Sub-word 단위 HMM을 이용한 한국어 대용량 어휘 인식

김 홍 수, 이 상 운, 이 건 웅, 홍 재 근
경북대학교 전자공학과

Large Vocabulary Speech Recognition Using Sub-word Unit HMM

HongSoo Kim, SangWoon Lee, KeonUng Lee, JaeKeun Hong
Dept. of Electronic Engineering, Kyungpook National University
E-mail ; alex@speech.knu.ac.kr

요 약

일반적인 한국어 대용량 어휘인식에 사용되는 triphone 모델은 한국어의 특성을 잘 표현한다는 장점이 있으나 인식시간이 길어지게 된다. 이러한 triphone 모델의 단점을 극복하기 위해 음절단위 HMM 모델을 사용하는 방법이 있는데 이 모델은 인식시간을 줄일 수 있으나 triphone 모델에 비해서 인식률이 낮다.

본 논문에서는 음성 인식시간을 단축시키고 조음현상을 고려하기 위하여 초성과 종성 자음은 각각의 biphones으로 나타내고 중성 모음은 1개의 monophone으로 나타내는 모델을 제안하였다.

PBW445 음성 데이터베이스에 대한 실험결과, 제안한 인식모델이 triphone 모델에 가까운 인식률을 나타내었으며, 인식시간을 크게 단축하였다.

I. 서론

음성인식에 필요한 주변기기의 발달과 음성인식의 사회적 필요성에 의해서 음성인식 기술은 많은 발전을 이루었다. 또한, 최근 음성인식 기술은 대용량 음성 인식기[1] 구현과 실용화를 위한 연구로 발전하고 있다. 이러한 대용량 음성 인식기에 사용되는 인식단위로는 triphone[1][2]과 같은 문맥종속 단위가 가장 우수한 인식 단위로 알려져 있다.

한국어는 영어와 달리, 초성과 중성, 종성으로 구성되어 있고, 2~3개 음소의 결합으로 발생하는 조음현상이 심하다[3]. 이러한 조음변화를 고려하기 위해 음소단위의 문맥의존 triphone 모델을 주로 사용하는데 이 모델은 단어 내외에서 일어나는 조음현상을 잘 표현할 수 있다. 그러나 triphone 모델은 단어들의 조음현상을 잘 표현하지만, 모델의 수가 많아지며 인식시간이 길어지는 단점이 있다. 이러한 음소모델의 단점을 극복하기 위한 한가지 방법으로 음절단위의 HMM 모델을 만드는 것이 있는데[3], 음절을 기본단위로 했을 때 인식시간은 줄어들 수 있으나 인식률은 triphone 모델에 비해서 떨어지게 된다[4]. 이것은 음절단위의 모델이 음절과 음절사이의 조음현상을 triphone 모델보다 잘 나타내지 못하기 때문이다.

음성의 특성상 모음은 자음에 비해 안정한 구간이 길며, 모음에 의해서 받는 자음의 변화가 자음에 의한 모음의 변화보다 더 크게 나타난다. 제안한 방법에서는 이러한 특성을 이용하여 초성+중성+종성으로 구성된 음절에 대해, 초성과 종성 자음의 biphone 모델 1개씩과 1개의 중성 모음을 나타내는 monophone 모델로 구성하였다.

본 논문에서, 단어의 인식에 사용하는 biphone 모델은 모델의 수를 줄임과 동시에 자음의 음소 변화를 잘 표현할 수 있다. 그리고, monophone 모델은 중성 모음의 스펙트럼상의 안정성을 표현한다. 단어는 biphone 모델과 monophone 모델을 결합한 형태이다.

기존의 인식모델과 제안한 인식모델의 비교를 위해서, 동일한 조건에서 인식실험을 수행하였다. 실험 데이터는 35,600개의 단어로 구성된 PBW445 데이터베이스를 사용하였고, 훈련 및 인식실험을 위해서 HTK 2.2(HMM Tool Kit)를 사용하였다.

II. Sub-word HMM

2.1 Triphones and Syllables HMM

음소 단위 HMM은 대용량 어휘의 음성인식에서 주로 사용되어 왔다. 음소란 서로 구별되어 쓰이지 않는 음들의 집합이라고 말할 수 있다. 특히 한국어의 경우에는 초성, 중성, 종성으로 구성된다. 초성과 종성은 자음으로, 중성은 모음으로 이루어진다. 초성자음의 개수는 19개이고, 중성모음은 20개, 종성자음은 7개이다. 본 실험에서 한국어 음성인식에 사용되는 음소의 개수는 묵음을 포함하여 41개이다.

이러한 음소모델은 대용량 어휘 인식에서 음성의 조음현상을 충분히 표현하지 못한다. 음성은 인접한 앞, 뒤 음성에 영향을 많이 받는다. 동일한 음소라도 전후의 환경에 따라 발생하는 다양한 음성학적 변화에 민감하다. 그래서 같은 음소에 대해서도 앞, 뒤 음성에 영향을 받아 그 발음이 조금씩 다르게 나타난다[5]. 이런 조음현상을 고려하기 위한 한 방법은 문맥의존 triphone HMM 모델을 사용하는 것이다. triphone이란 하나의 음소를 독립적으로 훈련 및 인식시키는 것이 아니라, 훈련 및 인식하고자 하는 음소의 문맥상 인접한 좌, 우 음소를 고려하는 것이다. 그 예로 표 1의 '가운데' 음성의 인식모델을 들 수 있다(**g**+a g-**a**+u a-**u**+n u-**n**+d n-**d**+e d-**e**). 이와 같이 인접한 음소를 고려하는 triphone HMM은 문맥상 이어지는 음성의 변화를 잘 모델링할 수 있다.

그러나, 조음현상을 잘 나타내기 위해서는 HMM 모델의 수가 많아지게 된다. 이를 위해 이론상으로 3000개 정도의 모델을 만들어야 한다. HMM 모델의 수는 음성인식시간에 비례하므로 그만큼 인식시간이 길어진다. 실제로 문맥의 제약으로 1000~2000개의 모델이 사용되고 유사한 HMM 모델들의 파라미터값들을 묶음으로써, 전체 모델의 수를 줄일 수 있으나 그 한계가 있다.

음성인식 시간을 단축하기 위한 방법으로 음절단위의 HMM을 만드는 방법이 있다. 음절의 정의는 더 이상 쪼갤수 없는 최소의 발음 가능한 단위이다. 음절은 2~3음소의 결합으로 구성되고[5][6], 음절내의 조음현상을 포함하기 때문에 고립단어 인식에서는 단어를 쉽게 음절 단위로 분해하고 구성할 수 있다[7]. 따라서 모델의 수는 줄어 들게 된다. 음절단위의 인식모델을 만들었을 때, 그 예는 표 1에서 확인할 수 있다

'가운데' ga + un + de

음절단위 HMM모델을 사용한 음성 인식기는 모델의 수가 적어서 인식시간을 단축시킬 수 있으나, triphone HMM을 이용한 음성 인식기보다 인식률이 저하된다. 이는 음절단위 모델이 음절과 음절사이의 음성변화를 잘 나타내지 못하기 때문이다. 따라서 음절의 HMM 모델수를 유지하면서, 음절간의 조음현상을 포함할 수 있는 새로운 인식단위가 필요하다. 새로운 인식단위 HMM 인식기는 triphone HMM 인식기와 같이 조음현상을 잘 나타내며 또한, 음절단위 HMM 인식기의 장점인 인식시간 단축을 동시에 가질수 있어야 한다.

2.2 Biphones and monophones HMM

현재의 대용량 어휘의 음성인식에서 인식시간의 단축과 인식률의 향상은 중요한 요소이고, triphone HMM과 음절단위 HMM은 서로 상반적인 관계를 가진다. 따라서, 본 논문에서 두 모델 각각의 장점을 가질 수 있는 인식단위를 제안하였다. 제안한 인식단위는 아래와 같이 자음으로 구성된 biphone HMM과 모음으로 구성된 monophone HMM을 가진다.

x	문맥 독립적 모음모델
x+y	오른쪽 y를 고려한 x 자음모델
x-y	왼쪽 x를 고려한 y 자음모델

초성과 종성의 자음은 인접한 모음이나 자음의 영향이 크고 그 변화도 크기 때문에 biphones HMM으로 모델을 훈련 및 인식시키고, 중성모음은 자음보다 발음시간이 길며 스펙트럼 변화가 적고 안정하기 때문에 하나의 독립된 음소모델로 인식단위를 훈련 및 인식시킨다. 이 모델을 인식단위로 사용한 HMM과 triphone 및 음절단위의 HMM의 비교를 표 1에 나타내었다.

표 1은 PBW445 음성데이타 중 '가운데' 음성을 각 인식단위별로 단어사전을 비교한 것인데, 이 음성데이타의 음소적 구분은 /g + a + u + n + d + e/ 이다. 표 1에서도 볼 수 있듯이, 제안한 모델은 그 모델수를 줄일 수 있으며 음성의 조음현상을 잘 표현할 수 있다.

표 1. 단어 '가운데'의 인식단위별 단어사전 비교

인식단위	triphones	biphone & monophone	syllable
단어사전	g+a	g+a	ga
	g-a+u	a	
	a-u+n	u	un
	u-n+d	u-n	
	n-d+e	d+e	de
	d-e	e	

III. 실험 및 결과

3.1 실험

본 연구에서는 PBW445 음성 데이터베이스를 사용하였다. 이 데이터베이스는 남성화자 21명과 여성화자 19명 각각이 445개의 단어를 2번씩 발음한 음성데이터이며 총 단어수는 35,600개이다.

인식실험은 HMM을 기반으로 한 음성인식 시스템 구현 및 실험을 위한 상용도구인 HTK2.2를 사용하였다. 음성신호의 앞, 뒤 묵음구간을 제거하기 위해 끝점 검출과정을 거친 후, 15ms Hamming window를 사용하여 5ms씩 중첩하여 12차 LPC 켑스트럼, 델타 켑스트럼, 델타 델타 켑스트럼 그리고 에너지, 델타 에너지와 델타 델타 에너지 파라미터들을 구하였다. 훈련과정에서 각각의 모델은 그림 1과 같이 left-to-right 3상태 HMM 모델을 사용하였다. 그림 1에서 처음 start상태와 끝 end상태는 각각의 모델들을 연결하는 역할을 한다. 상태 1에서 상태 3으로 천이되는 것은 묵음구간에서만 사용된다.

HTK는 인식단위의 경계 정보가 없는 연속으로 발성된 음성 데이터를 하위 단어 단위 HMM 모델의 훈련에 사용이 가능하므로 음성신호를 라벨링하는 단계를 거치지 않고 flat start procedure로 실험하였다. 실험에서 HMM에 의한 훈련은 Baum-Welch 알고리듬을 사용하여, 인식단위 HMM 모델을 재추정하였다. 각 인식단위별 HMM 훈련에서는 동일한 방법을 적용하였고, 실험은 동일한 환경에서 각각의 인식단위별(triphones, 음절단위, 제안한 인식단위) 인식률과 인식시간을 측정하였다.

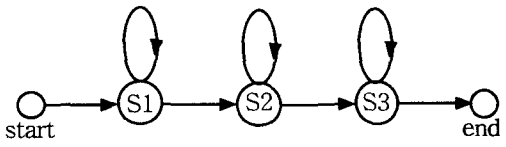

그림 1. 5상태 HMM 모델

3.2 실험결과

인식단위별 인식실험에는 PBW445 음성데이타 중에서 남성화자 3명과 여성화자 2명을 사용하였다. 인식에 사용되는 5명의 화자를 제외한 나머지 화자 35명의 음성데이타는 훈련에 사용하였다..

인식단위별 인식률 실험결과는 HTK의 인식률 산출 방법을 따른다. 인식률은 인식실험에 사용된 단어의 총수 (N), 인식된 단어수 (H), 삭제된 단어의 수 (D), 대체된 단어의 수 (S), 삽입된 단어의 수 (I)를 이용하여 다음과 같이 계산한다.

$$Correct = \frac{N-D-S}{N} \times 100\ (\%)$$

$$Accuracy = \frac{N-D-S-I}{N} \times 100\ (\%)$$

인식단위별 인식률과 인식시간 그리고 HMM 모델의 수를 표 2에 나타내었다. 인식단위별 인식률은 triphone 모델의 경우 97.2%, 음절의 경우 93.4%를 나타낸다. 그리고 제안한 인식단위를 사용한 인식률은 96.18%이다. 인식률은 triphone HMM이 가장 좋고, 음절단위 HMM이 가장 낮다. 표 2의 음성인식 시간은 사용하는 기기의 성능에 따라 달라지기 때문에, triphones HMM을 기준하여 음절단위 HMM과 제안한 인식단위 HMM의 인식시간을 상대적으로 측정하였다. 음성인식 시간은 triphone HMM을 기준으로 음절단위 HMM은 47%, 제안한 인식단위 HMM은 43%를 나타낸다. 동일한 조건하에서 인식시간은 인식단위 HMM 모델의 수와 비례함을 알 수 있다. 모델의 수는 이론상으로 나타나는 수가 아닌, PBW445 데이터베이스에서 문맥의존 HMM 모델의 수이다. 예를 들어 이론적 triphone 모델의 수는 3000여개이나, 실제로 본 실험데이타베이스에서 추출된 모델의 수는 묵음을 포함하여 1515개이다.

표 2. 인식단위별 인식률과 인식시간 비교

인식단위	triphone	biphone & monophone	syllable
단어인식률	97.2%	96.18%	93.4%
인식시간	1	0.43	0.47
unit 갯수	1515개	325개	475개

IV. 결론

본 논문에서는 대용량 어휘 고립단어 인식을 위하여 triphone 모델과 같이 문맥의 변화를 잘 나타내어 높은

인식률을 유지할 수 있고, 모델의 수가 음절단위의 모델수와 유사하여 인식시간을 단축시킬 수 있는 biphone과 monophone의 조합으로 구성된 모델을 제시하였다. 초성과 종성 자음은 biphone 모델을 적용하였고 중성 모음의 경우는 monophone 모델로 나타내었다. 그리고 제안한 인식단위 모델의 훈련과 인식실험은 HTK 연속음성인식 시스템을 사용하였다. PBW445 음성데이터 인식실험 결과, 제안한 모델이 triphone 모델에 가까운 인식률을 나타내었고 인식시간은 triphone 모델보다 57.3% 단축되었다. 그리고, 제안한 모델과 음절단위 모델을 비교했을 때 인식시간은 단축되고 인식률은 3% 향상되었다. 실험결과를 통하여, 제안한 인식단위 모델이 triphone 모델과 같이 문맥의 변화를 잘 고려함을 알수 있었고, 음절단위 모델보다 모델의 수가 적어서 인식시간이 단축됨을 확인할 수 있었다. 제안한 방법에서 단어사이의 조음현상도 고려하면 대용량 연속음성 인식시스템에 적용할 수 있을 것으로 기대된다.

V. 참고문헌

[1] C.-H. Lee, J.-L. Gauvain, R. Rieraccini, and L.R. Rabiner. "Large vocabulary speech recognition using subword units," *Speech Communication 13*, pp. 263-279, 1993.

[2] J. M. Kessens, M. Wester, and H. Strik, "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciations variation," *Speech Communication 29*, pp. 193-207, 1999.

[3] 김유진, 김회린, 정재호, "인식 단위로서의 한국어 음절에 대한 연구," *한국음향학회지*, 제16권, 제3호 pp. 64-72, 1997.

[4] J Hamaker, A. Ganapathiraju, J. Picone, and J. J. Godfrey, "Advances in alphadigit recognition using syllables," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 421-424, 1998.

[5] E. Aaron, Rosenberg, L. R. Rabiner, J. G. Wilpon, and D. Kahn, "Demisyllable-based isolated word recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-31, NO. 3, pp. 713-726, 1983.

[6] T. Ji, Z. Wang, and D. Lu, "A method for chinese syllables recognition based upon sub-syllable hidden markov model," *International Symposium on Speech, Image Processing and Neural Networks*, Vol. 2, pp. 730-733, 1994.

[7] C.-H. Lin, C.-H. Wu, P.-Y. Ting, and H.-M. Wang. "Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units," *Speech Communication 18*, pp. 175-190, 1996.