A Comparison of Discrete and Continuous Hidden Markov Models for Korean Digit Recognition

한국어 숫자음 인식을 위한 이산분포 HMM과 연속분포 HMM의 성능 비교 연구

Hyoung Jin Hong, Seung Ho Choi, Kee Cheol Lee, Ki Chul Kim, and Hwang Soo Lee
홍 형진, 최 승호, 이 기철, 김 기철, 이 황수

Dept. of Information and Communication Eng. KAIST Seoul Campus
한국과학기술원 정보 및 통신공학과

ABSTRACT

본 논문에서는 한국어 숫자음 인식에 대한 이산분포 HMM과 연속분포 HMM의 인식 성능을 비교하였다.

일반적으로 연속분포 HMM은 많은 계산량이 필요하고, 학습시 초기값에 매우 민감하다는 단점이 있지만, 이산분포 HMM의 VQ로 인한 왜곡을 제거함으로써 인식률을 향상시킬 수 있다.

여기서는 성능 비교를 위해서 mel-cepstrum의 분석 차수, 이산분포 HMM의 codebook 크기, 연속분포 HMM의 mixture 갯수등에 따른 인식성능을 비교하였다. 실험 결과 이산분포 HMM에서는 mel-cepstrum 벡터가 14차이고 codebook 크기가 64일 때 가장 좋은 성능을 나타냈으며, 연속분포 HMM에서는 mel-cepstrum 벡터가 16차이고 mixture가 3개일 때 가장 좋은 결과를 얻을 수 있었다. 특히 학습 데이타의 양이 적은 경우에는 연속분포 HMM이 이산분포 HMM보다 더 좋은 인식률을 나타내었다.

Ⅰ. 서론

음성에 관한 자연 과학적 연구에 박차가 가해진 것은 전기통신이 시작되면서 부터이다. 그 이후에 컴퓨터, 통신, 신호처리 기술등의 급속한 발달과 함께 인간과 기계사이의 보다 편리한 인터페이스의 필요성이 증가하면서, 인간에게 더욱 자연스럽고 친숙한 정보 전달 수단인 사람의 말, 즉 음성을 기계가 인식하는 음성인식에 대한 연구가 현저한 발전을 해왔다. 음성인식이란 음성속에 내재되어 있는 언어 정보를 추출하여 그 음성의 정보를 기계가 이해하는 과정을 말하는 것으로, 활용 분야로는 자판없는 타자기에서부터 자동 호텔예약장치, 자동 언어통역장치, 인간의 말을 알아듣고 말을 하는 인공지능을 가진 로보트에 이르기까지 실로 다양하다. 이 음성인식에 대한 연구는 수십년전부터 수행되었으나 현재의 음성인식 기술수준은 인간의 능력에 비하면 아직 보잘것이 없다.

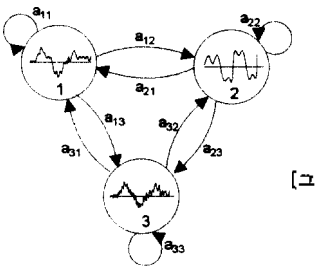
지금까지의 음성인식 방법을 크게 두가지로 나누면, 음성·언어학적인 지식을 이용하는 지식기반 접근방식과 입력된 음성 신호의 스펙트럼을 분석하여 미리 준비된 기준 패턴과 대조시켜 가장 유사한 것을 찾아내는 패턴 matching 방법으로 나눌 수 있다. 이 음성인식 방법중 현재 가장 많이 이용되고 있고 많은 연구가 진행중인 알고리즘은 패턴 matching 방법의 하나인 HMM(Hidden Markov Model)[7,8]과 신경회로망이다. 특히 HMM은 같은 음성의 다양한 파형으로부터 특징을 추출해 내어 그 음성 신호를 통계적으로 모델링는 알고리즘이다. 1970년대말부터 음성인식 시스템에 많이 이용되어온 HMM은 파라미터에 의한 모델링 방식으로서, 음성의 음향학적인 특성뿐만 아니라 시간축의 비선형적 왜곡까지 고려해주는 두 개의 stochastic process로 구성되어 있다. 이 HMM을 이용한 음성인식은 현재까지의 음성인식 방법중 가장 성공적인 것으로 간주되고 있으며, 신경회로망은 인간 두뇌의 생물학적 신경 계통을 모방한 인공 신경망을 이용하여 여러 연결마디들에 음성 특질들을 분산 분포시켜 음성인식을 실현하고자 하는 방식으로 최근 새로이 부각되고 있는 음성인식 방식이다.

본 연구에서는 한국어 숫자음을 인식하는 데 적합한 알고리즘을 찾고자 기존에 발표된 HMM 방식 중에서 이산분포 HMM과 연속분포 HMM 알고리즘을 비교 분석하였다. 본 논문은 2절에서는 HMM에 관해 개략적인 설명을 하고, 3절에서는 실험 내용과 결과를 설명한다. 마지막으로 4절에서는 본 연구의 배경과 결과에 대한 결론을 내린다.

Ⅱ. HMM을 이용한 음성인식[1,7,8]

HMM 방식은 러시아 수학자 Markov에 의해 1900년대 초에 소개된 hidden Markov model에 기초하여 음성이 Markov model로 표현될 수 있다는 가정하에 학습으로 얻은 모델 파라미터를 이용하여 음성을 인식하는 방식이다. 이 방식은 음성 분할(segmentation)과 음성 신호의 labeling을 동시에 통계적으로 해결할 수 있고, 알고리즘의 구조가 간단해서 구현하기 쉬우며, 일반적으로 다른 음성인식 알고리즘에 비해 뛰어난 성능을 나타낸다.

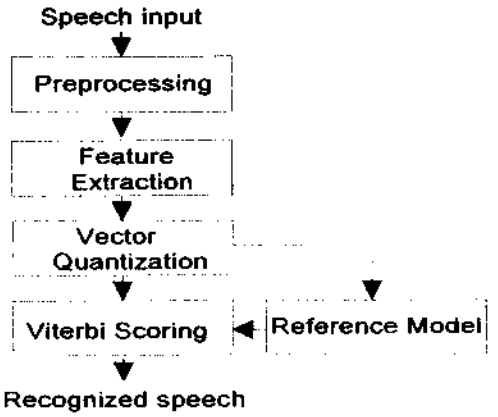
HMM은 두 개의 stochastic process로 구성되어 있으며, 이중 하나는 상태열(state sequence)로 은닉(hidden)되어 있고, 다른 것은 흔히 vector로 이루어진 관측열(observation sequence)로, 이 관측 vector를 모델링하는 방법에 따라 이산분포(discrete), 연속분포(continuous), 준연속분포(semicontinuous) HMM으로 나누고 있다 [1,5,6].



[그림1. HMM의 예]

위와같은 HMM은 Π, A, B와 세가지 파라미터로 정의할 수 있으며, 모델 파라미터 집합은 흔히 $\lambda = (\Pi, A, B)$로 표시한다. 여기서 $\Pi = \{\pi_1, \pi_2, \pi_3, \ldots, \pi_N\}$으로 t=0에서의 초기 상태 확률을 나타내며, $A = \{a_{ij} \mid 1 \le i, j \le N\}$ 로 상태 전이 확률을, $B = \{b_j(O_t) \mid 1 \le j \le N, 1 \le t \le T\}$로 관측확률을 나타낸다.이 HMM을 실제로 응용하여 인식에 이용하려면 다음과 같은 세가지 문제에 접하게 된다. 첫번째는 관측열 $O = O_1, O_2, O_3, \ldots, O_T$와 모델 $\lambda = (\Pi, A, B)$가 주어지면 $P(O|\lambda)$를 계산하는 문제이다. 이는 직접 계산하는 방법이 있으나 그것은 과다한 계산량을 요구하므로 효율적인 계산을 위해 forward-backward 알고리즘이나 Viterbi 알고리즘을 이용할 수 있다. 두번째는 관측열 O가 주어지면 $P(O|\lambda)$를 최대로 하는 파라미터 $\lambda = (\Pi, A, B)$를 추정하는 문제이다. 현재까지 이를 위한 분석적인 해결책이 알려지지 않아 반복적인 알고리즘을 통해 파라미터를 조정하고 있으며, Baum-Welch 알고리즘이나 segmental K-means 알고리즘이 이에 해당한다. 세번째는 주어진 관측열 O에 대해서 최적의 상태열을 구하는 문제로 이는 Viterbi 알고리즘을 많이 이용한다 [4].

이러한 HMM 모델링 방식은 확고한 이론적 기반을 바탕으로 음성인식에 성공적으로 사용되고 있으며, 다음 그림 2는 HMM을 이용한 일반적인 음성인식 시스템을 나타낸다.



[그림2. 음성인식 시스템]

2.1 이산분포 HMM

이산분포 HMM은 관측 vector를 정보 손실이 가장 적은 VQ codebook의 codeword index로 모델링하는 방법이다. 즉 위의 그림 2에서 음성 구간 검출 알고리즘에 의해 검출된 음성 신호는 매 프레임(frame)마다 특징(feature)파라미터를 구하는 과정을 거치며, 여기서 관측 vector 열을 얻는다. 이 관측열은 미리 준비된 codebook을 이용하여 VQ를 수행하는데, 이때 생기는 codeword index 열을 이산분포 HMM에서는 입력된 음성 신호의 관측열로 사용 한다. 따라서 이산분포 HMM의 관측열은 불연속적인 정수로 표시되며, 다음과 같이 표시할 수 있다.

$$b_j(O_t) = b_j(k), \quad v_k = \min_{1 \le v_i \le K} d(O_t, v_i) \quad \cdots\cdots (1)$$

K=codebook 크기

2.2 연속분포 HMM [10,12,13,14]

연속 분포 HMM에서는 관측 vector의 분포가 $B = \{b_j(X) \mid 1 \le j \le N\}$의 형태를 가진다. 이 때 $b_j(X)dx$ 는 관측 vector가 X와 $X+dx$사이에 있을 확률을 나타낸다. $b_j(X)$의 확률 밀도 함수는 Gaussian, Poisson, Gamma, Cauchy 등 여러가지 확률 밀도 함수를 사용할 수 있으며, 이 중 계산의 편이성때문에 Gaussian 밀도 함수가 가장 널리 사용되고 있다. 특히 여러개의 Gaussian 밀도에 가중치를 주고 합친 Gaussian 혼합 밀도 함수가 어떠한 연속 확률 밀도도 근사화할 수 있다는 강점을 지니고 있어서 남·녀의 음성 주파수 차이등의 다양한 변이성을 모델링하기에 적합하다. 이 Gaussian 혼합 밀도 함수는 다음과 같이 표현될 수 있다.

$$b_j(x) = \sum_{k=1}^{M} c_{jk} N(x, \mu_{jk}, U_{jk}) \quad \cdots\cdots (2)$$

여기서 $N(x, \mu, U)$ 평균 vector μ와 covariance 행렬 U를 가지는 정규분포함수이며, mixture 계수 c_{jk}는 다음 공리를 만족해야 한다.

$$\sum_{k=1}^{M} c_{jk} = 1 \qquad 1 \le j \le N \quad \cdots\cdots (3)$$

$$c_{jk} \ge 0, \qquad 1 \le j \le N, \ 1 \le k \le M$$

여기서 M은 mixture의 수이고 N은 state의 수이다. 따라서 분포함수는 다음과 같이 정규화된다.

$$\int_{-\infty}^{\infty} b_j(x)dx = 1, \ 1 \le j \le N \quad \cdots\cdots (4)$$

그러므로 식(2)의 분포함수를 통해 임의의 연속적인 분포에 대한 근사화도 가능하게 된다.

2.3 이산분포 HMM과 연속분포 HMM의 장단점

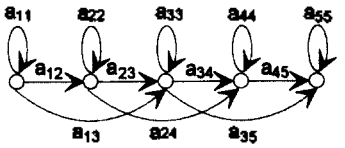
연속분포 HMM은 일반적으로 ML(Maximum Likelihood) 방법에 기초하여 각 파라미터를 계산하며, 관측열을 어떤 process로 가정하느냐에 따라 혼합 Gaussian process나 AR(Autoregressive) process로 나누어 계산할 수도 있다. 연속분포 HMM에 있어서는 학습 데이타만 충분하면, 하나의 mixture보다는 여러개의 mixture를 사용함으로써 더 나은 성능을 보일 수 있는 반면, 계산량이 크게 증가하게 된다. 또한 각 파라미터들의 초기값에 매우 민감하다는 단점이 있는데 이것은 segmental K-means 알고리즘을 이용하여 보완할 수 있다 [11]. 하지만 확률분포 함수를 parameter 방식에 기초하여 계산하는 연속분포 HMM은 이산분포 HMM보다는 필요한 학습 데이타의 수가 작다는 장점이 있다.

이산분포 HMM에서 음성의 관측신호는 VQ를 통해 가장 작게 왜곡되는 codeword로 변환되며, 이 codeword들의 열을 이산분포 HMM에서는 음성신호의 관측열로 간주한다. 따라서 연속분포 HMM보다는 계산량을 크게 줄일 수 있다는 장점이 있지만, VQ로 인해 음성신호에 생기는 왜곡이 인식률을 저하시키는 원인이 될 수 있다 [9,14].

III. 실험 및 결과

본 연구에서는 연속분포 HMM과 이산분포 HMM의 한국어 숫자음에 대한 인식 성능 실험을 위해 특징 vector의 분석차수를 변화시키면서 이산분포 HMM의 codebook 크기에 따른 인식률의 변화와 연속분포 HMM의 mixture 갯수에 따른 인식률의 변화를 분석하였다.

실험을 위한 음성 신호는 10kHz, 12bit로 A/D 변환되었고, 숫자음 /일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 공/을 남자 10명, 여자 10명이 10회씩, 잡음이 최대로 억제된 상태에서 녹음을 하였다. 이중 남자 5명, 여자 5명이 10회씩 발음한 데이타중 5회분을 학습에 이용하였으며, 나머지 학습에 이용되지 않은 데이타는 다중 화자종속 실험과 화자독립 실험을 수행하는 데 이용하였다. 음성의 특징 파라미터로서 mel-cepsturm을 사용하였고, 음성 신호를 모델링하기 위한 HMM 구조는 다음 그림 3과 같다. 또한 각 숫자음들은 단어 단위로 모델링 하였으며, 각 숫자음의 state 수에 차등을 두었다. 특히 연속분포 HMM에서는 full covariance 행렬을 diagonal covariance 행렬로 근사화하였다.



[그림3, HMM 구조]

실험 A. codebook 크기에 따른 이산분포 HMM의 인식률 변화

codebook 크기에 따른 이산분포 HMM의 인식률 변화는 다음 표1과 같으며, 표1에서 보듯이 mel-cepsturm의 차수가 14, codebook 크기가 64일 때 다른 경우보다 인식률이 다소 좋게 나타났다.

<표1. Codebook 크기에 따른 이산분포 HMM의 인식률>

분석 차수 / codebook 크기	10	12	14	16
32	92.8	92.0	93.0	94.8
64	95.6	96.0	96.4	94.8
128	94.6	97.0	96.4	95.6
256	97.8	97.4	97.4	98.2

- 다중화자종속 -

분석 차수 / codebook 크기	10	12	14	16
32	90.3	85.1	91.3	87.7
64	90.8	92.1	94.0	93.6
128	91.5	89.7	90.1	93.5
256	92.5	92.2	87.9	88.8

- 화자독립 -

실험 B. mixture 수에 따른 연속분포 HMM의 인식률 변화

화자독립 실험과 화자종속 실험에 사용한 데이타는 위에서 설명한 것과 같으며,다음 표2에 실험결과를 나타내었다. 표2에서 보듯이 연속 분포 HMM에서는 mixture의 수가 증가한다고 인식률이 증가하는 것이 아니라 어떤 임계치가 있어서 그 이상이 되면 오히려 인식률이 저하된다. 본 실험에서는 mixture 수가 4개 이상일 때 이러한 현상이 나타나고 있다.

<표2. Mixture 수에 따른 연속분포 HMM의 인식률>

분석 차수 / Mixture 수	10	12	14	16
1	83.6	90.6	93.8	93.6
2	88.8	91.8	93.8	94.0
3	91.4	95.4	95.6	94.2
4	87.8	95.6	95.2	93.2
5	85.2	94.2	94.8	91.4

- 다중화자종속 -

분석 차수 / Mixture 수	10	12	14	16
1	85.0	83.2	89.9	92.8
2	89.9	86.2	86.1	92.5
3	91.8	86.4	86.8	92.8
4	81.9	88.4	83.3	90.4
5	80.0	88.2	82.7	88.2

- 화자독립 -

실험 C. 학습 데이타의 크기에 따른 이산분포 HMM과 연속분포 HMM의 인식률 변화

본 실험에서는 학습 데이타의 크기가 변할 때, 이산분포 HMM과 연속분포 HMM의 인식률 변화를 살펴 보았다. 위의 실험 A와 B에서는 하나의 숫자음을 모델링하기 위해 50개의 token을 사용하였는데, 실험 C에서는 하나의 숫자음을 모델하기 위한 학습 데이타의 token 수를 10으로 한 경우에 14차와 16차 특징 vector에 대해 실험하였다. 실험 A와 B의 결과로부터 이산분포 HMM의 VQ codebook 크기를 64로, 연속분포 HMM의 mixture 수를 3개로 제한하여 비교하였다. 그 결과는 다음 표 3과 같으며, 표에서 알 수 있듯이 한 개의 숫자음을 모델하기 위해 50개의 token(약 2000-3000개의 관측 vector)을 이용한 실험 A와 B에서는 이산분포 HMM의 성능이 더 좋았으며, 학습 데이타의 수가 적은 경우에는 연속분포 HMM이 이산분포 HMM보다 더 좋은 인식률을 보이고 있다.

<표3. 학습 데이타의 크기에 따른 인식률>

실험 내용 \ 분석 차수		14	16
50 tokens / 1 model	DHMM	96.4	94.8
	CHMM	95.6	94.2
10 tokens / 1 model	DHMM	90.8	90.2
	CHMM	91.4	91.4

- 다중화자종속 -

실험 내용 \ 분석 차수		14	16
50 tokens / 1 model	DHMM	94.0	93.6
	CHMM	86.8	92.8
10 tokens / 1 model	DHMM	87.5	89.4
	CHMM	90.6	90.8

- 화자독립 -

참고로, 계산량에 있어서는 연속분포 HMM의 covariance 행렬을 근사화 하였으나, 3개 이상의 mixture에서는 이산분포 HMM에 비해 10배 이상의 학습 시간을 요구했으며, 인식시에는 4배이상의 계산시간을 필요로 했다.

Ⅳ. 결론

본 연구에서는 연속분포 HMM과 이산분포 HMM의 알고리즘을 한국어 숫자음의 인식에 사용하여, 한국어 인식에 적합한 알고리즘을 찾고자 하였다. 실험 결과에서 알 수 있듯이 이산분포 HMM에서는 mel-cepsturm 차수를 14로, codebook의 크기를 64로, 연속분포 HMM에서는 mel-cepstrum 차수를 16으로, mixture 수를 3으로 선택하는 것이 적절해 보인다. 일반적으로 연속분포 HMM은 이산분포 HMM과는 달리 VQ로 인한 음성 신호의 왜곡이 없으므로 작은 량의 학습 데이타만으로도 높은 인식 성능을 얻을 수 있으나, 학습 데이타의 양이 충분하면 이산분포 HMM의 인식 성능이 더 좋을 수 있다.

향후 잡음이 섞인 음성및 mel-cepstrum 이외의 다른 특징 파라미터에 대한 이산분포 HMM 및 연속분포 HMM, 준연속분포 HMM의 성능을 비교 분석하려고 한다.

[참 고 문 헌]

[1] X. D. Huang, Y. Ariki, and M. A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, England, 1990.

[2] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping for Connnected Word Recognition," IEEE Trans. on ASSP, Vol. 29, No. 2, pp. 284-297, April 1981.

[3] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," IEEE Trans. on ASSP, Vol. 32, No. 2, pp. 263-271, April 1989.

[4] G. D. Forney, "The Viterbi Algorithm," Proc. of the IEEE, Vol. 61, No. 3, pp. 268-278, March 1973.

[5] K. F. Lee, Automatic Speech Recognition, Kluwer Acadenic Publshers, Norwell, M. A., U.S.A, 1989.

[6] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Inc. Englewood Cliffs, New Jersey 07632

[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," Proc. of the IEEE, Vol. 77, No. 2, February 1989.

[8] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, Vol. 3, No. 1, pp. 4-16, January 1986.

[9] X. D. Huang and M. A. Jack, "Semi-continuous Hidden Markov Models for Speech Signals," Computer Speech and Language, Vol. 3, No. 3, pp. 239-251, July 1989.

[10] B. H. Juang and L. R. Rabiner, "Mixture Autoregressive Hidden Markov Models and for Speech Signals," IEEE Trans. on ASSP, Vol. 33, No. 6, pp. 1404-1413, December 1985.

[11] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Segmental K-Means Training Procedure for Connected Word Recognition," AT & T Technical Journal, Vol. 65, No. 3, pp. 21-31, May 1986.

[12] L. A. Liporace, "Maximum Likelihood Estimation for Multivariate Observation of Markov Sources," IEEE Trans. on Information Theory, Vol. IT-28, No. 5, September 1982.

[13] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of Isolated Digits using Hidden Markov Models with Continuous Mixture Densities," AT & T Technical Journal, Vol. 64, No. 6, July-August 1985.

[14] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M Sondhi, "Some Properties of Continuous Hidden Markov Model Representations," AT & T Technical Journal, Vol. 64, No. 6, July-August 1985.