연속 음성 인식을 위한 PTM 음절 모델

김봉완, 이용주
원광대학교 SiTEC, 원광대학교 전기,전자 및 정보공학부

Phonetic Tied-Mixture Syllable Model for CSR

Bong-Wan Kim, Yong-Ju Lee
SiTEC, Wonkwang Univ.
bwkim@sitec.or.kr, yjlee@wonkwang.ac.kr

요약

최근 연속 음성 인식에서의 성능 향상을 위하여 음절을 인식 단위로 사용하고자 하는 노력들이 보고되고 있다.[1][2] 그러나 음절의 경우 음소에 비해 학습성이 좋지 않고 모델의 수가 많으므로 음절 경계에서의 문맥 종속 모델링이 어렵다는 단점을 갖고 있다. 본 논문에서는 음절의 이러한 단점을 극복하기 위하여 모노폰과 트라이폰을 이용하여 음절 모델을 합성하는 방법을 제안한다. 제안된 모델은 트라이폰에 비하여 평균 55%, PTM에 비하여 평균 13%의 인식 속도 향상을 보이며, 동일한 속도일 경우 PTM, 트라이폰 모델 모두에 대하여 ERR이 약8% 향상됨을 볼 수 있었다.

1. 서론

최근 대어휘 연속 음성 인식 시스템의 경우 음소가 탈락되거나 약화되는 경우가 많이 발생함으로써 인식 시스템의 성능을 저하시킴에 따라 음소보다 긴 음절과 같은 단위들을 인식의 단위로 검토하고자 하는 노력들이 보고된 바 있다.[1][2] 또한 음절을 인식 단위로 사용할 경우 음소보다 음성 인식을 위한 어휘 트리(lexical tree) 또는 네트워크를 보다 단순하게 구성할 수 있으므로 인식 속도의 향상을 기대할 수 있다.

한국어에서 발생 가능한 음절의 수는 이론적으로 3,520개이나 실제로 쓰이는 음소의 연결에 제약이 있어 이보다 훨씬 적은 것으로 알려져 있으며 [3]에 의하면 사전에 수록된 65,973개의 표제어에 나타난 음절의 총 가짓수는 1,453개로 조사된 바 있다. 따라서, 영어의 음절에 비해 기본 단위의 수가 현저히 적어 어휘 독립 음성 인식을 위한 단위로서의 장점을 갖고 있다.

그러나 음절은 음소에 비해 학습용 데이터베이스에서 각 단위별 출현 횟수가 현저히 적어 충분한 학습이 이루어지지 않는 문제점이 있다.

본 논문에서는 이러한 음절의 단점을 극복하기 위하여 상태 공유 트라이폰(shared-state cross-word triphone)과 다중 혼합(multiple mixture)을 갖는 음소 모델로부터 Phonetic tied-mixture Syllable(PTMS) 모델을 합성하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 PTMS 모델 합성 방법, 그리고 합성된 모델의 문맥 종속 모델링에 대하여 기술한다. 3장에서는 제안된 모델을 이용하여 연속 음성 인식 실험을 수행하고 그 결과를 트라이폰, PTM 등과 인식률 및 속도의 측면에서 비교하고, 4장에서 결론을 기술한다.

2. PTMS Model

2.1 PTMS 모델의 합성

PTMS 모델을 합성하기 위하여 본 논문에서는 트라이폰과 모노폰의 형태(topology)가 동일하다고 가정한다.

제안된 PTMS 모델은 기존 음소 모델의 학습성을 공유하고, 학습 데이터에서 출현하지 않은 음절 모델의 생성 문제를 해결하기 위하여 음소 결정 트리(phonetic decision tree)를 이용하여 음절을 구성하는 음소 모델들을 결합함으로써 합성한다. 합성된 음절 모델을 구성하는 각 상태에 모노폰의 다중 혼합(multiple mixture)을 공유하도록 구성한다. 모노폰의 다중 혼합을 공유하는 방법은 PTM 모델의 합성 방법[5]과 동일하다.

그림 1에 이러한 과정을 나타내었다.

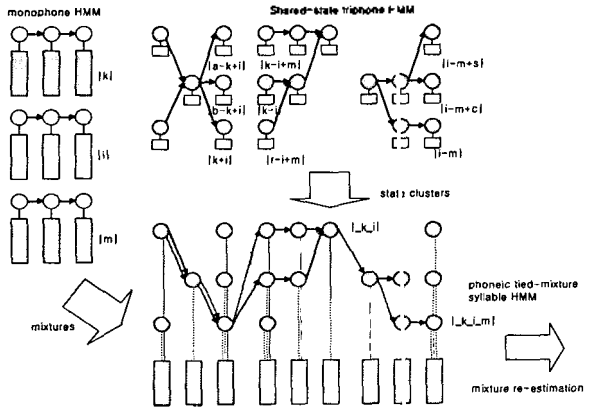

그림 1. PTMS 모델의 합성

음절 모델은 1~3개의 음소로 이루어지며, 본 논문에서는 음절을 구성하는 음소들을 표기하기 위하여 '_' 기호를 사용하였다. 즉 |_k_i| 음절은 |k| 음소와 |i|음소로 이루어져 있음을 의미한다. PTMS 모델의 합성을 위한 구체적 절차는 다음과 같다.

1. 각 상태 당 많은 가우시안 혼합(Gaussian mixture)를 갖는 모노폰을 학습한다. 파라미터 공유(Parameter tying)은 적용하지 않는다.
2. 상태 공유 트라이폰(state-shared cross-word triphone)을 학습한다. 이 모델에서의 혼합 성분은 상태의 공유를 결정하기 위해서만 사용되므로 단일 혼합(single mixture)으로도 충분하다.
3. 생성하고자 하는 음절을 구성하는 음소들을 상태 공유 트라이폰에서 찾아서 연결하여 음절 모델을 합성한다. 예를 들어 |_k_i_m|이라는 음절을 합성하기 위해서는 상태 공유 트라이폰에서 |k+i|, |k-i+m|, |i-m| 음소 모델을 찾아 연결하여 새로운 |_k_i_m|이라는 음절 모델을 생성한다. 이 때 공유되어 있는 상태는 공유 상태를 그대로 유지한다.
4. 모노폰의 혼합들을 대응하는 상태들에 할당한다. 2번단계에서 공유된 상태(tied state)들은 혼합 성분과 가중치(mixture weight)를 모두 공유하며, 공유되지 않은 상태(non-tied state)는 혼합은 혼합 성분만을 공유하며 가중치는 공유하지 않는다.
5. 혼합 성분과 가중치를 재추정하기 위하여 훈련 과정을 다시 수행한다.

2.2 자주 출현하는 음절 모델의 상태 공유 해제

위에서 합성된 음절 모델들은 트라이폰의 공유 상태 정보와 모노폰의 다중 혼합을 공유함으로써 학습 데이터에서 출현하지 않은 음절 모델을 생성할 수 있는 방법과 음절의 낮은 학습성을 보완하여 준다. 그러나 학습 데이터에서 자주 출현하는 음절 모델들은 그 음향적 특성에 대한 데이터가 충분하므로 이를 별도로 모델링하는 것이 가능하다. 따라서 본 논문에서는 학습데이터에서 자주 출현하는 음절 모델들은 공유된 상태를 해제하여 훈련 과정을 다시 수행함으로써 혼합 가중치를 별도로 모델링하도록 하였다.

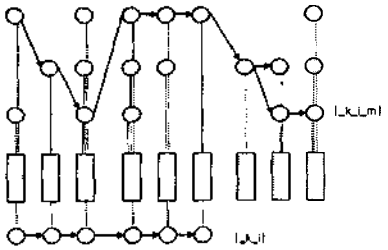

그림 2. 자주 출현하는 음절 모델의 상태 공유 해제

위의 그림에서 |_k_i| 음절 모델은 기존 공유된 상태를 해제하여 별도의 상태에 자신만의 혼합 가중치를 갖게되며 이후의 학습과정을 통하여 혼합 가중치를 재추정한다.

2.3 문맥 종속 모델링

대어휘 음성 인식을 위해서 대부분의 인식기에서는 문맥 종속 모델링을 사용하고 있다. 그러나 음소의 경우 기본 단위의 수가 적으므로 이러한 문맥 종속 모델링이 어렵지 않게 구현될 수 있으나, 한국어 음절의 경우 약 1,500여개에 달하므로 문맥 종속 모델링을 수행하기 어렵다. 따라서, 이러한 문제를 해결하기 위하여 앞 또는 뒤의 음절 전체를 문맥의 구성요소로 사용하는 것이 아니라 앞 음절의 경우 음절을 구성하고 있는 가장 나중 음소를 문맥 구성 요소로 사용하고, 뒷 음절의 경우 음절을 구성하고 있는 가장 첫 음소를 문맥 구성 요소로 사용하는 음소 문맥 종속 음절 모델링 방법이 제안된 바 있다.[4]

본 논문에서는 음소 종속 음절 모델링을 위하여 어휘 내에서 좌측 음소만을 문맥으로 사용하여 음절 모델(Word-internal left-phone dependent PTMS 모델)을 구성한다. 이렇게 모델링 할 경우 약 40,000 ~ 60,000 여개의 모델만으로 문맥종속 모델링이 가능하다.

그러나 학습 과정에 자주 출현하지 않은 음절의 경우 문맥 종속 모델을 안정되게 학습시킬 데이터의 양도 충분치 않거니와, 공유된 상태를 통해서 다중 혼합 및 가

중치가 완전히 공유되고 있으므로 문맥 종속 모델과 문맥 비종속 모델의 구분이 무의미하다.

따라서 본 논문에서는 문맥 종속 음절 모델의 경우 2.2절에서 기술한 상태 공유가 해제된 음절에 대해서만 문맥 종속 음절 모델링이 수행되며 그렇지 않은 모델에 대해서는 문맥 종속 모델을 생성하지 않고 단순히 문맥 비종속 모델을 공유하여 사용한다.

3. 인식 실험 및 고찰

3.1 음성 데이터 베이스 및 인식 시스템

인식 실험을 위하여 400명의 화자가 1인당 평균 104문장을 발성한 Dictation 용 음성 DB를 사용하였다.[6] 발성된 총 문장은 20,833문장으로 대규모 텍스트 코퍼스로부터 고빈도 10,000어절로 구성된 문장을 추출하여 수집된 DB이다. 학습을 위하여 총360명(남성 180명, 여성 180명)의 화자의 데이터를 사용하였다. 테스트를 위해 사용된 데이터의 화자의 양은 40명(남성 20명, 여성 20명)으로 총 472분 분량이다.

음성 DB는 발성 내용에 대하여 철자 및 발음 전사가 이루어져 있으며 본 실험에서는 동일 어휘라 할지라도 다르게 발성된 어휘는 모두 다른 어휘로 취급하여 인식 실험을 수행하였으며, 인식 사전에 포함된 어휘는 13,118개이다.

음성 인식을 위한 사전의 구성은 발음 전사를 이용하여 음소 인식 사전과 음절 인식 사전을 구성하였으며 어휘의 수, 음절을 구성하는 음소의 수와 그 순서는 동일하게 구성하였다. 실험을 위한 인식 시스템은 two-pass decoder 방법을 사용하는 Julius을 이용하였으며 2-gram, 3-gram 언어 모델을 적용하여 인식 성능을 측정하였다.[7]

음성 DB 전체에서 출현한 음절 모델의 개수는 총 1,125개 이며 이 중 학습 데이터에서 출현하지 않은 음절 모델의 개수는 10개이다. 1,125개의 음절 중 360명분의 학습 데이터(총 37,506 문장)에서 2,000이상 출현한 음절은 75개에 불과하며 100회 미만 출현한 음절의 개수도 661개에 달한다. 전체 음절당 평균 출현 횟수는 604.1로 음소에 비해 학습성이 현저히 떨어진다는 것을 알 수 있다. 2.2절과 2.3절에서 기술한 공유 상태 해제와 문맥 종속 모델링은 학습 데이터에서 2,000회 이상 출현한 음절 75개에 대해서만 적용하였다.

음성 인식을 위하여 사용된 특징 파라미터는 12차 MFCC, 12차 차분 MFCC, 그리고 정규화된 대수 에너지 성분을 포함하여 총 25차의 특징 파라미터를 사용하였다. 이후 성능 비교를 위해 기술되는 모든 항목에서 약자들의 의미는 다음과 같다.

표1. 성능 비교를 위한 모델 들

단위	모델	설명
음소	CTM16	cross-word triphone, 16 mixture
	WTM16	word-internal triphone, 16 mixture
	PTM(CT)	cross-word triphone based PTM
	PTM(WT)	word-internal triphone based PTM
음절	PTMS	제안된 음절 모델
	PTMS(UT)	제안된 음절 모델 + 공유 상태 해제
	PTMS(LC)	제안된 음절 모델 + 공유 상태 해제 + word internal left phone dependent modeling

3.2 음소 모델들의 성능

트라이폰 모델들은 모두 상태당 16개의 혼합을 갖도록 구성되어 있으며 PTM 모델들은 모두 64개의 혼합을 갖는 모노폰으로부터 트라이폰을 이용하여 생성된 것이다. 크로스 워드 트라이폰의 경우 인식률이 PTM에비해 다소 우월하지만 인식 속도에 있어서는 현격한 차이를 보임을 알 수 있다. PTM은 크로스 워드 트라이폰에 비해 다소간의 인식률저하가 있기는 하지만 거의 실시간으로 인식이 이루어짐을 볼 수 있다.

표2. 음소 모델과 음절 모델의 비교

인식 단위	Model	# models	# states	# mixtures
음소	CTM16	12,089	4,943	79,088
	WTM16	4,801	4,106	65,695
	PTM(CT)	12,089	4,943	8,640
	PTM(WT)	4,801	4,107	8,640
음절	PTMS	1,127	1,583	8,640
	PTMS(UT)	1,127	2,000	8,640
	PTMS(LC)	2,124	8,114	8,640

표3. 음소 모델들의 인식 성능

Models	Beam Width	Word accuracy		X RT
		2-gram	3-gram	
CTM16	1500	87.87	93.71	2.42
	800	86.54	92.05	1.43
WTM16	1500	82.86	89.51	2.26
	800	79.74	86.05	1.39
PTM(CT)	1500	87.52	93.04	1.08
	800	86.26	91.38	0.82
PTM(WT)	1500	84.65	91.04	0.99

3.3 제안된 PTMS 모델의 성능

PTMS 모델은 단일 혼합을 갖는 상태 공유 트라이 폰(shared-state cross-word triphone)과 64개의 혼합을 갖는 모노폰으로부터 합성한후 훈련 과정을 거친 것이다.

표4. 제안된 음절 모델의 성능

Models	Beam width	Word acuracy		X RT
		2-gram	3-gram	
PTMS	1500	86.16	91.34	0.88
	800	84.38	89.03	0.72
PTMS(UT)	1500	86.67	92.02	0.90
	800	85.08	89.7	0.72
PTMS(LC)	1500	87.83	92.71	0.90
	800	86.16	90.59	0.75

표 4를 보면 제안된 음절 모델이 음소 모델에 비하여 전체적으로 인식 속도가 향상되었음을 볼 수 있다. 트라이폰과 비교하여 50%이상의 속도 향상을 볼 수 있고 PTM(CT)와 비교해 보아도 13%정도 속도가 향상되었음을 알 수 있다. PTMS(LC)의 경우 어휘 내부에서만 문맥 종속 모델링을 수행하고 있는 WTM16, PTM(WT)에 비하여 ERR에 있어 18% 이상의 향상이 있음을 볼 수 있으며, CTM16의 빔폭 800과 PTMS(LC)의 빔폭 1500을 비교해 보면 PTMS(LC)가 인식 속도, 인식률 모두에 있어서 향상되었음을 알 수 있다.

표5. PTM과 PTMS(LC)의 인식 성능 비교

Beam width	Model	Word accuracy		X RT
		2gram	3gram	
2000	PTM(CT)	87.43	93.46	1.22
	PTMS(LC)	88.04	93.05	1.04
1500	PTM(CT)	87.52	93.04	1.08
	PTMS(LC)	87.83	92.71	0.90
1000	PTM(CT)	86.80	92.08	0.90
	PTMS(LC)	86.93	91.60	0.78
800	PTM(CT)	86.26	91.38	0.82
	PTMS(LC)	86.16	90.59	0.75

Phonetic tied mixture 기법을 유사하게 사용하는 PTM(CT)와 PTMS(LC)를 비교한 표 5를 보면 빔폭 1,000인 PTMS(LC) 모델이 빔폭 800인 PTM(CT)보다 속도가 빠르면서도 인식률은 오히려 더 향상되었음을 볼 수 있으며, 이러한 양상은 다른 빔폭에서도 지속됨을 볼 수 있다. 동일한 속도를 나타낸 경우 8%정도의 ERR 향상이 있음을 알 수 있다.

4. 결론

본 논문에서는 한국어 음성 인식을 위한 인식 단위로서 음절을 사용하기 위하여 다중 혼합을 갖는 모노폰과 상태 공유 트라이폰으로부터 음절 모델을 합성하는 방법을 제안하였다. 제안된 모델은 트라이폰에 비하여 평균 55%, PTM(CT)에 비하여 평균 13%의 인식 속도 향상을 보이며, 동일한 속도일 경우 PTM(CT), CTM16 모두에 대하여 ERR이 8%정도 향상됨을 볼 수 있었다.

음절의 경우 음소에 비해 보다 긴 음성 구간을 모델링하며, 제안된 음절 모델의 경우 음소보다 상태의 수가 많으므로 상태들 간 다중 경로를 통하여 음절 내의 발음 변이를 모델링하기 쉬운 장점을 갖고 있으므로 이에 대한 추가 연구가 필요하다. 또한 제안된 음절 모델의 어휘간 문맥 종속 모델링이 가능하도록 음성 인식기의 문맥 확장(context expansion) 방법을 개선하는 방안에 대한 연구도 필요하다.

참고문헌

1. Aravind Ganapathiraju, Joseph Picone, et al., "Syllable-Based Large Vocabulary Continuous Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 9, (4), 358-366, 2001

2. H. Bourlard, H. Hermansky, N. Morgan, "Copernicus and the ASR challenge - Waiting for Kepler," Proc. DARPA Speech Recognition Workshop, 157-162, 1996

3. KBS, 표준한국어발음대사전, 어문각 1993

4. 김봉완, 이용주, "연속 은닉 마코프 모델을 이용한 한국어 음성 인식을 위한 효율적 음절 모델링," 한국음향학회지, 22권 1호, 23-27, 2003

5. Akinobu Lee, Tatsuya Kawahara, et al., "A New Phonetic Tied-Mixture Model For Efficent Decoding," Proc. ICASSP '00, 1269-1272, 2000

6. 김봉완, 최대림, 김영일, 이광현, 이용주, "SiTEC의 공동 이용을 위한 음성 코퍼스의 구축 현황 및 계획," 대한음성학회 말소리, 제46호, 175-185, 2003

7. T. Kawahara, A. Lee, T. Kobayashi, et al., "Free software toolkit for Japanese large vocabulary continuous speech recognition," Proc. ICSLP 2000, Vol. 4, 476-479, 2000