비선형 집단화와 완화기법을 이용한 VQ/HMM에 관한 연구

정 희 석°, 강 철 호
광운대학교 전자통신공학과

A Study on VQ/HMM using Nonlinear Clustering and Smoothing Method

Chung, Heui-Suck°, Kang, Chul-Ho
Kwangwoon University
E-Mail : hsjung@explore.kwangwoon.ac.kr

요 약

본 논문에서는 이산적인 HMM(Hidden Markov Model)을 이용한 고립단어 인식 시스템에서 입력특징 벡터의 변별력을 향상시키기 위해 수정된 집단화 알고리듬을 제안하므로써 K-means나 LBG 알고리듬을 이용한 기존의 HMM에 비해 2.16%의 인식율을 향상시켰다. 또한 HMM학습과정에서 불충분한 학습데이타로 인해 발생되는 인식율저하의 문제를 해소하기 위해 개선된 smoothing 기법을 제안하므로써 화자독립 실험에서 3.07%의 인식율을 향상시켰다.

본 논문에서 제안한 두가지 알고리듬을 모두 적용하여 최종적으로 실험한 VQ/HMM에서는 기존의 방식에 비해 화자독립 인식실험 결과 평균 인식율이 4.66% 개선되었다.

Ⅰ. 서 론

음성은 인간의 가장 자연스러운 의사소통 수단이다. 따라서 음성인식은 인간과 기계간의 자연스런 의사소통을 이룰 인터페이스를 제공한다.

현재 연구되어지고 있는 음성인식의 방법으로는 벡터 양자화(Vector Quantization), 시간적인 정합을 이용한 DTW(Dynamic Time Warping) 알고리듬, 확률적인 방법으로 잘 알려진 Hidden Markov Model(HMM), 그리고 신경망에 의한 인식 등이 있다. 그 중에서도 HMM은 음성의 시간적인 변이성을 통계적인 확률모델로 분석하므로써 높은 인식율을 보여 1980년대 이후 활발한 연구가 진행되어져 오고 있으며 최근에는 신경망과 함께 결합하여 다양하게 연구되고 있다.

본 논문에서는 VQ/HMM을 기초한 음성인식시스템에서 확률적으로 개선된 smoothing 기법을 적용하여 학습데이터의 불충분으로 인한 인식율 저하문제를 해결하고, 입력 특징벡터간의 변별력을 향상시켜 인식율을 개선하기 위한 집단화 알고리듬에 대한 연구 및 실험을 하였다.

Ⅱ. VQ/HMM의 기본이론

1. Vector Quantization

벡터 양자화는 무한한 수의 특징 벡터를 유한한 수의 이산 벡터 공간으로 사상시키는 부호화 방법으로서 1980년대 이후 음성인식분야에 적용되기 시작했다. 이는 입력된 음성신호의 많은 정보량을 상대적으로 매우 적은 수의 코드벡터들로 사상시킴으로써 정보량을 줄이고 이에 따른 연산량을 감소시키는 효과를 갖는다[1]. 그러나 이러한 과정에서 정보의 손실이 발생하며 이를 양자화 오차라 한다. 따라서 이를 최소화할 수 있는 최적의 코드벡터를 생성하는 집단화(clustering) 알고리듬이 중요시되고 있다.

자율학습에 의한 집단화 알고리듬으로는 동적 집단화(Dynamic clustering)와 계층적 집단화(Hierarchical clustering)의 두 가지의 기본 형태로 나누어 볼 수 있다. 동적 집단화 방법은 정해진 클러스터의 수에 따라 입력된 모든 특징 벡터들이 안정된 분할을 이룰 때까지 반복적으로 클러스터 멤버와 클러스터 중심값을 갱신하는 것으로써 특히 음성 신호처리 분야에서의 특징 벡터 집단화 알고리즘으로 널리 이용되고 있다[1][2].

본 논문에서는 이산적인 HMM에서 일반적으로 많이 이용되는 동적 집단화 방법으로써 K-means 알고리듬을 이용하여 제안한 알고리듬과의 인식율을 비교하였다.

2. Hidden Markov Model

HMM은 관측할 수 없는 "hidden" process와 음성 신호로부터 이러한 hidden process의 상태로 유도되는 음향학적 벡터를 연결하는 관측 과정(observation process)으로 구성된다. 따라서 HMM에서는 관측할 수 없는 음성의 통계적인 특성을 관측 가능한 벡터열을 통해 추정하므로써 음성의 통계적인 변이성을 잘 반영하고 있다[3][4][5][6].

임의의 음성 특징벡터의 관측열 $O=(o_1 o_2 \ldots o_T)$이 사실임을 가정할 때 주어진 N-states HMM 모델에서의 상태열이 $q=(q_1 q_2 \ldots q_T)$, 라면 결국 관측열의 확률은 다음 식 (1)와 같이 주어진다.

$$P(O \mid q,\lambda) = \sum_{q=q_0} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

$$= \sum_{q=q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \qquad (1)$$

관측벡터에대한 전향변수를 $\alpha_t(i) = p(o_1 o_2 \ldots o_t,\ q_t = i \mid \lambda)$로 정의하면 다음의 과정을 통해 식 (2)와 같이 관측열의 확률을 구할 수 있다.

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1})$$

여기서, $1 \le t \le T-1$, $1 \le j \le N$

$$p(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad (2)$$

또한, 후향변수를 $\beta_t(i) = p(o_{t+1} o_{t+2} \ldots o_T \mid q_t = i, \lambda)$로 정의할 때 다음의 과정을 통해 식 (3)과 같이 관측열의 확률을 얻을 수 있다.

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \qquad (3)$$

여기서, $t = T-1, T-2, \ldots, 1$, $1 \le i \le N$

임의의 관측열의 확률을 최대로 하기위한 HMM 학습 알고리듬으로는 일반적으로 Maximum Likelihood Estimation(MLE)을 이용한다. 여기서는 Baum에의해 확률적으로 증명된 Baum-Welch 재추정 알고리듬을 적용하여 주어진 모델의 확률을 최대로하여 학습한다[3][4].

모델 파라미터의 재추정식은 다음 식 (4)(5)(6)와 같이 표현된다.

$$\overline{\pi_i} = \frac{\alpha_0(i)\beta_0(i)}{\sum_{j=1}^{N} \alpha_T(j)} \qquad (4)$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} b_j(o_t) \beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i) \beta_{t-1}(i)} \qquad (5)$$

$$\overline{b_i(k)} = \frac{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i) \delta(o_t, v_k)}{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i)} \qquad (6)$$

식 (6)에서 표현한 관측심볼의 출력확률에서 delta함수는 다음 식 (7)로 정리된다.

$$\delta(o_t, v_k) = \begin{cases} 1, & o_t = v_k \\ 0, & o_t \ne v_k \end{cases} \qquad (7)$$

결국, HMM모델의 음성인식 과정은 주어진 관측열에 대한 최대 확률분포를 갖는 모델을 결정하는 것으로 이루어진다. 여기서 전향/후향확률에 의한 연산을 이용하여 상태경로를 추적하는 경우 인식율이 다소 우수한 반면 주어진 모든 상태에서의 출력 심볼의 확률을 전부 추정하므로 계산량과 복잡도가 증가한다. 따라서 실시간을 요구하는 인식과정에서는 일반적으로 동적 프로그램 기술로서 잘 알려진 Viterbi decoding 방법을 이용하여 상태경로의 변이와 최적의 모델을 추정함으로써 인식할 수 있다[3][5].

Ⅲ. 제안한 화자독립 VQ/HMM 알고리듬

1. Discriminative Pattern Clustering 알고리듬

K-means 알고리듬은 K개로 정해지는 집단의 수만큼 입력 패턴벡터를 양자화하여 각 패턴벡터와 중심값 벡터와의 유클리드 거리를 최소화하는 중심값 벡터를 추정한다[2][3]. 그러나 이러한 집단화 방법은 일정한 벡터축상에서 정규분포를 갖는 패턴들의 집단화에서는 우수한 성능을 보이지만 정해진 수 K의 값이 충분히 크지 못할 때 상대적으로 변별력이 저하된다.

따라서 본 논문에서는 각 클러스터의 멤버벡터의 수를 제한하여 유클리드 거리를 최소로 하므로써 입력 패턴벡터의 변별력을 향상시켰다. 즉, 특징파라미터의 분포가 벡터평면상에서 특정부분에 밀집될 경우 집중된 영역에서는 많은 수의 클러스터 중심점을 할당하므로써 유사한 음성구간에 대한 변별력을 향상시키는 비선형적 집단화방법을 제안하고자 한다. 본 논문에서는 다음과 같은 단계로 입력 패턴벡터를 집단화한다.

<단계 1> 초기 중심값 : 모든 입력벡터로부터 하나의 중심점 벡터를 연산한다.

$$C_0^{(i)} = \frac{\sum_{n=1}^{N} v_n^{(i)}}{N} \quad , \quad 1 \le i \le p$$

p : 벡터 차수 v_n : 패턴벡터 C_0 : 초기 중심값 벡터

<단계 2> 클러스터 분할 : 가장 많은 멤버함수를 갖는 클러스터를 분할하여 중심값을 이동시킨다.

$$C_l^{+} = C_l(1+\varepsilon)$$
$$C_l^{-} = C_l(1-\varepsilon) \quad , \quad 0.01 \le \varepsilon \le 0.05$$

<단계 3> 중심점 갱신 : 주어진 클러스터 중심값과의 유클리드 거리가 최소가 되도록 입력벡터를 재할당하고 중심값을 갱신한다.

$$l^* = \arg\min_{l} d(v_n^{(i)}, C_l^{(i)})$$

여기서, $d(\cdot)$: 유클리드 거리
l^* : 임의의 입력벡터에대한 클러스터의 인덱스

$$C_l^{(i)} = \frac{\sum_{l=1}^{L} v_l^{(i)}}{L} \quad , \quad L : l \text{ 번째 클러스터 멤버의 수}$$

<단계 4> 집단화된 클러스터의 수가 정해진 수 K보다 작으면 단계 2로 되돌아가 반복하고 K개이면 작업을 종료한다.

2. 제안한 DHMM Smoothing 기법

HMM은 주어진 관측열의 관측시퀀스가 사실이라는 가정하에 최대확률분포를 갖도록 Baum-Welch 재추정

알고리듬을 통해 학습하므로 학습하고자하는 데이터가 불충분할 경우 심각한 오류를 일으킨다[4][6]. 따라서 본 논문에서는 출력심볼의 발생확률을 재추정하는 과정에서 각 심볼별 확률분포를 추가하여 심볼의 출력확률을 smoothing하므로써 불충분한 학습데이타로 인한 인식율의 저하를 방지하고 재추정 파라미터의 신뢰성을 향상시켰다.

다음 식 (8)과 (9)은 본 논문에서 제안하고 있는 출력확률분포의 재추정과정에서 심볼의 발생확률을 고려한 재추정식을 보여준다.

$$\overline{b_i}(k) = \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i) \cdot f_{O_t}(v_k)}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)} \quad (8)$$

$$f_{O_t}(v_k) = \frac{\text{관측열에서 심볼 } v_k\text{가 발생할 기대치}}{\text{학습 데이타 열에서의 심볼의 총 수}} \quad (9)$$

식 (8)에서 표현한 학습 심볼의 발생확률은 HMM 입력단의 벡터양자화 과정에서 얻을 수 있으며 이를 통해 화자의 특이성에 따른 인식율 저하의 문제를 극복하고 안정된 화자독립 시스템을 구현할 수 있다.

다음 그림 1은 본 논문에서 이용하고 있는 VQ/HMM 시스템의 전체 구성도이다.

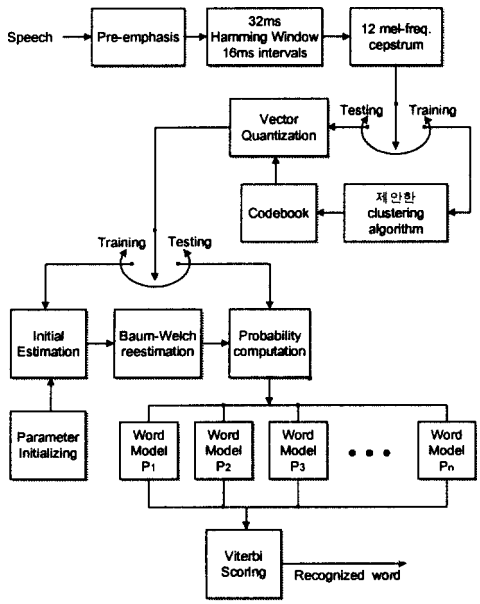

그림 1. 전체 시스템 구성도

Fig. 1. Block-diagram of total system

Ⅳ. 모의 실험 및 결과

본 논문에서 사용한 음성데이타는 아래 표 1과 같이 10개의 윈도우 명령어로 구성하였으며 남성화자 30명으로부터 각 3회씩 발성한 900개의 고립단어를 학습데이타로 사용하였고 남성화자 22명으로부터 각 4회씩 발성하여 얻은 880개의 고립단어를 이용하여 테스트하였다. 각 음성은 mono 8[KHz]로 표본화되었으며 각 샘플당 16[bit]의 resolution으로 디지털 변환되었다.

표 1. 모의 실험에서 사용된 고립단어

Table 1. Isolated words used in simulation

구분	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9
단어	시작	위로	아래로	오른쪽	왼쪽	실행	메뉴	탭	나가기	닫기

본 논문에서는 이용한 음성특징벡터로는 음성의 스펙트럼에 기초한 선형예측계수를 liftering이나 weighting 과정을 통해 쉽게 완화시켜 스펙트럼의 변이성을 제거하여 음성의 정적특성을 강조하므로써 높은 인식율을 보여주는 12차 Mel-frequency cepstrum 계수(MFCC)를 이용하였다.

1. 기존의 VQ/HMM에 대한 인식실험 및 결과

기존의 VQ/HMM에 의한 모의실험에서는 K-means 집단화 알고리듬을 이용해 각 음성특징벡터에 대해 128개의 코드북벡터를 생성하고 이산적인 HMM(DHMM)의 관측벡터로 인가하였고 학습시 재추정 임계값을 0.0005로 선정하여 각 단어별 HMM 모델을 생성하였다.

기존의 VQ/HMM을 이용한 본 실험에서의 화자독립시 평균 인식율이 90.23%의 결과를 보였다.

표 2는 VQ/HMM 학습과정에서 각 단어별 모델을 생성하는데 소요되는 반복횟수를 보여준다. 이는 재추정 알고리듬의 임계값에 의해 좌우되는데 본 논문에서 설정하고 있는 임계값에 대해 기존의 VQ/HMM은 평균 417.2회의 반복횟수가 소요되었다.

2. 제안한 VQ/HMM에 대한 인식실험 및 결과

음성 특징벡터의 변별력을 향상시키고자 본 논문에서 제안한 비선형 집단화 알고리듬을 적용한 실험에서는 화자독립시 평균인식율이 92.39%를 보여 기존의 VQ/HMM에 비해 2.16% 인식율이 향상되었다.

또한, 불충분한 학습 데이터로부터 한정된 수의 벡터 시퀀스를 학습할 경우 발생하는 인식율 저하를 극복하기 위해 본 논문에서 제안한 확률적인 smoothing 기법을 적용한 화자독립에서의 인식실험결과 화자독립시 평균인식율이 93.30%를 보여줌으로써 기존의 VQ/HMM에 비해 3.07% 개선되었다.

최종적으로 제안한 두가지의 개선된 알고리듬을 적용한 경우 화자독립시 기존의 VQ/HMM에 비해 4.66% 향상된 94.89%의 높은 인식율을 보였다.

또한, 표 2와 표 3에서는 기존의 VQ/HMM과 제안한 두가지의 알고리듬을 적용한 VQ/HMM에서의 학습시 각 단어별 모델 생성을 위한 재추정 반복횟수를 나타내었다. 제안한 알고리듬의 경우 인식율의 향상뿐만아니라 기존의 VQ/HMM에 비해 동일한 임계값에서 모델 파라미터가 갱신되는 동적 범위를 줄여줌으로써 오히려 평균 반복횟수도 217회 감소한 200.2회의 평균 반복횟수를 보여주고 있다.

그림 2에서는 기존의 VQ/HMM을 이용한 화자독립 인식실험에서의 평균인식율과 본 논문에서 제안한 두가지 알고리듬에 대한 각각의 평균 인식율 및 두가지 알

고리듬을 모두 적용한 평균 인식율을 비교하였고 그림 3에서는 기존의 VQ/HMM과 제안한 각 알고리듬을 적용한 VQ/HMM에서의 단어별 인식율을 비교하였다.

특히 그림 3에서는 제안한 비선형 집단화 알고리듬이 발성구간에 따라 유사한 단어의 변별력을 향상하는데 우수한 성능을 가진다는 것을 잘 보여준다.

그림 4에서는 기존의 VQ/HMM과 제안한 알고리듬을 적용하여 확률적으로 smoothing된 VQ/HMM에서의 학습시 단어별 반복횟수를 비교하여 나타내었다.

표 2. 기존의 VQ/HMM 학습시 재추정 반복횟수
Table 2. The number of iterations for learning of original VQ/HMM

단어 모델	시작 _0	위로 _1	아래로 _2	오른쪽 _3	왼쪽 _4	실행 _5	메뉴 _6	탭 _7	나가기 8	닫기 _9	평균
반복 횟수	357	740	843	314	509	464	332	283	197	133	417.2

표 3. 제안한 알고리듬에 의한 학습시 재추정 반복횟수
Table 3. The number of iterations for learning of VQ/HMM with proposed algorithm

단어 모델	시작 _0	위로 _1	아래로 _2	오른쪽 _3	왼쪽 _4	실행 _5	메뉴 _6	탭 _7	나가기 _8	닫기 _9	평균
반복 횟수	88	228	451	307	225	57	47	191	151	257	200.2

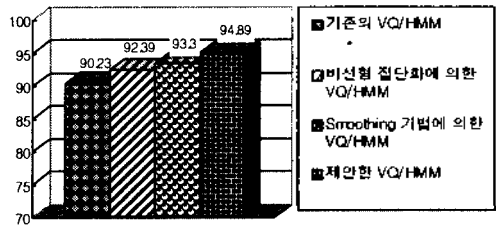

그림 2. 전체 인식율 비교
Fig. 2. Comparison of total recognition rate

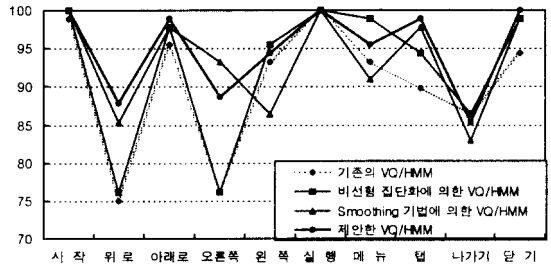

그림 3. 화자독립 실험에서의 단어별 인식율 비교
Fig. 3. Comparison of recognition rates on speaker-independent experimentation

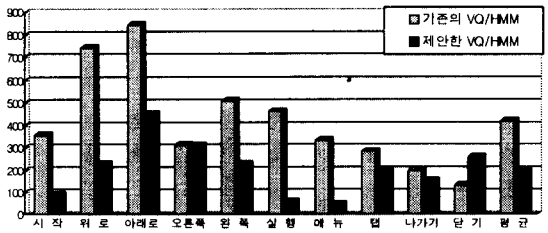

그림 4. 학습시 반복횟수 비교
Fig. 4. Comparison of numbers of iterations for learning

V. 결 론

본 논문의 연구에서는 고립단어 인식을 위해 기존의 VQ/HMM에서의 집단화 알고리듬을 수정하여 유사한 음성구간의 변별력을 향상시켜주기 위한 비선형적 집단화 알고리듬을 제안하였으며, HMM 학습과정에서 불충분한 학습데이터로 인한 인식율 저하를 막기 위한 확률적인 smoothing 기법을 제안하였다. 그 결과 화자독립시 4.66%의 평균인식율의 향상을 보였다.

이는 HMM 학습과정에서 불충분한 학습데이터로 인한 인식율저하문제를 극복하고 있음을 잘 보여주고 있다. 특히 단어별 인식율에서 보여준 것과 같이 비교적 상시간 지속되는 모음구간의 유사성이나 동일모음의 발성음에서 초성, 종성자음의 변별력 문제로 인해 많은 오류를 발생시키는 "왼쪽", "메뉴", "닫기"의 인식율을 기존의 방식에 비해 각각 2.27%, 5.68%, 4.54% 향상시켰고 단어구간이 짧아 특징벡터의 변별력이 저하된 "탭"의 인식율에서도 4.55%의 향상을 보였다. 이는 실험과정에서 "왼쪽"의 발성음이 주로 "오른쪽"으로 오인식되고, "메뉴"의 발성음이 "위로"로 오인식되었던 점, 그리고 "닫기"라는 발성음이 "나가기"로 오인식되었던 점을 감안할 때 음성특징 파라미터의 추출과정에서 유사한 단어구간에서의 변별력을 상대적으로 향상시킨 결과를 잘 말해준다.

HMM 모델의 학습과정에서는 상태의 수, 상태별 출력심볼의 수, 재추정 알고리듬에서의 임계값 등에 의해 학습 반복횟수와 학습시간을 결정하게 되는데 동일한 조건하에서 제안한 알고리듬의 경우 HMM 학습과정에서 소요되는 재추정 알고리듬의 평균 반복횟수를 반감시켜 학습시 더욱 빠른 속도로 수렴하는 결과를 가져왔다. 이는 Baum-Welch 재추정 과정에서 갱신되는 모델의 확률에 대한 동적범위가 제안한 알고리듬의 경우 더욱 작아지므로 기존의 방식에 비해 오히려 빠른속도로 수렴됨을 나타낸 결과이다.

참 고 문 헌

[1] Allen Gersho, Robert M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[2] John R. Deller, Joho G. Proakis, Jphn H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.

[3] Valtcho Valtchev, *Discriminative Methods in HMM-based Speech Recognition*, University of Cambridge, 1995.

[4] L. Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc., 1993.

[5] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[6] X. D. Huang, Y. Ariki, M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh information tech., 1990.