HMM 부모델을 이용한 단어 인식에 관한 연구

신 원 호*, 김 원 구**, 윤 대 희*, 차 일 환*
*연세대학교 전자공학과, **군산대학교 전기공학과

A Study on Word Recognition using sub-model based Hidden Markov Model

Won-Ho Shin*, Weon-Goo Kim**, Dae-Hee Youn*, Il-Whan Cha*
Dept. of Electronics Eng. Yonsei Univ.
Dept. of Electrical Eng. Kunsan National Univ.

abstract

In this paper the word recognition using sub-model based Hidden Markov Model was studied. Phoneme models were composed of 61 phonemes in terms of Korean language pronunciation characteristic.

Using this, word model was maded by serial concatenation. But, in case of this phoneme concatenation, the second and the third phoneme of syllable are overlapped in distribution at the same time. So considering this, the method that combines the second and the third phoneme to one model was proposed. And to prevent the increase in number of model, similar phonemes were combined to one, and finially, 57 models were created. In experiment proper model structure of sub-model was searched for, and recognition results were compared. So similar recognition results were maded, and overall recognition rates were increased in case of using parameter tying method.

I 서 론

인간이 기계에게 의사를 전달하는 방법은 여러 가지가 있으나 가장 자유스러운 방식으로는 인간과 기계 사이의 직접적인 대화를 고려할 수 있다. 현재의 키보드를 통한 인간과 기계의 의사 소통(man machine interface)을 가장 초보적인 단계라고 한다면, 음성을 통한 대화(음성 인식, 음성 합성)는 가장 진보된 방식이라 할 수 있다.

음성 합성이 주어진 문장을 음성으로 바꾸어 내는 작업이라면 음성 인식은 음성을 문장으로 변환하는 작업이라 할 수 있는데, 음성 인식이 완벽하게 이루어질 수 있다면 기계를 이용하는 사용자 측면에서 누리는 편리함은 헤아릴 수 없을 정도로 많아질 것이며, 공상 과학 영화에서 볼 수 있었던 인간과 대화하는 로보트의 출현도 기대할 수 있을 것이다. 물론 기계에 의한 음성 인식이 현재까지 해결되지 않은 여러 문제점 및 제한 요소로 인하여 광범위한 상용화는 이루어지지 않고 있지만 많은 사람들의 노력에 의하여 상당한 발전이 이루어 졌다.

음성 인식을 위한 연구는 dynamic time warping(DTW)[1], hidden markov model(HMM)[2], 신경 회로망등 꾸준히 이루어져 왔는데 본 논문에서는 현재 음성 인식에 가장 많이 이용되고 있는 HMM을 이용하여 부모델(sub-model)[7] 단위의 HMM을 이용한 단어 인식 시스템을 구성하였다.

또한 우리말에 적합한 단위 모델을 설정하고자 하였으며, 구성한 시스템에 대한 성능을 비교하여 성능 개선 방법을 제시하였다.

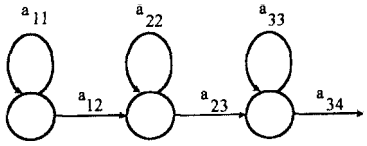
II에서는 HMM의 음소 및 단어 모델 구성에 대하여 서술하였다. 즉 실험에 이용한 모델 구조와 모수 tying방법, 중성과 종성 연결한 모델 구성 방법을 설명하였으며 III에는 실험 결과를 비교하고, IV에서 결론을 맺었다.

II HMM의 음소 및 단어 모델 구성

A. 음소 HMM의 모델 구조

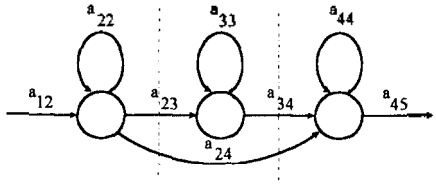
HMM은 관측 확률 분포를 어디에서 얻느냐에 따라 두 가지 다른 형태의 구조가 존재하는 데, 상태에서 얻는 경우를 Moore machine(state emitter), 상태 이전(transition)중에 얻는 경우를 Mealy machine(transition emitter)라 한다[9]. 일반적으로 두개의 모델을 접속하는 경우 접속이 용이한 Mealy form을 이용하는 것이 보편적이다. 그러나 상태 이전 경로가 많아질 경우 이에 따라 관측 확률 분포가 늘어나는 단점을 가지고 있다. 인식해야 할 어휘의 수가 많아질 경우 부모델 단위의 모델을 연결하여 단어 등의 모델을 구성하는데 우리말의 경우 음소로 음소(phoneme)로 구분할 수 있는 초중,종성을 이용하여 부모델을 구성하는 것이 일반적이다. 그림[1]은 실험에 이용한 일반적인 형태의 HMM 모델 구조이다.

본 연구는 한국과학재단의 연구비 지원으로 이루어 졌읍니다.



그림[1] 일반적인 형태의 HMM 모델 구조

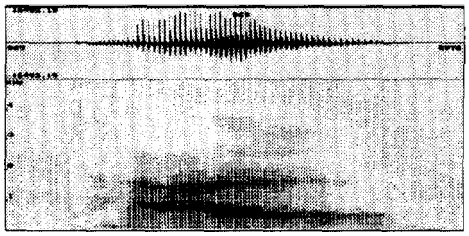
이때 상태의 개수나 천이 정도는 음소의 주파수 특성 및 그 분포 변화를 이용하여 결정하게 된다. 또한 위에서 Mealy form의 단점으로 지적한 상태 수 증가에 따른 확률 분포 수의 증가를 막기 위한 방법으로 모수 tying[5][8]을 이용할 수 있다. 예를 들어 음소의 상태 분포를 좌우 변화 구간 및 안정 구간의 3가지[4]로 나눌 경우 이를 그림[2]와 같이 나타낼 수 있다. 각 상태 수에 따른 이의 분할을 아래에 나타내었다.



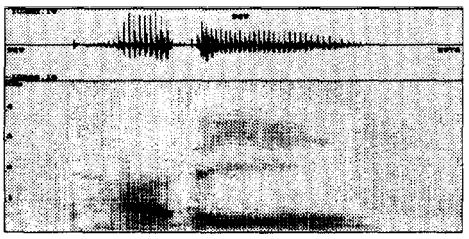
그림[2] 모수 tying을 이용한 HMM 모델 구조

B. 제안한 부모델(sub-model) 구조

앞에서 언급한 바와 같이 초, 중성 및 종성의 음소를 이용하여 구성할 경우 가장 문제되는 것은 연결 부분에서 발생되는 조음현상(coarticulation)이다. 특히 문맥 독립(coutext independent)적인 모델을 이용하는 경우 이 문제는 피할 수 없으므로 대용량의 인식 시스템에서는 문맥 의존적인(context dependent) 모델을 이용한다[6]. 그런데 우리말의 경우 발음 현상이 음절을 중심으로 구분되어지며 중성과 종성의 경우에서는 같은 시간대에 주파수 영역에서 중첩되어 나타남을 알 수 있다. 아래에 "다리"와 "달"이란 신호를 살펴보면 음소 연결을 통해 구성할 경우 "ㄷ ㅏ ㄹ ㅣ", "ㄷ ㅏ ㄹ"와 같이 유사한 구조로 연결이 되어지나 주파수 영역 분포에 있어서는 그 형태가 다르며 달의 경우 "ㅏ"와 "ㄹ"이 중첩되어 있음을 볼 수 있다. 따라서 문맥 의존 모델의 이전 단계로 중성과 종성의 음소를 결합한 경우를 이용하면 음소 연결 시에 나타나는 문제점을 어느 정도 해결할 수 있다. 또한 중성과 종성을 함께 모델링하므로 접속시키는 모델의 수도 간소화 시킬 수 있게 된다.



그림[3] "달"의 신호와 스펙트럼



그림[4] "다리"의 신호와 스펙트럼

이 때 단점으로는 각각의 중성과 종성의 조합에 따른 경우를 따로 모델링해 주어야 하므로 모델의 수가 증가하게 된다. 본 논문에서는 실험에 이용된 61개의 음소를 위에 설명한 방법을 이용하여 모델을 구성하였으며 이때 61개의 음소중 발음 특성이 유사한 것을 통합하여 57개의 모델을 생성하였다. 이 때 고려해야 할 사항으로는 상태 수인데 일반적인 음소를 이용하여 모델을 구성할 경우 동일하거나 비슷한 수의 상태 개수를 이용하게 된다. 그러나 중성과 종성이 함께 연결된 경우 음운 현상이 보다 길어지므로 상태의 수를 달리하여야 한다. 모수 tying을 이용한 음소의 경우 전후에 변화 구간과 안정 구간의 3가지 형태를 갖으며 자음과 같이 지속 시간이 짧은 경우 두개의 분포 형태를 갖기도 한다. 따라서 중성과 종성이 연결된 경우에는 보다 긴 상태 수를 갖도록 해야 하며, 각 모델에 따라 다른 개수의 상태를 갖는 모델 구조로 변화시켜 주어야 한다.

III 실험 및 결과 고찰

A. 음성 신호 전 처리, 특징 추출 및 데이터 베이스의 구성

본 논문에서는 계산량 및 모수의 용량이 다른 형태에 비해 작은 이산 분포 HMM을 이용하여 모델을 구성하였다. 실험에 이용할 데이터를 위하여 한 남성 화자가 조용한 실험실 환경에서 발음한 음성 신호를 10 Khz로 샘플링하였다. 이를 0.95의 값으로 프리엠퍼시스(pre-emphasis)[2]한 후 14차의 lpc(linear predictive coding) 켑스트럼(cepstrum)을 구하였다. 실험에 이용된 특징 벡터는 위에 구한 켑스트럼과 이의 차분 값을 이용한 차등 켑스트럼(differenced cepstrum)[8]이다. 차등 켑스트럼은 전후 20ms의 시간 차를 갖는 켑스트럼 간의 차를 이용하였다. 이산 확률 분포를 갖는 HMM을 구성하기 위해 에너지 등의 특징 벡터를 이용하여 녹음된 신호의 음성 구간을 추출하였다. 이를

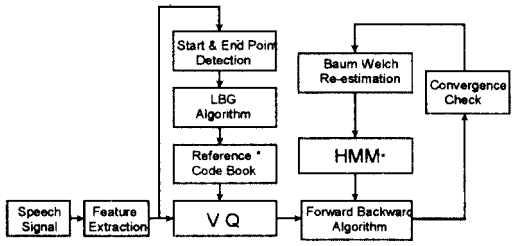
LBG algorithm[3]을 이용하여 256개의 code를 갖는 벡터로 구성하였다. 그리고 얻어진 codebook을 이용하여 학습 데이터들을 부호화 하였다. 이 때 챌스트럼의 경우 신호 전후의 묵음 및 소음 구간에 해당하는 챌스트럼을 codebook에 추가하여 258개의 codebook을 이용하는 방식을 취하였다. 그러할 경우 묵음과 음성 부분과의 구분을 보다 명확히 할 수 있게 된다. 차등 챌스트럼의 경우에는 256 code를 그대로 이용하였다. 표[1]은 학습에 이용된 61개의 단어이다[10].

바람	입술	나비	빨래
피리	다리	받고	과도
딸	튼	가을	기둥
동굴	까닭	칼	사라
이제	찌개	처음	소리
쉬파리	시간	싸리	씨알
하늘	동해	마음	나리
예닐곱	등지	달	구리
달력	이웃	이발	메밀
세상	애기	해방	쉬나물
쉬항	위문	뒤	되풀이
된장	외길	외국	바다
말씀	어머니	열	건강
고리	노인	눈동자	분사람
글	근세	양식	완성
의사			

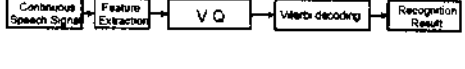
표[1] 실험에 이용한 단어의 구성

B. 음성 인식 시스템의 구성

위에 설명한 HMM을 이용하여 음성 인식 시스템을 구성하였다. 학습에는 Baum-Welch 알고리즘[2]을 이용하였다. 각 모델의 초기 모델은 균등한 확률 분포를 갖는 형태로 초기화하였다. 인식은 Viterbi decoding[2] 방법을 이용하여 최대 확률 값을 갖는 경우를 구하도록 하였다. 그림[5][6]에는 학습 및 인식의 블럭다이어그램을 나타내었다.



그림[5] 학습 과정 블럭다이어그램



그림[6] 인식 과정 블럭다이어그램

C. 결과 고찰

실험에서는 61개의 모델 구조와 제안한 57개의 모델 구조를 갖는 경우에 대하여 상태 수의 변화에 따른 인식률의 변화를 비교하였다. 먼저 tying을 고려하지 않은 경우 그림[1]과 같은 구조를 이용하였는데 음소(또는 단위 모델)와 전후의 묵음 구간 모델의 상태 수에 따른 인식 결과를 비교하였다.

표[2]는 61단어 구조를 이용한 경우에 대한 인식 결과인데, 묵음 모델의 경우, 이의 분포가 음소의 경우보다 단순한 경우가 많으므로 작은 수의 상태 수를 가진 경우가 보다 적합하리라고 생각할 수 있다. 하지만 상태의 수가 너무 적은 경우에는 인식률을 오히려 저하시키는 결과를 초래함을 볼 수 있다. 실험에서는 챌스트럼만을 이용한 경우와 차등 챌스트럼을 함께 이용한 경우를 비교하였는데 차등 챌스트럼을 함께 이용한 경우가 인식률에 있어서도 좋은 결과를 보여 주었다. 따라서 이 두 가지 특징 벡터를 모두 고려하기로 하였다. 실험에서는 학습때 나타나지 않은 데이터를 고려하여 최저 관측 확률 분포를 $\frac{1}{10^6}, \frac{1}{10^5}, \frac{1}{10^4}, \frac{1}{10^3}$로 제한한 경우에 대하여 비교하였다. 표[2]의 괄호 안의 값은 2번째 후보까지를 고려한 인식 결과이다.

묵음 모델 상태 수 \ 음소 모델 상태 수		3	4
2	1.e-6	91.3 (98.9)	92.9 (97.8)
	1.e-5	92.3 (98.4)	94.0 (97.8)
	1.e-4	93.4 (98.4)	94.5 (97.8)
	1.e-3	94.0 (97.8)	95.1 (98.4)
	avg.	92.75 (98.375)	94.125 (97.95)
3	1.e-6	91.8 (98.4)	92.3 (97.8)
	1.e-5	92.3 (98.4)	94.0 (98.9)
	1.e-4	92.9 (98.9)	95.1 (98.9)
	1.e-3	94.0 (97.8)	95.6 (98.9)
	avg.	92.75 (98.375)	94.25 (98.625)
4	1.e-6		94.0 (97.3)
	1.e-5		94.5 (97.3)
	1.e-4		94.5 (97.8)
	1.e-3		95.1 (97.8)
	avg.		94.525 (97.55)

표[2] 모델 상태 수에 따른 인식 결과

제안한 57개의 모델 구조를 갖는 경우에 대해서도 이와 동일한 실험을 수행하였는데, 표[2]의 결과와 거의 같은 결과를 얻었으나 큰 향상은 없었다. 다음으로는 모두 tying 방법을 이용하여 실험하였다. 이 경우에도 음소(단위) 모델과 묵음 모델의 상태 수를 변화시켜 가며 실험하였는데 큰 향상은 얻을 수 없었다. 이처럼 인식률의 향상이 두드러지지 않은 이유 가운데 한가지는 각 모델당 동일한 상태 수의 설정을 생각할 수 있는데, 다음 실험에서는 제안한 모델의 지속 시간을 고려하여 다른 상태 수를 갖도록 지정하여 학습하였다. 그 결과 61개의 모델 구조를 이용하여 얻은 인식 결과보다 향상된 결과를 얻었다. 이때 각 모델 상태 수를 초성의 경우 3개, 중성과 종성이 연결된 경우 4개의

구분된 분포 특성을 갖도록 설정하였다. 다음 표[3]은 이의 인식 결과이다. 실험에서는 이들 상태 수를 달리 설정하여 인식 결과를 비교하였으나 아래의 경우 가장 좋은 결과를 얻었다.

최저 확률	1.e-6	95.1 (98.4)
	1.e-5	95.1 (98.9)
	1.e-4	96.2 (98.4)
	1.e-3	96.2 (98.4)

표[3] 상태 수를 달리한 경우, 제안한 모델 구조의 인식 결과

이상에서 살펴본 바와 같이 제안한 모델 구조의 경우 각 모델에 적합한 상태 구조를 할당할 경우, 61개의 모델 구조에 비하여 향상된 인식률을 보임을 알 수 있었다. 따라서 각 모델 상태 수의 설정이 인식에 영향을 미침을 알 수 있는데, 상태 변환 확률 A_{ij}의 경우 a_{ii}의 값이 상태 i에 지속하는 정도를 추정하게 된다. 상태 i에 머무르는 평균 지속시간은 $1/(1 - a_{ii})$으로 모델 상태 수를 충분히 크게 한 뒤 학습된 모델의 상태 변환 확률 행렬 대각 성분을 분석하여 보면 몇 개 정도의 상태 수를 갖는 것이 적절한 지 추정할 수 있다. 그러나 상태 수를 적절히 선택한다고 하여 음운 현상을 적절히 모델링하였다고 할 수는 없을 것이다. 본 논문에서는 지속 기간 등에 대하여는 고려하지 않았는데 고려하지 않은 사항들이 전체 인식 성능에 미치는 영향 등도 감안하여야 하기 때문이다. 따라서 앞으로 이러한 요인을 함께 고려하여 정확한 음소의 음운 특성을 모델링하도록 하여야 할 것이다.

VI 결 론

본 논문에서는 부모델(sub-model) 단위의 Hidden Markov Model을 이용하여 단어 인식을 위한 연구를 수행하였다.

우리 말을 발음 특성에 따라 61개의 음소로 분류하여 음소 모델을 구성하였으며, 이들을 이용하여 직렬 접속(serial concatenation)의 방법으로 단어 모델을 구성하였다. 그러나 이들이 직렬로 연결될 경우 우리말의 중성과 종성이 스펙트럼 분포에 있어 거의 같은 시간대에 분포함을 고려하여 중성과 종성을 하나의 모델로 구성하는 방법을 제안하였다. 이에 따라서 증가되는 모델을 개수를 보완하기 위해 61개의 음소중 유사한 것들을 하나의 모델로 통합하여 57개의 모델을 생성하였다.

실험에서는 각 부모델에 적합한 모델 구조를 설정하고자 하였으며, 모델 구조에 따른 인식 결과를 비교하였다. 또한 위에 제안한 구조를 실험하여 61개의 모델 구조의 경우와 거의 동일한 인식률을 얻었으며 모수 tying을 이용한 경우 전체적으로 향상된 결과를 얻었다.

참고 문헌

[1] H.Sakoe, S. Chiba,"Dynamic programming optimization for spoken word recognition," *IEEE Trans on ASSP Proc.,ASSP-26(1)* 43-49 Feb 1978

[2] L.R.Rabiner "A Tutorial To Hidden Markov Models And Selected Applications in Speech Recognition",*Proc of IEEE*, Vol. 77,No 2,pp.257-285 Feb 1989

[3] Y. Linde, A. Buzo, R.M. Gray,"An Algorithm for Vector Quantization",IEEE Trans. Com., Vol. Com-28,No.1,pp. 84-95,1980

[4] K.F.Lee, Large-Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINIX System, Ph.D. dissertation, Computer Science Department, CMU, 1988

[5] K.F.Lee, H.W.Hon, R.Reddy "An Overview of the SPHINIX Speech Recognition System" *IEEE Trans on ASSP* vol.38. No 1,Jan 1990

[6] R.Schwartz, et al. "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech" *Proc of ICASSP* 1985 vol 3 pp 1205-1209

[7] H.Murveit, M.Weintraub "1000 word Speaker-Independent Continuous Speech Recognition using Hidden Markov Models" *IEEE ICASSP* vol 1 pp 115-118 Apr 1988

[8] X.D.Huang, Y.Ariki, M.A.Jack edits: Hidden Markov Models for Speech Recognition Edinburgh University Press 1990

[9] J.R.Deller,Jr, J.G.Proakis, J.H.L.Hansen Discrete-Time Processing of Speech Signals pp. 680-681 Macmillan Publishing Company 1993

[10] 표준 한국어 발음 대사전 pp. 31-32 어문각