자동 음소 분할 성능 개선을 위한 음소 모델링에 관한 연구

박혜영, 김형순
부산대학교 전자공학과

A Study of Phoneme Modeling for Improvement of Automatic Segmentation Performance

Hae Young Park, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
E-mail : {phyoe, kimhs}@pusan.ac.kr

요 약

본 논문에서는 Hidden Markov Model(HMM)을 이용하여 corpus 기반 TTS에 사용할 DB를 자동 음소 분할 해주는 시스템을 구현하였다. HMM을 이용해서 음소 분할할 경우 HMM을 모델링 하는 방법에 따라 많은 성능의 차이가 난다. 따라서 본 논문에서는 HMM 모델링 방법에 따른 몇 가지 실험 및 성능 평가를 하였다. 실험 결과 음성 인식과는 달리 HMM모델링 시 triphone 모델보다 monophone 모델의 성능이 더 우수하였으며, 에너지 기반의 후처리를 통해 성능 향상을 얻을 수 있었다.

1. 서 론

최근의 음성 합성 기술은 컴퓨터의 성능 향상과 저장 장치에 대한 가격의 하락으로 인해 기존의 conventional TTS에서 보다 더 자연스러운 합성음을 만들어 낼 수 있는 corpus 기반의 TTS로 전환이 이루어지고 있다. 이에 따라, 최근 TTS는 대용량 음성 데이터 베이스에 대해 음소 분할된 corpus 구축을 필요로 하고 있다. 그러나 이러한 음소 분할 작업을 사람이 직접 수행할 경우 몇 가지 문제점이 있다[1]. 첫째로 이 과정은 스펙트로그램 판독 및 반복되는 듣기평가를 통해 이루어지므로 매우 지루한 작업일 뿐만 아니라 많은 시간이 소요되게 된다. 둘째로 수작업에 의한 음소 분할은 높은 수준의 음성학적 지식을 요하며, 소수의 음성학 전문가에 의존할 수 밖에 없다. 셋째로 음소 경계 선정을 위한 구체적인 판단기준을 미리 정해놓더라도 상당 부분의 경우 주관적인 판단을 전혀 피할 수 없으며, 이에 따라 음소 경계 선정과정에서의 일관성이 보장되지 못한다[2].

음소 분할 작업을 자동으로 할 수 있다면 위에서 언급한 문제들이 해소 될 수 있으며, 대용량 corpus 기반의 TTS를 구축하는데 있어 수작업에 의한 업무량과 시간을 단축할 수 있다. 이에 따라 HMM 기반의 자동 음소 분할 시스템이 corpus 기반 TTS에 많이 사용되고 있으며, 본 논문에서는 HMM기반의 자동 음소 분할 시스템의 성능을 개선하기 위해 HMM 모델링에 관한 몇 가지 실험을 하였다.

본 논문의 구성은 다음과 같다. 2절에서 baseline 시스템 구성을 위한 고려 사항들에 대해 살펴 보고, 3절에서는 본 논문에서 자동 음소 분할 성능을 개선하기 위해 사용한 여러 가지 모델링 방법 및 실험에 대해서 기술한다. 4절에서는 각각의 모델링 방법에 대한 실험 결과를 비교분석하고, 5장에서 결론을 맺는다.

2. Baseline 시스템 구성

Corpus 기반 TTS의 음질 향상을 위해서는 정확한 음소 경계 정보를 얻는 것이 중요하다. 이를 위해 HMM을 모델링 할 경우 음소 분할 단위, 모델 topology, 음성 특징 파라미터 등 고려해야 할 사항이 많다. 본 논문에서 자동 음소 분할을 위한 baseline 시스템을 구성하기 위해 고려한 사항은 다음과 같다.

2.1 음소 분할 단위 선정

우리말의 음소의 수는 자음 19개, 모음 21개이다. 이 중에 모음의 경우 현대 우리말에서 발음 구분이 불분명해지고 있는 "ㅔ"와 "ㅐ", "ㅖ"와 "ㅒ", 그리고 "ㅚ"와 "ㅙ"와 "ㅞ"는 각각 동일한 음소인 것으로 간주하였다. 그 결과 모음 중에서 음소 분할 단위로 선정된 것은 17개이다. 그리고 자음에 있어서 여러 가지 음운 현상에 의해 변이음들이 나타날 수 있는데 본 논문에서는 폐쇄

음의 불파음화에 대한 변이음(g', d', b')과 유음에 있어 'r'(탄설음)의 'l'(설측음)되기에 대한 변이음을 따로 고려했다. 따라서 최종 음소 분할 단위로 선정된 것은 묵음(silence) 및 짧은 휴지 구간(short pause)에 대해 2개, 자음 23개, 모음 17개로 총 42개이다.

2.2 음성 특징 분석 파라미터 선정

일반적으로 음성 인식에서는 매 10ms마다 20ms 구간의 음성신호로부터 음성특징 파라미터를 추출하는 방식이 널리 사용되고 있다. 그러나 정교한 음소 분할 및 레이블링을 위해서는 보다 미세한 음성분석 시간단위가 필요하다. 참고로, TIMIT 음성 데이터베이스 구축시 사용된 자동 음소 분할에서는 2.5ms의 시간단위가 사용된 것으로 알려지고 있으며[3], 시간단위가 5ms를 넘지 않도록 설정하는 것이 좋을 것으로 판단된다. 본 논문에서는 5ms마다 20ms구간의 음성 신호로부터 12차 Mel Frequency Cepstrum Coefficient(MFCC)와 log energy, 그리고 이들의 delta 계수 및 acceleration 계수들로 총 39차 특징벡터를 추출하였다.

2.3 음소 모델 구성

HMM에 의해 모델을 구성하기 위해서는 관찰확률 분포를 이산분포, 연속분포 또는 준연속분포 중에서 선정해야 하며, 상태수와 천이방식 등 HMM topology와 더불어 일부 파라미터의 tying 여부를 정해야 한다. 이러한 사항의 결정은 음소 분할 실험을 통한 성능평가에 따라 이루어져야 할 것이다.

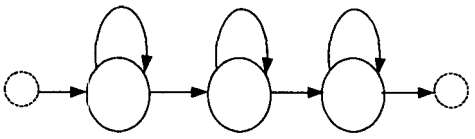
본 논문의 baseline 시스템에서 HMM은 연속분포이며 상태수는 3개로 구성하였으며 천이방식에 따라 그림 1에 나타나 있는 두 가지 형태로 구성하였다. 하나는 일반적인 음소 모델을 위해 skip path 없이 self transition과 단지 이웃하는 state로 transition하는 형태로 모델링 한 경우이고, 또 다른 하나는 짧은 휴지 구간(short pause)에 주위 잡음에 의해 나타나는 짧은 burst나 화자의 발성 특성에 따라 나타나는 숨소리 등 다양하게 나타나는 잡음 성분을 모델링 하기 위해 skip path가 있고 transition이 음소 모델 보다 복잡한 형태로 모델링 한 경우이다.

2.4 발음 사전 구성

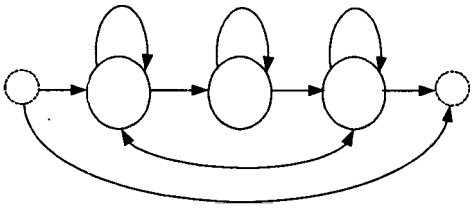
동일한 단어라도 다양한 형태로 발음될 수 있기 때문에 경우에 따라서는 복수 개의 발음사전을 사용할 수 있다. 하지만 모든 가능한 발음을 정확하게 표시할 수 없다. 따라서 본 논문에서는 하나의 단어에 대해 하나의 발음 사전을 고려하였으며, 단어 뒤에 휴지구간이 오는 경우와 오지 않는 경우 두 가지 형태로 구성하였다.

3. HMM 모델링 방법에 대한 실험

본 논문에서는 좌우 context를 고려한 triphone과 고려하지 않은 monophone에 대한 실험과 Baun-Welch 알고리즘으로 모델링 한 다음 음소 경계 정보를 얻어서 다시 segmental K-means 알고리즘으로 모델링 한 경우의 성능을 살펴 보았다.



(a) 일반적인 변이음에 대한 HMM 구성



(b) Short pause를 위한 HMM 구성

그림 1. 음소 모델을 위한 HMM 구성

3.1 HMM 단위에 관한 실험

음성 인식에서는 동일한 음소일지라도 좌우 cocntext에 따라 특성이 다르기 때문에 좌우 context를 고려한 triphone의 성능이 monophone의 성능보다 일반적으로 우수하다. 그러나 [3]에서는 음성 인식 성능과 음소 분할 성능 사이에는 상관관계가 없는 것으로 나타나있다. 실제 음소 분할을 위한 HMM의 모델링 단위 및 mixture 개수는 실험을 통한 성능평가에 따라 선정할 필요가 있다.

본 논문에서는 triphone의 tied state 수에 따른 음소 분할 성능 분석과 더불어 triphone과 monophone의 성능을 비교하였으며, 각각의 경우에 대해 mixture 수를 변화시키면서 실험하였다.

Triphone으로 모델링 할 경우 좌우 문맥을 고려하기 때문에 모델의 수가 엄청나게 많아지게 되며, 실제 주어진 데이터보다 추정해야 할 파라미터 수가 너무 많아지게 된다. 이 문제의 해결을 위해 state를 tying 하는 방법이 사용되며, 여기에는 tree based clustering(TBC)과 data driven clustering(DDC)의 두 가지가 있다. 본 논문에서는 이 중에서 TBC를 사용하였다. TBC의 경우 tied state의 개수를 제한하기 위해 두 가지 threshold가 사용되는데 하나는 각 node에 할당된 관찰벡터의 최소 개수이며 또 다른 하나는 clustering 전후의 likelihood의 변화값인 delta likelihood(DL)이다. 본 실험에서는 최소 관찰벡터의 수는 고정시킨 상태에서 DL값을 변화시킴으로써 tied state수를 조절하였다. 또한, 모델을 상세하게 표현하기 위해 mixture 수에 따른 실험도 하였다.

한편, 음소 분할은 음성 인식과 달리 하나의 음소에 대해 좌우에 오는 음소의 영향을 고려해서 모델링 하는 것보다는 고려하지 않고 모델링 하는 것이 음소 분할 성능을 개선시킬 수 있을 것이다. 이에 따라 monophone

모델에 대해서도 mixture 수를 늘려 가면서 실험하였다.

3.2 2단계 모델링에 의한 실험

2단계 모델링에 의한 방법은 1단계 모델링 즉, Baum-Welch reestimation 모델링(BW)에서 얻어진 음소 경계정보를 이용해서 segmental K-means reestimation으로 다시 HMM을 모델링 하는 방법이다. Segmental K-means으로 모델링 할 경우 초기모델로 두 가지를 사용했는데, 하나는 Baum-Welch reestimation 모델링 마지막 단계에서 만들어진 모델을 사용하는 경우(method 1)이고 또 다른 하나는 전체 DB에 대해 평균과 분산을 모델 파라미터로 초기화 하는 경우(method 2)이다. 2단계 모델링의 전체 구성도는 그림 2와 같다.

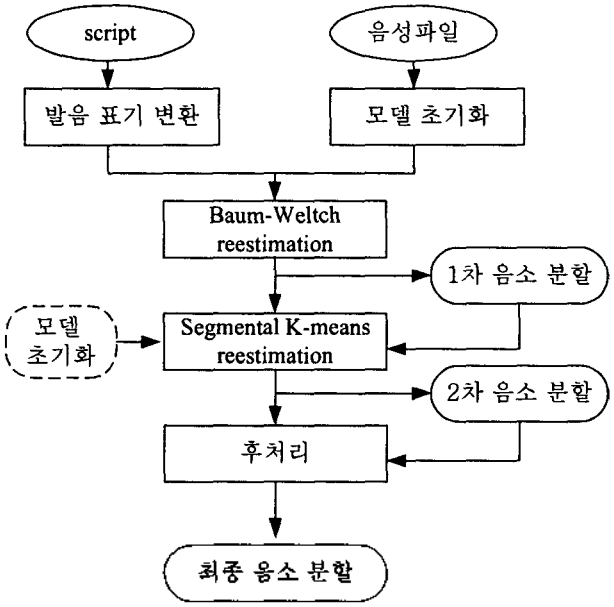


그림 2. 2단계 모델링에 의한 음소 분할 시스템의 구성도

4. 실험 결과 및 분석

음소 분할 시스템을 구현하고 그 성능을 향상시키기 위해서는 객관적인 성능평가가 필요하다. 본 논문에서는 corpus 기반의 TTS를 만들기 위해 수집된 5,175문장의 여성 데이터 베이스 중에서 50문장을 수작업으로 음소 분할 한 다음 자동 음소 분할 결과와 비교하여 성능평가를 하였다. 성능평가 방법은 수작업에 의한 음소 분할 정보와 비교해서 50문장에 있는 전체 음소 개수에 대해 오차 범위가 10ms 또는 20ms 이내에 들어오는 음소의 비율로 평가하였다.

4.1 수작업에 의한 음소 분할

수작업에 의한 음소 분할 작업은 POW(phonetically optimized word) 여성 DB에 대해 Baum-Welch reestimation 알고리즘 과정을 거쳐 나온 화자 독립 HMM을 이용해 Viterbi segmentation을 통해 얻은 음소 경계 정보를 수정하였다. 수작업에 의한 음소 분할 작업은 전문 레이블링 교육을 이수한 두 명의 전자공학과 학생이 하였다. 음소 분할 기준은 SERI와 ETRI가 지원한 연구과제를 수행하는 과정에서 원광대에서 1996년부터 개발하고 사용해 왔던 기준안을 참고로 하였다.

4.2 실험 결과

표 1은 HMM을 triphone으로 모델링 할 때 DL값을 변화시키면서 tied state 수를 조절할 때의 음소 분할 성능을 나타낸 것이다. 표의 제일 처음에 있는 DL=186.7은 POW 남성 DB에 대해서 여러 가지 방법으로 실험한 후 인식 성능이 가장 우수했을 때의 DL값이다. 음성 인식의 경우 tied state수를 조절함에 따라 성능이 상당히 달라지지만, 표 1의 결과를 보면 tied state수에 따른 음소 분할 성능의 변동은 거의 없었다.

표 1. Triphone 모델에서 DL값에 따른 음소 분할 성능 (총 state 수 : 41982)

DL		186.7	200	400	800	1000
Tied state 수		12301	11867	7513	4511	3859
오차범위	≤ 10ms	54.2	54.0	53.6	53.7	53.8
	≤ 20ms	82.1	82.2	82.2	81.9	82.3

표 2는 triphone과 monphone 모델에 대해 mixture 수를 1개에서 5개까지 증가시키면서 한 실험의 결과이다. 이 때 triphone은 표 1에서 성능이 우수한 경우인 DL이 186.7인 경우에 대해서 실험을 하였다. Mixture 수가 동일하게 1개일 경우 monophone의 성능이 triphone의 성능 보다 우수하며, triphone의 mixture 수를 늘리더라도 monophone의 음소 분할 성능이 우수하다.

표 2. Mixture수에 따른 성능

(a) Triphone

Mixture 수		1	2	3	4	5
오차범위	≤ 10ms	54.2	51.5	52.9	52.2	50.9
	≤ 20ms	82.1	79.0	80.2	78.8	78.0

(b) Monophone

Mixture 수		1	2	3	4	5
오차범위	≤ 10ms	59.6	62.9	63.3	61.2	60.1
	≤ 20ms	83.6	84.1	84.0	83.5	82.8

표 3에서 (a)는 2단계 모델링 방법에서 2차 모델링시 초기 모델에 따른 성능을 나타내었고, (b)는 2단계 모델링에서 얻어진 음소 경계 정보를 에너지 기반 후처리를 한 결과를 나타내었다. 자동 음소 분할 경계에서 묵음과 어떤 음소가 오는 경우 그 경계가 묵음 쪽으로 치우치는 경향이 있었다. 이 경우에 자동 음소 분할의 경계를 묵음과 음성의 경계쪽으로 옮겨 주기 위해 에너지를

이용한 후처리를 도입하였다. 에너지에 의한 후처리는 자동 음소 분할에서 묵음 구간이 검출된 부분에서만 하게 되는데, 묵음 앞에 오는 음소가 모음이나 비음일 경우 final lengthing으로 인해 묵음 구간의 에너지와 차이가 크지 않기 때문에 이 부분에서는 후처리를 하지 않았다.

표 3을 보면 1차 모델링에 의해 음소 분할한 경우(BW)보다 1차 음소 경계 정보를 이용해서 segmental K-means로 2차 모델링 한 경우(method 1, method 2)의 성능이 더 우수하다. 그리고 전체 DB에 대한 평균과 분산을 초기 모델로 사용한 경우(method 2)가 1차 모델링에서 만들어진 마지막 모델을 사용하는 경우(method 1) 보다 성능이 더 우수한데, 이유는 이 경우에 사용한 초기 모델이 local maximum에 빠졌기 때문이라고 추정된다. 또한 에너지 기반의 후처리를 한 경우 모델링 방법과 관계없이 전체적으로 성능이 개선되었다.

표 3. Monophone 모델에서 모델링 방법에 따른 성능

(a) 후처리를 하지 않은 경우

Modeling		BW	Method 1	Method 2
오차범위	≤ 10ms	59.6	61.7	69.1
	≤ 20ms	83.6	84.2	82.8

(b) 후처리를 한 경우

Modeling		BW	Method 1	Method 2
오차범위	≤ 10ms	62.4	64.2	71.3
	≤ 20ms	85.3	86.3	84.2

표 4는 표 3(a)에서 가장 좋은 성능을 나타낸 모델(Method 2)에 대해 후처리 후 음소그룹 pair별 자동 음소 분할 성능을 나타낸 것이다. 여기에서 파열음, 불파음, 마찰음, 파찰음, 유음, 이중모음, 단모음의 8가지 음소그룹을 선정하였다. 표에서 보면 성능의 저하가 크게 일어나는 경우를 두 가지로 나누어 볼 수 있다. 첫번째는 불파음 다음에 파열음, 마찰음, 파찰음이 오는 경우인데, 이 경우에 불파음의 폐쇄 구간과 파열음, 마찰음, 파찰음의 폐쇄구간이 중첩이 되어 구별되지 않기 때문에 수작업으로 음소 분할 시 두 음소 사이의 묵음 구간의 반을 경계로 정했기 때문이다. 두 번째는 비음과 비음, 유음과 유음, 모음과 모음 혹은 모음과 유음이 연속해서 오는 경우이다. 이 경우는 두 음소 사이의 경계를 사람의 눈과 귀로도 구별하기 어렵다. 즉, 수동 음소 분할이 어려운 부분은 자동 음소 분할도 어렵다는 것을 나타낸다.

한편, 모음의 경우 자음에 비해 평균적인 음소 지속 시간이 길고, 특정 모음의 경우 무성음-화가 일어나 그 음소의 특성이 나타나지 않는 경우도 있다. 이러한 이유로 인해 모음에 대한 HMM의 topology는 자음과 다른 형태로 구성하는 것에 대해 검토할 필요가 있다.

표 4. 2단계 모델링에 대한 음소그룹 pair 별 결과
(오차범위 10ms 이내인 경우)

후행 / 선행	파열	불파	마찰	파찰	비음	유음	이중모음	단모음
파열	-	-	-	-	-	-	97.4	91.9
불파	12.5	-	37.5	0.0	0.0	-	-	100
마찰	-	-	-	-	-	-	83.3	85.6
파찰	-	-	-	-	-	-	100	93.1
비음	70.0	-	76.5	45.5	3.2	50.0	83.9	91.6
유음	7.1	-	75.0	16.7	66.7	7.1	71.4	54.7
이중모음	74.1	57.1	50.0	50.0	71.0	31.5	7.1	20.0
단모음	66.5	73.0	73.1	67.9	76.8	38.7	20.8	39.6

5. 결 론

본 논문에서는 corpus 기반 합성기를 구축하기 위한 준비 작업으로 합성 단위를 자동으로 분할해주는 방법에 대해서 논의하였다. 이를 구현하기 위해 HMM 모델링 방법에 따른 여러 가지 실험을 하였다. 음소 분할에 있어서 음성 인식과는 달리 triphone의 tied state 개수에 따른 영향은 미비했으며, triphone보다는 monophone 모델 이용시 더 좋은 성능을 나타내었다. 본 논문에서 구현한 자동 음소 분할의 성능은 후처리를 하지 않은 경우 오차범위가 10ms인 경우에 69.1%이며, 오차범위가 20ms인 경우에 82.8%의 성능을 보였으며 에너지 기반의 후처리를 한 경우 각각의 오차 범위에 대해 71.3% 및 84.2%의 성능을 얻었다.

향후 연구과제로 음소 그룹별로 HMM topology를 음소 그룹별로 다르게 구성하는 방안과 음소그룹 pair 특성에 따른 후처리 방법에 대해 연구가 진행될 예정이다.

참 고 문 헌

[1] H. C. Leung and V. Zue, "A procedure for automatic alignment of phonetic transcription with continuous speech," in Proc. ICASSP, pp.429-432, Apr. 1984.

[2] B. Eisen, H. Tillmann and C. Draxler, "Consistency of judgments in manual labeling of phonetic segments : the distribution between clear and unclear cases," in Proc. ICSLP, pp.871-874, Oct. 1992.

[3] A. Ljolie and M. D. Riley, "Automatic segmentaion and labeling of speech," in Proc. ICASSP, pp.473-476. Apr. 1991.

[4] P. Caralho, I. Trancoso and L. Oliveira, "Automatic segment alignment for concatenative speech synthesis in Portuguese," 10th Portguese Conference on Pattern Recognition, RECPAD, pp.221-226, Feb. 1998.

[5] 홍성태, 김제우, 김형순, "자동 음성분할 및 레이블링 시스템의 성능향상," 대한음성학회지, 제35-36권, pp. 175-189, 1998년 12월.