대분류기법을 이용한 음성인식 시스템의 속도향상

전화성[01], 김길연[2], 윤영선[1], 오영환[1]
[1]한국과학기술원 전산학과, [2]에스엘투 주식회사 음성인식팀,

The Performance Improvement of Speech recognition system using Hierarchical Classification Method

Hwaseong Jeon[01], Kilyeon Kim[2], Young-Sun Yun[1], Yung-Hwan Oh[1]
[1]Spoken Language Lab., Dept. of Computer Science, KAIST
[2]Speech recognition team, SL2 co.,LTD.
hsjeon@slworld.co.kr

요 약

본 논문에서는 HMM 학습모델을 이용하여 1445단어 음성인식기를 구현하고, 대분류기법을 이용하여 그 성능을 향상시키는 방법에 대하여 연구를 수행하였으며, 속도개선에 중점을 두었다. 속도개선을 위해서 HMM모델에 계층적 대분류 기법을 적용시켰다. HMM의 상태수가 많을수록 속도가 저하된다는 점을 고려하여, 적은 상태수의 HMM모델로 후보를 정하고, 가변적으로 해당하는 상태수의 HMM모델로 목적단어를 인식하는 방법을 제안하였다. 후보를 정하는 방법을 후보수와 특징파라미터의 종류와 수를 고려하여 다양하게 설정, 실험하여 가장 이상적인 경우를 찾아내었다.

1. 서론

현재 연구되고 있는 음성 인식의 방법으로는 신경망에 의한 인식(ANN, Artificial Neural Net), 시간적인 정합을 이용한 DTW(Dynamic Time Warping)알고리즘, 확률적인 방법으로 알려진 HMM(Hidden Markov Model)등이 있다. 그 중에서도 HMM을 이용한 방법이 가장 인식율이 높은 것으로 알려져 있다.[3]

기존은 연구들은 인식율에 하나에만 중점을 둔 연구였다. 보통 인식율을 높이려면 처리속도가 길어지는 것이 보통이다. 본 논문은 인식율을 원래의 시스템에 대해 근접하게 유지하면서, 처리속도를 빠르게 하는 연구이다. 이를 위해 HMM모델에 대한 계층적인 대분류기법을 적용하였다.

보통 인식처리시간은 인식모델에 대한 탐색시간이 대부분이다. 본 논문에서의 대분류기법이란, 상태수가 다른 HMM모델을 다단계로 적용하여, 탐색 공간을 줄이는 방법이다. 탐색공간을 줄이기 위해 적은 상태수의 HMM모델을 이용해 후보를 선발하는데 ,이 때 후보의 개수와 관련 특징파라미터 개수의 설정을 유동적으로 변화시켜 가장 이상적인 경우를 찾는다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서의 음성인식 시스템의 구조에 대해 살펴본다. 3장에서는 음성인식시스템의 속도향상을 위한 계층적 대분류기법의 개념과 방법에 대해서 살펴본 후, 4장에서 이를 기반으로 한 인식률과 속도에 대한 실험과 결과를 보인다. 이 결과를 기반으로 5장에서 결론을 맺는다.

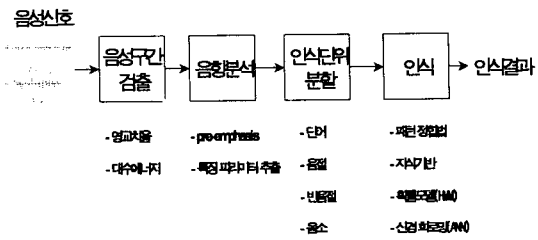


그림 1. 음성인식과정

2. 음성인식 시스템의 설계

2.1 음성인식기의 구조

음성인식 시스템의 구조는 음성구간검출 단계 , 음향분석단계, 인식단위분할단계, 인식 단계로 나뉘어 진다. 영교차율과 대수에너지 등을 이용하여 인식을 위한 음성구간을 검출해 낸 후 이를 전처리과정과 특징파라미터 추출과정을 통해 인식을 위한 정보를 추출해 낸다. 이러한 정보를 벡터양자화 과정을 통해 몇 개의 대표 벡터로 표현을 한 후, 이의 산출물인 1차원 코드색인열을 HMM모델을 이용하여 학습시킨다. 본 논문의 인식기는 1445개의 중권종목단어를 HMM모델로 학습시켰다.[2]

이 외에도 특징파라미터 추출 후 대표패턴을 가지고 동적 정합을 하는 방법(DTW)과 신경망에 의한 방법(ANN)등으로 인식을 할 수 있다.[3][5]

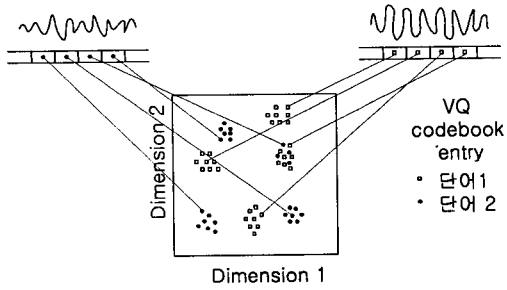


그림 2. 벡터양자화

2.2. 음향분석과 인식기의 학습

본 시스템은 A-law coding형식으로 입력된 발성음을 linear PCM으로 변환 후, energy를 이용하여 끝점검출을 한다. 이 후 5개의 특징파라미터(5 streams)를 사용한다. 12 차 Mel-Cepstrum과 12 차 Delta, 12 차 Acceleration Cepstrum, 그리고 3차의 Merged Energy(energy + delta energy + delta delta energy)와 청각신호가 반영된 5차의 PLP (perceptual linear prediction) 와 RASTA Cepstrum을 이용하여 코드북을 생성한다.

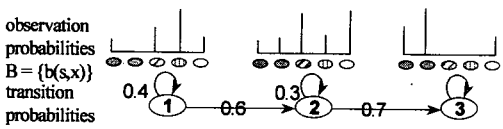
코드북의 생성은 벡터양자화(VQ ,Vector Quantization)과정을 통해 이루어 진다. 이는 단어를 몇 개의 대표 벡터로 표현하여 정보를 압축시킨다. 방법은 가장 가까운 대표벡터와의 왜곡거리의 합을 확인점수로 사용하는 방식이며, text-dependent에서도 사용가능하다. Mel-Cepsrum, Delta-Cepsrum, Acceleration Cepstrum, PLP는 256차의 코드북을 생성하며, Merged Energy는 64차의 코드북을 생성한다. 이의 결과물은 Code word sequence이며, 1차원 코드색인열로 표현이 된다.

인식기의 학습은 HMM모델을 이용한다. 계층적 대분류기법을 이용하기 위해 각 단어당 두종류의 HMM 학습을 시킨다. 1단계는 후보를 선출하기 위해서 3개의 상태를 가진 HMM모델을 학습시킨다. 2단계는 각각의 단어의 음절길이에 부합하는 상태수의 HMM 모델을 학습시킨다. 앞으로 이를 가변적 상태모델이라 하겠다. 단어의 음절길이에 따라 2음절일 경우는 7개, 3,4음절일 경우는 15개, 5음절이상일 경우는 21개의 상태수를 같는다. 3 상태 HMM과 가변적 상태 HMM은 각각 단어의 개수인 1445개가 형성된다. 이와 같이 두가지로 학습시키는 이유는 HMM모델이 상태수에 따라 탐색하는 시간이 큰 차이를 보이기 때문이다. 1445개의 가변상태 HMM을 탐색하는데는 많은 시간이 걸린다.

이를 극복하기 위해 적은 3개의 상태수로 후보(N-best)를 정한 후, 단어에 부합하는 가변 상태 HMM모델로 해당 단어를 선출하는 방식을 제안한다. 긴 처리시간을 요구하는 가변상태 HMM 탐색을 모든 단어에 대해 적용하지 않으므로 많은 시간을 줄일 수 있다. 자세한 대분류기법의 적용방식에 대해서는 3장에서 설명하겠다.

2.3. SL2 증권정보 DB

본 시스템에서 학습에 사용한 DB는 SL2 증권정보 DB이다. SL2 증권정보 DB는 A-law coding형식으로 1445개의 단어를 각각 160명의 남녀화자발성한 음성녹음파일 데이터베이스다. 단어별로 분할되어 있으며, 서울, 대전, 대구, 부산, 경기, 광주간 전화망상에서 수집되었다.



$$HMM(\lambda) := (A, B, \pi)$$

$$\Pr(O \mid \lambda) = \sum_{S} \Pr(O, S \mid \lambda)$$

$$= \sum_{S} a(\pi, s^1) \prod_{t=1}^{T} b(s^t, o^t)\, a(s^{t-1}, s^t)$$

그림 3. HMM 모델

3. 인식실험에 대한 계층적 대분류 기법의 적용

학습단계와 동일한 음성검출단계를 거친 후 5종류의 특징파라미터를 추출한다. 이를 12차 코드북을 통과시켜 1차 코드 색인열 생성한다. 생성된 색인열은 1차적으로 1445개의 3상태 HMM을 통과하며 우도비값을 구해낸다. 구해낸 우도비값 중 N개의 후보를 추출한다. 추출된 N개의 후보는 가변 상태 HMM을 적용하여 최대 우도비 값을 해당 단어로 인식한다. 후보를 추출할 때에는 적용되는 특징파라미터의 개수를 다양한 경우로 분류하여 실험하였다. 4장에서는 후보의 개수와 적용하는 특징파라미터의 종류를 다양하게 변화시켜 가장 최적화된 결과를 얻어내는 실험을 수행하였다.

4. 실험 및 결과

4.1. 실험방법

실험은 3상태 HMM으로부터 추출된 후보의 개수를 100, 50, 20, 10으로 다양하게 변화시켰다. 그리고 후보를 뽑을 때 특징파라미터의 개수도 다양한 경우로 실험하였다. 후보의 개수와 파라미터의 개수 모두 인식시간과 인식율에 영향을 미치는 요소이다. 이를 적절히 조절함으로서 이상적인 인식속도와 인식율을 구해내는 것이 목표이다. 표1은 N-best 인식테스트를 위한 후보의 개수와 특징파라미터의 개수 그리고 계층의 개수별로 실험요소를 정리하였다. Mel Cepstrum은 M, Delta Mel-Cepstrum은 D, A(Accelation) Mel-Cepstrum은 A, Merged Energe는 E, PLP(Perceptual linear prediction + RASTA) Cepstrum는 P로 표현하였다.

4.2. 실험조건

장비는 컴팩서버 5500을 이용하였다. 사양은 Dual CPU(550Hz) , 1G의 RAM과 ,18G의 HDD이다. 이러한 장비에 60회선 analog Dialogic Board를 이용하여 60회선이 동시에 접속되어, 60개의 쓰레드가 처리되면서 걸리는 시간을 측정하였다. 테스트 인원 및 테스트 환경은 무작위 남자10명, 여자10명이 주식종목명 1445개 발성하였으며, 일반 가정환경 및 회사환경에서 발성하였다.

4.3. 실험결과

대분류기법을 적용하지 않은 모델의 인식율은 94.8%가 나왔으며, 평균 인식시간은 7.1초였다. 이를 기준으로 표2의 결과를 평가해 볼 때 대분류기법이 속도향상에 효율적임을 보여준다. 결과 중 가장 이상적인 값은 100,10-이,오의 결과이며, 인식율은 1.3% 낮아졌지만, 속도는 거의 10배가 빨라졌다.

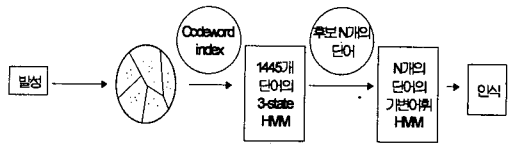


그림 4. 대분류기법의 적용

후보개수 -특징파라미터개수	N-best Modeling 과 특징파라미터 종류		N단어에 대한 Full state HMM 적용
100-오	100 Best 특징파라미터: M, D, A, E, P		100단어 Full state HMM 적용
50-오	50 Best 특징파라미터: M, D, A, E, P		50단어 Full state HMM 적용
10-오	10 Best 특징파라미터: M, D, A, E, P		10단어 Full state HMM 적용
100-이	100 Best 특징파라미터: M, D		100단어 Full state HMM 적용
50-이	50 Best 특징파라미터: M, D		50단어 Full state HMM 적용
10-이	10 Best 특징파라미터: M, D		10단어 Full state HMM 적용
100,10-이,오	100 Best 특징파라미터: M, D	10 Best 특징파라미터: M, D, A, E, P	10단어 Full state HMM 적용
100,10-이,삼	100 Best 특징파라미터: M, D	10 Best 특징파라미터: M, D, A	10단어 Full state HMM 적용
50,10-이,오	50 Best 특징파라미터: M, D	10 Best 특징파라미터: M, D, A, E, P	10단어 Full state HMM 적용
50,10-이,삼	50 Best 특징파라미터: M, D	10 Best 특징파라미터: M, D, A	10단어 Full state HMM 적용
100,20-이,오	100 Best 특징파라미터: M, D	20 Best 특징파라미터: M, D, A, E, P	20단어 Full state HMM 적용
100,20-이,삼	100 Best 특징파라미터: M, D	20 Best 특징파라미터: M, D, A	20단어 Full state HMM 적용
50,20-이,오	50 Best 특징파라미터: M, D	20 Best 특징파라미터: M, D, A, E, P	20단어 Full state HMM 적용
50,20-이,삼	50 Best 특징파라미터: M, D	20 Best 특징파라미터: M, D, A	20단어 Full state HMM 적용

표 1 . 실험과정

5. 결론

본 논문은 기존의 음성인식 시스템의 인식율을 유지하면서 속도를 개선시키는 연구였다. HMM모델에 계층적인 대분류기법을 적용하였다. 많은 상태를 천이하는데 걸리는 시간을 줄이는데 목적을 두었다. 상태수를 3개로 최소화시킨 HMM모델을 통해 후보를 선발하고, 선발된 후보에 대해서만 모든 상태수를 가진 가변적 상태 HMM을 적용하였다. 후보수와 특징파라미터의 종류와 수를 모두 고려하여 다양하게 실험하여 이상적인 경우를 찾을 수 있었다. 보통 탐색공간을 줄이면 인식율이 불안정해진다. 그러나 본 연구에서는 안정적으로 처리시간을 줄였다는데 그 의미가 있다.

실험	인식율(%)	속도(sec)
100-오	93.7	2.97
50-오	90.9	1.91
10-오	86.7	0.81
100-이	87.8	2.33
50-이	87.8	1.35
10-이	85.5	0.38
100,10-이,오	**93.5**	**0.79**
100,10-이,삼	92.7	0.71
50,10-이,오	88.1	0.63
50,10-이,삼	87.9	0.59
100,20-이,오	93.5	1.19
100,20-이,삼	92.9	1.09
50,20-이,오	90.3	0.98
50,20-이,삼	89.4	0.81

표 2. 실험결과

참고문헌

[1] H, Ney, D. Mergel, and A. paeseler, "Data Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition," IEEE Trans. on Signal Processing, Vol.40,No.2, pp.272-281, Feb 1992.

[2] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition," Prentice-Hall 1993.

[3] X.D Huang, Y. Ariki, M.A. Jack. "Hidden Markov Models for Speech Recognition,"

[4] E.G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. "Interpolation of maximum likelihood predictors in stochastic language modeling," Eurospeech' 97, pp2731-2734, 1997.

[5] Simon Haykin, "Neural Networks."

[6] V. Warnke, S. Harbeck, E. Noth, H. Niemann, and M. levit. "Discriminative estimation of interpolation parameters for language model classifiers," ASSP'99, pp. 525-528, 1999.

[7] 최환진, 윤성진, 오영환, "음향분절모델과 상대거리에 기반한 평활화를 이용한 한국어 단어 인식에 관한 연구", 한국정보과학회지, 제 23권, 제2호, pp.168-176, 1995