제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

HMM훈련 알고리즘에 따른 음소인식률 비교 연구

구 명완
한국통신 멀티미디어연구소 음성언어연구실

A Comparative Study on the phoneme recognition rate with regard to HMM training algorithms

Myoung-Wan Koo
Spoken Language Research Team, Multimedia Technology Research Lab., Korea Telecom

mwkoo@smm.kotel.co.kr

요 약

본 논문에서는 HMM(Hidden Markov Model) 훈련 방법에 따른 음소인식률의 변화에 대하여 기술한다. 음성모델은 이산 확률 밀도 혹은 연속 확률 밀도를 갖는 HMM을 사용하였으며, 훈련 알고리즘으로서는 forward-backward와 segmental K-means 알고리즘을 사용하였다. 연속 확률 밀도는 N개의 mixture로 구성되어 있는데 1개의 mixture에서 N개의 mixture로 확장할 경우에서는 이진 트리(binary tree) 방식과 one-by-one 방식을 사용하였다. 여러 가지의 조합을 이용하여 음소인식 실험을 수행한 결과 연속 확률 분포를 사용하고 one-by-one 방식을 사용한 forward-backward 알고리즘이 가장 우수한 결과를 나타내었다.

1. 서 론

현재 사용하고 있는 음성인식 알고리즘은 크게 2종류로 나누어질 수 있다. 첫 번째는 DTW(dynamic time warping) 알고리즘이다[1]. 이 알고리즘은 1960년 말 일본과 구 소련 과학자들이 제안하였으며 화자 종속 알고리즘으로 현재까지 실용화 시스템에 사용되고 있다. 특히 구조가 간단하고 ASIC 칩으로 만들기가 쉬워 저 가격의 인식 장치에 많이 이용되고 있다.

두 번째 알고리즘으로 HMM이 있다[2]. 이 알고리즘은 IBM에서 음성인식에 사용하기 시작하였으며 1980년 이후 대부분의 연구소와 학교에서 음성인식 시스템의 기본 구조로 채택되었다. 기본 아이디어는 음성의 특징을 이중 스토캐스틱으로 표현하여 단어의미에 해당되는 확률 값을 EM(Estimation-Maximization) 알고리즘으로 구하는 것이다. HMM 음성 모델 방식은 출력 확률을 모델링 하는 방식에 따라 이산 HMM, 연속 HMM 및 반 연속 HMM 등 세 종류의 모델 방식으로 나누어 진다. 이산 HMM은 계산량이 적어 실시간 시스템의 구현이 쉽다는 장점을 가지고 있으나 음성 모델링 하는 능력이 떨어 진다는 단점이 있다. 연속 HMM은 음성 모델 능력이 가장 우수하나 훈련시키기 위해서는 많은 데이터가 필요하며 계산량이 많아서 실시간 처리가 어렵다는 단점이 있다. 계산량을 줄이기 위하여 covariance matrix를 diagonal matrix로 정의하는 방식을 사용하기도 한다[3]. 마지막으로 반음소 HMM을 사용하는 방식이 있다. 이 방법은 이산 HMM을 사용할 때 사용되는 VQ codeword에 연속 확률 밀도를 갖도록 모델링 하여 주는 방식이다. 음성 모델 능력은 이산 HMM과 연속 HMM의 중간이나 확률값을 구하는 방식이 조금 복잡하다.

본 논문에서는 음성 모델링 하는 방식에 따라 음소 인식률의 변화를 알고자 한다. 2장에서는 HMM 모델링

에 대한 설명을 하고 3 장에서는 음소 인식 시스템에 대하여 기술한다. 그리고 4 장에서 실험 결과에 대하여 분석하고, 마지막으로 5장에서 결론을 맺는다.

2. HMM 모델링

2.1 이산 HMM 모델

이산 HMM 모델은 이산 확률 분포를 사용하며 이를 구현하는 방식으로 VQ(vector quantization) 알고리즘을 사용한다. 4개의 코우드북이 사용되며 파워에 관계되는 코우드 북은 64개의 코우드 워드로 이루어지며 나머지 3개의 코우드 북은 각각 256개의 코우드 워드로 구성된다.

2.2 연속 HMM 모델

연속 확률분포를 사용하는 연속 HMM모델은 출력 확률 밀도로 주로 다음과 같이 Gaussian mixture density를 사용한다

$$bj(x) = \sum_{k=1}^{M} Cjk\ N(x, ujk, Ujk) \qquad (1)$$

식(1)에서 N(x,u,U)는 평균 벡터 u와 covariance 행렬 U를 갖는 D-dimentronal normal density이다. 이때 mixture Cjk는 다음과 같은 공식을 만족해야 한다.

$$\sum_{k=1}^{M} Cjk = 1, \quad 1 \le j \le N \qquad (2)$$

$$Cjk \ge 0, \quad 1 \le j \le N, \quad 1 \le k \le M \qquad (3)$$

여기서 M은 mixture 수이고 N은 state 수이다. 따라서 출력확률 밀도는 다음과 같이 정규화 된다.

$$\int_{\infty}^{\infty} bj(x)dx = 1\ , \quad 1 \le j \le N \qquad (4)$$

가. 훈련 알고리즘

훈련알고리즘은 forward-backward 알고리즘과 segmental K-means 알고리즘을 사용한다. forward-backward 알고리즘은 출력값이 나올수 있는 모든 가능한 경로에 해당되는 모든 likelihood값을 고려하여 최대 likelihood값이 나올 수 있도록 HMM 모델을 추정하는 알고리즘이다. 이 알고리즘은 계산량이 많이 소요되기 때문에 벨 연구소에서는 segmental K-means 알고리즘을 제안하였다[4].

Segmental K-means 알고리즘은 훈련시 비터비 알고리즘을 수행하여 훈련 DB를 출력 확률에 관계되는 상태로 정렬한다. 이때 각 상태에서는 훈련 DB의 해당 프레임이 할당되기 때문에 이 프레임들로부터 각 상태의 출력 확률 밀도를 구할 수 있다. 이 알고리즘은 계산량이 감축되고, 훈련과정에서 인식에 사용되는 알고리즘을 사용한다는 장점이 있으나, 안정적인 출력확률 밀도를 구하기 위해서는 DB양이 많이 필요하다.

나. Mixture 개수 확장 방법

식(1)을 구하기 위해서는 M mixture에 해당되는 평균과 covariance 행렬의 초기값이 필요하다. 이 초기값은 임의로 제공할 수 있으나, 안정적인 확률밀도값을 구하기 위해서 1 mixture부터 시작해서 M mixtrue로 확장해 나가는 방식을 주로 사용한다. 이때 확장해 나가는 방식은 이진트리(binary tree) 방식과 one-by-one 방식으로 나누어 진다. 이진 트리방식은 mixture 개수를 2의 배수로 확장하는 방법이고, one-by-one 방식은 최고 높은 Cjk를 갖는 mixture k를 선택한 후 이 확률분포를 2개로 나누어서 mixture 개수를 증가시킨다. 그림 1에는 이진 트리방식의 mixture 확장방식을 나타내었으며, 그림 2에서는 one-by-one방식의 확장방법을 나타내었다.

3. 음소인식 시스템

음소인식 시스템은 일반적인 단어인식 시스템과 구조적으로 동일하다. 후보단어가 1개의 음소로 구성된 단어라고 정의하면 기존의 단어인식기를 그대로 사용할 수 있다[5].

3.1 특징 추출

전화망을 통해 들어온 음성은 8㎑로 표본화되고, $(1-0.95z^{-1})$의 전달함수를 갖는 필터를 사용하여 pre-emphasis된다. 이 음성은 20msec의 길이의 프레임(frame)단위로 분할된다. 이 프레임은 10msec씩 중첩된다. 자기상관계수(autocorrelation)방법을 사용하여 14차 LPC 분석을 수행하고, 이 LPC 계수를 이용하여 켑스트럴(cepstral)계수를 구한다. 이 계수는 아래 수식의 창(window)$W_c(m)$을 사용하여 weighting 된다.

$$Wc(m) = 1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right), \quad 1 \le m \le Q \qquad (5)$$

구해진 음성 특징은 다음과 같이 4 개의 그룹으로 나누어 지며 이산 HMM에서는 4 개의 코우드 북에 각각 사용되며 연속 HMM에서는 38차 특징 벡타로 사용된다.

1) 12 차 로그 켑스트럼
2) 12 차 델타 로그 켑스트럼
3) 12 차 델타 델타 로그 켑스트럼
4) 로그 파우워, 델타 로그 파우워

3.2 음소 모델

음소모델은 7개의 노드(node)와 12개의 변이(transition)를 갖는 그림 3과 같은 topology로 구성된

다. 이 때 변이들은 3개의 그룹으로 묶어서 같은 그룹에 있는 변이는 동일한 출력 확률을 갖도록 하였다.

4. 음소인식 실험 방법

4.1 음성 DB

사용한 음성 DB는 전화망을 통해 얻은 단어로부터 음소를 자동으로 분할해 주는 자동음소분할기를 사용하여 음소단위로 구축한 것이다. 표 1에는 훈련과 인식 음소 개수를 나타내었다. 사용한 음소는 59개로 구성되었다.

표 1. 음소 DB

	훈련용	인식 실험용
갯 수	701,159	176,100

4.2 음소인식 실험방법

이산 HMM과 연속 HMM의 메모리 량이 동일하도록 연속 HMM의 mixture 개수가 8 이 되도록 하였다. 음소인식 실험은 다음과 같은 방식으로 수행하였다.

방법1: 이산 HMM방식
방법2: 연속 HMM방식, Segmental K-means
방법3: 연속 HMM방식, 이진트리 확장, forward-backward
방법4: 연속 HMM방식, one-by-one 확장, forward-backward

4.3 음소인식 실험결과

음소인식 실험결과는 표 2에 나타나 있다.

표 2 음소인식 실험결과

	방법1	방법2	방법3	방법4
인식률(%)	54.49	53.08	53.43	54.29

인식실험 결과를 보면 이산 HMM방식(방법1)과 one-by-one 확장방식을 이용한 연속 HMM방식(방법4)이 가장 우수한 결과를 얻었다. 이 때 묵음을 제외한 음소 인식률은 표 3에 결과가 나타나 있다.

표 3 음소인식 실험결과(묵음제외)

	방법1	방법2	방법3	방법4
인식률(%)	50.86	51.84	52.16	52.45

표 3을 보면, 이산 HMM방식보다 연속 HMM방식이 우수함을 나타내고, 연속 HMM방식 중에는 forward-backward 알고리즘에 one-by-one mixture 확산 방식을 채택 했을 경우가 인식율이 가장 우수하였다. 실제로 묵음은 단어의 시작과 끝에만 있으므로, 표 2보다는 표 3의 결과가 더욱 의미가 있다.

5. 결론

본 논문에서는 HMM 모델 및 훈련방식에 따른 음소인식률의 변화에 대해서 기술하였다. 이산 HMM과 연속 HMM을 고려하였으며, 훈련방식으로 forward-backward 알고리즘과 segmental K-means 알고리즘을 고려하였다. 또한 M mixture 확률밀도를 구하기 위해 1 mixture로부터 확장하는 방법으로 이진 트리방식과 one-by-one 방식을 비교하였다. 인식성능 실험결과 forward-backward, one-by-one방식을 채택한 연속 HMM 방식이 가장 우수한 결과를 나타내었다.

참고문헌

[1] H. Sakoe, S. Chiba, "Dynamic programming algorithmoptimization for spoken work recognition, " IEEE Trans. ASSP-26, pp.43-49, 19789

[2] K. F. Lee, Automatic speech recognition:the development of the SPHINX. Kluwer Academic Publisher, 1989

[3] C. H. Lee et al., "Acoustic modeling for subword units for speech recognition," in Proc. 1990 ICASSP, pp.721-724, Apr. 1990

[4] C. H. Lee et al., "Acoustic modeling for large vocabulary speech recognition," Computer Speech and Language, No. 4, pp. 127-165, 1990

[5] M. W. Koo et al., "KT-STOCK: A speaker independent, large vocabulary speech recognition system over the telephone," in Proc. 1994 ICSLP, Sep. 1994

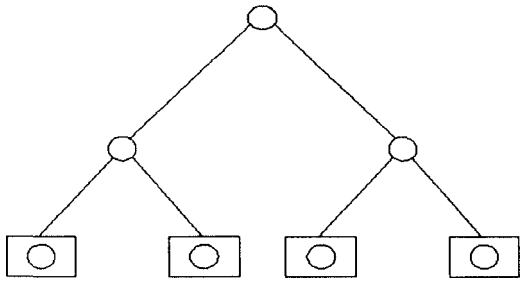

그림 1. Binary-tree 방식

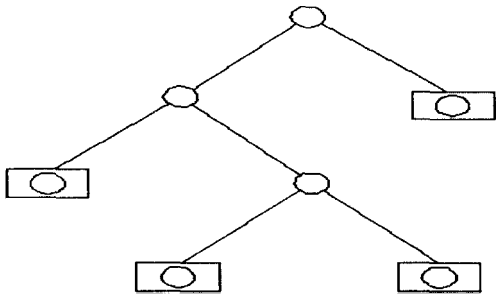

그림 2. One-by-one 방식

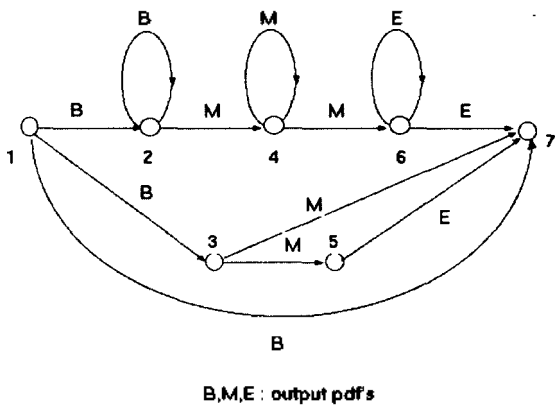

그림 3. 음소 HMM 모델