-1)+2d_j(\tau)+d_j(\tau+1)\}/4 & \text{else} \end{cases} \tag{16}$$

그림 3은 평활화 한 경우와 평활화 하지 않은 경우의 학습에 따른 지속시간 확률의 변화를 나타낸다.



<그림 3> /갈/의 지속시간 확률의 학습에 따른 변화

3.2 동적특징 파라메터를 부여한 HMM

연속출력 확률분포 HMM에서는 음성의 정상 구간의 HMM을 1개의 상태에 대응시킨다. 음성 패턴의 동적특징을 무시하지 않기 위해서는 어떤 특징 파라메터 변화의 동적모델을 조합에 넣거나 상태수를 증가시켜 자기루프의 천이회수를 감소시켜야 한다.

본 실험에서는 특징 파라메터로 음성의 정적특징량 각각의 차원에 대해서 식(18)로 계산되는 시간축 방향의 선형 회귀계수에 의한 스펙트럼의 동적특징량을 이용하였다. 학습과 인식에 있어서 2개의 특징 파라메터인 정적특징량과 회귀계수에 대해서 식(4)로 각각 출력확률을 계산하고 식(17)로써 최종 출력확률을 구한다. 여기서 정적특징량과 회귀계수의 상관은 0으로 한다.

$$b_{ij}(o_t) = b_{ij}^{cep}(o_t) \cdot b_{ij}^{reg}(o_t) \tag{17}$$

시계열 $x(t)$에 있어서 시간 t를 중심으로한 $2n+1$ 폭으로 한다. 본 실험에서는 $n=5$로 하여 11폭으로 하였다. 선형회귀계수 $\Delta x(t)$는 다음식으로 계산된다.

$$\Delta x(t) = \frac{\sum_{i=-n}^{n} i \cdot x(t+1)}{\sum_{i=-n}^{n} i^2} \tag{18}$$

t: 프레임 t: 회귀계수 폭 i: 회귀계수 차수

3.3 혼합연속출력 확률분포 HMM

연속출력 확률분포 HMM은 음성구간을 단일 정규분포로 근사화해서 파라메터를 추출하였다. 그러나 어떤 상태에 대응하는 음성 구간을 한개의 정규분포로 근사화 할 수 없는 경우가 있기 때문에 복수개의 단일 정규분포를 조합시켜 하나의 복잡한 분포로 근사화 하기 때문에 학습패턴에 가장 근사된 표현을 할 수 있다.

따라서 혼합연속출력 확률분포로 HMM모델을 구성하면 상태 i에서 상태 j로의 천이에서 출력벡터 o_t의 출력확률 $b_{ij}(o_t)$는 다음과 같이 M개의 연속분포의 가중치 합에 의해서 표현된다.

$$b_{ij}(o_t) = \sum_{m=1}^{M} \lambda_{ijm} b_{ijm}(o_t) \tag{19}$$

단

$$\sum_{m=1}^{M} \lambda_{ijm} = 1, \qquad \int b_{ijm}(o)do = 1 \tag{20}$$

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

$$b_{ijm} = b_{ij}(o_t, \mu_{ijm}, \Sigma_{ijm}) \tag{21}$$

로 된다. 천이 확률의 추정식은 식(2)와 같으며

$$\gamma_t(i,j,k) = \frac{\alpha(i,t-1)a_{ij}b_{ijm}(o_t)\beta(j,t)}{P(o \mid M)} \tag{22}$$

로 정의하면 λ_{ijm}, μ_{ijm}, Σ_{ijm}의 추정식은 다음 식과 같다.

$$\lambda_{ijm} = \frac{\sum_t \gamma_t(i,j,m)}{\sum_t \sum_m \gamma_t(i,j,m)} \tag{23}$$

$$\mu_{ijm} = \frac{\sum_t \gamma_t(i,j,m)\, o_t}{\sum_t \gamma_t(i,j,m)} \tag{24}$$

$$\Sigma_{ij} = \frac{\sum_t \gamma_t(i,j,m)(o_t - \mu_{ijm})(o_t - \mu_{ijm})^t}{\sum_t \gamma_t(i,j,m)} \tag{25}$$

Ⅳ. 결과 및 고찰

4.1 실험 데이타 및 분석 방법

음성 데이타는 남성 화자 7명이 5회 발성한 한국어 단음절(100개)을 사용하였다(7명×5회×100개). 음성의 분석 조건은 표 1과 같다.

표1. 음성의 분석조건

샘플링 주파수	10 KHz
창함수(해밍창)	20 ms
프레임 주기	5 ms
분 석	14차 LPC 분석
특징 파라메터	10차 LPC mel cepstrum계수

표2. 한국어 단음절(100개)

가	간	갈	감	거	게	고	구	기	
나	난	날	남	네	노	누	니		
다	단	달	담	데	도	동	두	디	
라	란	람	로	롱	루	리			
마	만	말	메	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	제	조	종	주	지	
차	찬	찰	참	체	초	주	치		
키	칼								
타	탄	탈	탐	토	투	티			
파	판	팔	포	피					
하	한	할	함	후	후	히			

4.2 실험 결과 및 고찰

본 실험에서는 7명이 5회 발성한 음절중 3회분을 화자종속 데이타로 하였고, 학습에 참여하지 않은 2회분을 화자 독립 데이타로 하였으며, 재추정에 의한 최대 학습

회수는 10회로 제한하였다.

표 3은 각 파라메터에 따른 인식 결과를 나타내었다. 화자 종속에서는 연속출력 확률분포 HMM보다 지속시간 확률분포를 갖는 HMM의 경우 0.70%, 동적특징 파라메터를 부여한 HMM의 경우 1.06%, 혼합연속출력 확률분포 HMM의 경우 1.64%의 인식률이 향상되었다.

화자 독립에서는 연속출력 확률분포 HMM보다 동적특징 파라메터를 부여한 HMM의 경우 1.4%, 혼합연속출력 확률분포 HMM의 경우 2.36%, 지속시간 확률분포를 갖는 HMM의 경우 2.78%의 인식률이 향상되었다.

또한 상태수 증가에 따라서도 인식률이 향상되었으며 상태수 4와 5사이에서 가장 큰 폭으로 인식률이 향상되었다. 그러나 연속출력 확률분포 HMM에서 상태수가 7일때 화자독립의 경우 오히려 인식률이 상태수 6보다 감소되었고 상태수를 증가할 경우 학습 및 인식 시간이 많이 소요되었다.

<표3> 각 파라메터에 따른 인식률의 비교 [단위:%]

상태수	CHMM[1]		DD+CHMM[2]		CHMM+RGC[3]		MIX+CHMM[4]	
	종속	독립	종속	독립	종속	독립	종속	독립
4	96.92	82.85	98.23	85.11	98.35	84.72	99.59	87.39
5	98.01	85.59	98.40	88.80	99.23	87.16	99.65	88.12
6	98.62	87.73	99.06	89.35	99.39	88.29	99.86	88.59
7	98.89	87.40	99.55	91.40	99.72	89.01	99.90	88.93
평균	98.11	85.89	98.81	88.66	99.17	87.29	99.75	88.25

1) 연속출력 확률분포 HMM
2) 지속시간 확률분포를 갖는 HMM
3) 동적특징 파라메터를 부여한 HMM
4) 혼합연속출력 확률분포 HMM(MIX=2)

V. 결 론

본 논문에서는 연속출력 확률분포 HMM 모델구조의 단점을 보안하기 위해 (1) 지속시간 확률분포를 갖는 HMM, (2) 동적특징 파라메터를 부여한 HMM, (3) 혼합연속출력 확률분포 HMM을 구성하여 파라메터 및 상태수에 따라 비교 실험하였다.

실험결과 화자 종속에서는 연속출력 확률분포 HMM보다 지속시간 확률분포를 갖는 HMM의 경우 0.70%, 동적특징 파라메터를 부여한 HMM의 경우 1.06%, 혼합연속출력 확률분포 HMM의 경우 1.64%의 인식률이 향상되었다.

화자 독립에서는 연속출력 확률분포 HMM보다 동적특징 파라메터를 부여한 HMM의 경우 1.4%, 혼합연속출력 확률분포 HMM의 경우 2.36%, 지속시간 확률분포를 갖는 HMM의 경우 2.78%의 인식률이 향상되었다.

또한 상태수 증가에 따라서도 인식률이 향상되었으나 연속출력 확률분포 HMM에서 상태수가 7일때 화자독립의 경우 오히려 인식률이 상태수 6보다 감소되었다. 이처럼 상태수를 증가할 경우 학습 및 인식 시간이 많이 소요되고 인식률이 오히려 감소될 수도 있으므로 시행착오적인 방법으로 최적인 HMM모델을 구성해야 한다.

참 고 문 헌

[1] 김종호, "마르코프·체인", 동국대학교 출판부, 1987

[2] 中川聖一, "確率 モデルに よる 音聲認識" 電子情報通信學會, 東京, 1988

[3] 平田好充, 中川聖一, "連續 HMMを 用いた 有聲破裂音識別" 信學會全國大會 講論集A-18 1989.3

[4] 平田好充, 橋本泰秀, 中川聖一, "混合連續 出力 HMMを用いた 有聲破裂音の 識別" 信學會 全國大會講論集 A-18, 1989.9

[5] 中川聖一, 橋本泰秀, "HMM法と ベイズ確率を 用いた連續音聲の セグメンテ-ション," 信學論 J72D(II), 1-10, 1989

[6] 中川聖一, 平田好充, 橋本泰秀, "連續出力 分布型 HMMに よる 日本語 音韻認識" 日本 音聲 學會誌, 1991.

[7] X.D.Hung, "Phoneme Classification Using Semicontinuous Hidden Markov Models" IEEE Trans on Signal Process, VOL.40, 1992

[8] Seiichi Nakagawa, Yoshi Ueda, Takashi Seino, "Speaker-independent, Text_independent Language Identification by HMM" ICSLP-92, pp.196-199, 1992.

[9] 은종관, "연속분포 HMM을 이용한 한국어 연속 음성인식 시스템 개발", 한국음향학회 VOL.13 NO.1 pp.24-31, 1994

[10] 박창호, 김상범, 허강인, "CHMM을 이용한 음소와 단어 단위 음성 인식에 관한 연구", 한국통신학회 부산·경남지부 학술논문 발표회 논문집, 제1집, pp.67-71, 1994.

[11] 이종진, 안점영, 허강인, "Hidden Markov Model을 이용한 한국어 연속음성 인식", 한국통신학회 부산·경남지부 학술논문 발표회 논문집, 제 1 집, pp.72-76, 1994.