자동 음성 분할을 위한 음향 모델링 및 에너지 기반 후처리

박혜영(부산대), 김형순(부산대)

<차 례>

<Abstract>

Acoustic Modeling and Energy-Based Postprocessing for Automatic Speech Segmentation

Hyeyoung Park, Hyungsoon Kim

Speech segmentation at phoneme level is important for corpus-based text-to-speech synthesis. In this paper, we examine acoustic modeling methods to improve the performance of automatic speech segmentation system based on Hidden Markov Model (HMM). We compare monophone and triphone models, and evaluate several model training approaches. In addition, we employ an energy-based postprocessing scheme to make correction of frequent boundary location errors between silence and speech sounds. Experimental results show that our system provides 71.3% and 84.2% correct boundary locations given tolerance of 10 ms and 20 ms, respectively.

* 주제어: 자동음성분할(automatic speech segmentation), 음향 모델링(acoustic modeling), 후처리(postprocessing), 음성합성(speech synthesis)

1. 서 론

연속음성을 음소 단위로 분할하고 레이블링하는 일은 음성학 및 음성공학의 여러 연구 분야에서 매우 유용한 일이다. 특히, 최근의 음성 합성 기술은 기존의 conventional TTS에서 보다 자연스러운 합성음을 생성시키는 코퍼스(corpus) 기반의 TTS로 전환이 이루어지고 있으며, 이를 위해서는 대용량 음성 데이터 베이스에 대해 음소 단위로 분할된 코퍼스의 구축이 필수적으로 요구된다.

이러한 음소 분할 작업을 사람이 직접 수행할 수도 있으나, 여기에는 많은 문제가 수반된다[1]. 우선 이 과정은 스펙트로그램 판독 및 반복되는 듣기평가를 통해 이루어지므로 매우 지루한 작업일 뿐만 아니라 많은 시간이 소요되게 된다[2]. 또한 수작업에 의한 음소 분할은 높은 수준의 음성학적 지식을 요하며, 음소 경계 선정을 위한 구체적인 판단기준을 미리 정해놓더라도 상당 부분의 경우 주관적인 판단을 전혀 피할 수 없어서, 음소경계 선정과정에서의 일관성이 보장되지 못한다[3].

음소 분할 작업을 자동으로 할 수 있다면 위에서 언급한 문제들이 해소 될 수 있으며, 대용량 코퍼스 기반의 TTS를 구축하는데 있어 수작업에 의한 업무량과 시간을 단축할 수 있다. 이에 따라 음성인식에 널리 사용되고 있는 Hidden Markov Model(HMM) 기반의 자동 음성 분할 시스템에 대한 연구가 상당히 이루어지고 있으며, 이 기술이 코퍼스 기반의 TTS에 실제로 사용되고 있다[4]-[7]. 본 논문에서는 HMM기반의 자동 음성 분할 시스템의 성능을 개선하기 위해 HMM 모델링 방법 및 후처리에 관한 몇 가지 실험을 하였다.

본 논문의 구성은 다음과 같다. 2절에서 baseline 시스템 구성을 위한 고려 사항들에 대해 살펴보고, 3절에서는 본 논문에서 자동 음성 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

코퍼스 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. 음성 특징 분석 파라미터 선정

일반적으로 음성 인식에서는 매 10ms마다 20ms 구간의 음성신호로부터 음성특

징 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[4], 시간단위가 5ms를 넘지 않도록 설정하는 것이 좋을 것으로 판단된다. 음성인식을 위한 음성특징분석 파라미터는 음소변별력이 뛰어나면서 음성학적으로 중요하지 않은 변화요인에 둔감한 특성을 가지는 것이 요구된다. 지금까지 음성인식 효과적으로 사용되어 온 음성특징 파라미터로는 LPC cepstrum 또는 Mel-frequency cepstrum과 이들의 시간 축 미분 값들을 들 수 있으며, 단구간 에너지와 그 미분치도 중요한 정보로 활용된다. 본 논문에서는 5 ms마다 20 ms구간의 음성 신호로부터 12차 Mel Frequency Cepstrum Coefficient(MFCC)와 log energy, 그리고 이들의 delta 계수 및 acceleration 계수들로 총 39차 특징 벡터를 추출하였다.

2.2. 음소 분할 단위 선정

음성을 분할하기 위한 기본단위로는 음소, 유사음소(phoneme-like unit), 변이음 등이 사용될 수 있으며, 이는 코퍼스 기반의 TTS에 사용되는 음성 합성단위가 어떤 것인가와 밀접한 관련이 있다.

우리말의 음소는 자음의 경우 15개의 장애음(ㅂ, ㅃ, ㅍ, ㄷ, ㄸ, ㅌ, ㅅ, ㅆ, ㅈ, ㅉ, ㅊ, ㄱ, ㄲ, ㅋ, ㅎ)과 3개의 비음(ㅁ, ㄴ, ㅇ), 그리고 1개의 유음(ㄹ)을 합쳐 총 19개로 구성되며, 모음의 경우 단모음과 이중모음을 합쳐서 21개이다. 본 연구에서는 이들 음소들 중에서 현대 우리말에서 발음 구분이 불분명해지고 있는 "ㅔ"와 "ㅐ", "ㅖ"와 "ㅒ", 그리고 "ㅚ"와 "ㅙ"와 "ㅞ"는 각각 동일한 그룹으로 간주하였으며, 그 결과 모음 중에서 음소 분할 단위로 선정된 것은 17개이다. 그리고 자음에 있어서 여러 가지 음운 현상에 의해 변이음들이 나타날 수 있는데 본 연구에서는 폐쇄음의 불파음화에 대한 변이음(g', d', b')과, 유음에 대해 'r'(탄설음)의 'ㅣ'(설측음)에 대한 변이음을 고려했다. 따라서 최종 음성 분할 단위로 선정된 것은 묵음(silence) 및 짧은 휴지 구간(short pause)에 대해 2개, 자음 23개, 모음 17개로 총 42개의 단위이다. 표 1에 본 논문에서 사용한 음성 단위 목록과 각각의 기호를 나타내었다.

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. 이와 관련해서 HMM topology

의 길이, 즉 모델의 시작에서 끝으로 가는 동안 거쳐야 하는 최소 상태수는 매우 중요한 역할을 하며 음소의 최소 지속기간보다 작아야 한다.

<표 1> 본 논문에서 사용된 음성분할 단위 및 기호 목록

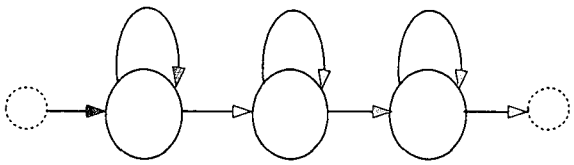
번 호	기 호	설 명	번 호	기 호	설 명
1	g	ㄱ	22	r	ㄹ
2	g'	불파음화 ㄱ	23	l	[r]의 [l]되기
3	G	ㄲ	24	a	ㅏ
4	d	ㄷ	25	ja	ㅑ
5	d'	불파음화 ㄷ	26	v	ㅓ
6	D	ㄸ	27	jv	ㅕ
7	b	ㅂ	28	o	ㅗ
8	b'	불파음화 ㅂ	29	jo	ㅛ
9	B	ㅃ	30	u	ㅜ
10	s	ㅅ	31	ju	ㅠ
11	S	ㅆ	32	U	ㅡ
12	z	ㅈ	33	i	ㅣ
13	Z	ㅉ	34	E	ㅐ+ㅔ
14	c	ㅊ	35	je	ㅒ+ㅖ
15	k	ㅋ	36	wE	ㅚ+ㅞ+ㅙ
16	t	ㅌ	37	wa	ㅘ
17	p	ㅍ	38	wi	ㅟ
18	h	ㅎ	39	wv	ㅝ
19	m	ㅁ	40	Wi	ㅢ
20	n	ㄴ	41	sp	짧은 휴지 구간
21	N	ㅇ	42	silence	묵음

그러나 적절한 음소 지속기간에 대한 모델 없이 topology 길이가 너무 짧은 것도 바람직하지 않다. 실제로 파열음에서의 burst의 길이는 5-6ms에 불과할 경우도 있으며 길 경우 30-40ms에 이르기도 한다. 음소지속기간에 대해 구체적인 모델을 설정하는 것도 성능에 큰 영향을 줄 것으로 판단된다. 또한 HMM을 monophone 모델과 triphone 모델 중 어떤 것을 선택하느냐에 따라서도 성능이 달라질 것이다. 이러한 사항의 결정은 실제 음소 분할 실험을 통한 성능평가에 따라 이루어지도록 한다.

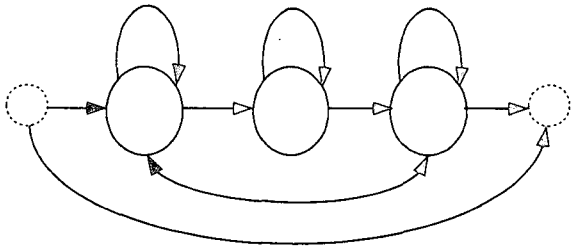
그리고, HMM topology등은 음성분석 시간단위가 얼마인가에 따라서도 영향을 받기 때문에 음성분할과 음성인식에 사용하는 모델 구조는 별도로 구성하는 것도 고

려해야 한다. 음소 그룹에 따라 서로 다른 topology를 두는 방법도 검토할 수 있다.

본 논문에서 구성한 자동 음성 분할 baseline 시스템에서는 3개의 상태와 각각 연속확률분포를 가지는 HMM을 사용하였으며, 그림 1에 나타낸 바와 같이 일반적인 변이음과 짧은 휴지 구간(short pause)에 대해 각각 다른 topology를 가지도록 구성하였다. 일반적인 변이음 모델은 skip path 없이 자기 자신 또는 바로 이웃하는 상태로의 천이만 허용되는 형태로 모델링 하였으며, 짧은 휴지 구간(short pause)의 경우 주위 잡음에 의해 나타나는 짧은 burst나 화자의 발성 특성에 따라 나타나는 숨소리 등 다양하게 나타나는 잡음 성분을 모델링 하기 위해 skip path가 있고 상태천이가 일반 음소 모델 보다 복잡한 형태로 구성하였다.



(a) 일반적인 변이음에 대한 HMM 구성



(b) Short pause를 위한 HMM 구성

<그림 1> 음소 모델을 위한 HMM 구성

3. 성능개선을 위한 접근 방법

본 논문에서는 자동음성분할 시스템의 성능향상을 위해 음향 모델링 방법과 에너지 기반의 후처리 방식을 검토하였다. 먼저 HMM의 기본 단위로 triphone 모델과 monophone 모델을 검토하였으며, 모델링 방법으로 Baum-Welch 알고리즘과 segmental K-means 알고리즘의 조합에 의한 접근 방식을 검토하였다. 그 외에도, 묵음과 음성, 그리고 음성과 묵음 사이의 경계에서 상당 부분의 오류가 발생하는 것을 감안하여 에너지 기반의 후처리 방식을 도입하였다.

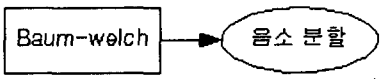
3.1. HMM 단위 선정

음성 인식에서는 동일한 음소일지라도 좌우 문맥(context)에 따라 특성이 다르기 때문에 좌우 문맥을 고려한 triphone의 성능이 monophone의 성능보다 일반적으로 우수하다. 그러나 음성 인식 성능과 음소 분할 성능 사이에는 별다른 상관관계가 없다는 보고된 바 있어서[4], 실제 음소 분할을 위한 HMM의 모델링 단위는 단순히 음성인식에서 우수한 성능을 얻은 결과를 그대로 사용하기 보다 실험을 통한 성능평가에 따라 선정할 필요가 있다.

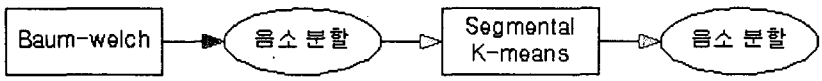
따라서 본 논문에서는 monophone 모델과 triphone 모델 각각에 대해 mixture 수를 바꾸어 가면서 모델을 구성하고 자동 음성 분할 성능을 평가하도록 하였다. Triphone으로 모델링 할 경우 훈련용 데이터의 크기에 비해 모델의 수와 추정해야 할 파라미터 수가 너무 많아지는 문제의 해결을 위해 상태들을 tying 하는 방법이 사용되며, 여기에는 tree based clustering(TBC)과 data driven clustering(DDC)의 두 가지가 있다. 본 논문에서는 이 중에서 TBC를 사용하였다. TBC의 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 논문에서는 최소 관찰벡터의 수는 고정시킨 상태에서 DL 값을 변화시킴으로써 tied state수를 조절하였다.

3.2. HMM 모델링 방법

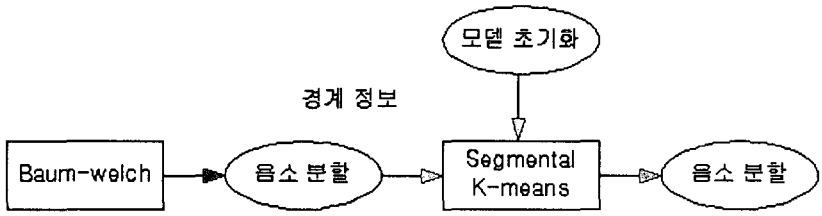
본 논문에서는 세 가지 HMM 모델링 방법에 대해서 자동 음소 분할 성능을 평가하였다. 첫 번째 모델링 방법은 가장 대표적인 방법인 Baum-Welch reestimation 모델링을 한 경우(BW)이고 다른 두 가지 방법은 첫 번째 모델링 방법에서 얻어진 음소 경계정보를 이용해서 segmental K-means 알고리즘을 이용하여 HMM을 다시 훈련하는 방법이다. 이때 초기 모델로 어떤 것을 사용하느냐에 따라 두 가지를 검토하였는데, 하나는 Baum-Welch reestimation에 의해 만들어진 최종 모델을 초기 모델로 사용하는 경우(BW-SKM)이고, 나머지 하나는 이전 모델을 무시하고 전체 음성 DB에 대해서 얻어진 평균 벡터와 공분산 행렬로 다시 초기화하는 경우(SKM)이다. 본 논문에서 검토한 세 가지 HMM 모델링 방법이 그림 2에 나타나 있다.



(a) Baum-Welch (BW)



(b) Baum-Welch & Segmental K-means (BW-SKM)

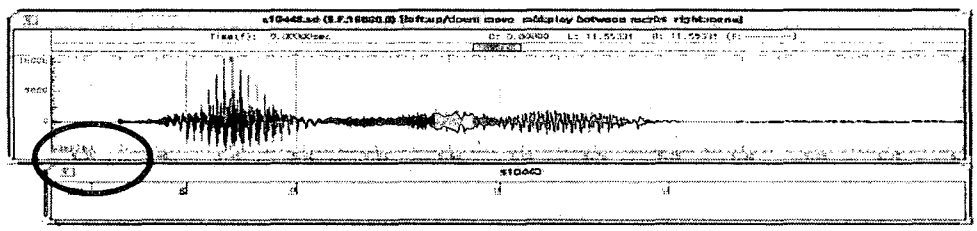


(c) Segmental K-means (SKM)

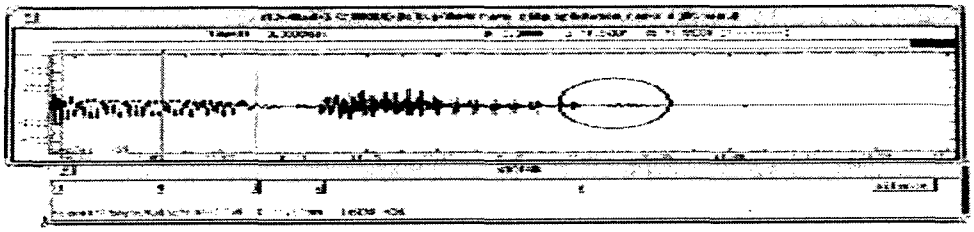
<그림 2> 본 논문에서 검토한 HMM 모델링 방법들

3.3. 에너지 기반의 후처리

HMM을 이용하여 자동 음소 분할을 하는 경우 묵음과 음성 사이, 그리고 음성과 묵음 사이에서 상당한 경계 오차가 발생하며, 그림 3에 그 예가 나타나 있다. 그림 3(a)는 어절의 시작부분에서 묵음 다음에 음소가 오는 경우를 보여주며 자동 분할시 음소 구간에 묵음 구간을 많이 포함하고 있다. 그림 3(b)에서는 어절의 끝 부분에서 음소 다음에 묵음이 오는 경우 에너지가 천천히 작아지는데 자동 분할시 묵음 구간에 음성부분을 많이 포함한 예를 보여주고 있다. 이러한 상황에서 프레임 에너지를 이용하여 적절한 후처리를 해 줌으로써 음성분할 성능의 향상을 기대할 수 있다.



(a) 어절의 시작 부분. 묵음+음소



(b) 어절의 끝 부분, 음소+묵음

<그림 3> 자동 음소 분할의 문제점

에너지 기반의 후처리를 위한 threshold는 묵음과 자음, 묵음과 모음, 음소와 묵음이 오는 세 가지 경우로 나누어서 각각 다른 값을 사용하였다. 그 이유는 어절의 시작 부분의 경우, 묵음 다음에 자음으로 시작하는 경우가 모음으로 시작하는 경우에 비해 에너지가 작기 때문에 동일한 threshold의 적용이 바람직하지 않기 때문이다. 그리고 어절의 끝 부분에서는 어절의 시작부분과 달리 에너지가 서서히 작아지면서 fluctuation이 나타나는 경우가 많아서, 이 경우 에너지를 분석하기 위한 window 크기를 어절의 시작부분과 다르게 적용하는 것이 필요하다. 따라서 본 논문에서는 묵음과 연결되는 어절의 시작부분과 끝 부분에 다른 크기의 window를 사용해서 에너지를 구하고 threshold를 적용하였다. 그림 4는 어절의 시작부분에서 묵음 다음에 모음이 오는 경우에 프레임별 log 에너지 분포와 threshold를 나타내었다. 본 논문에서 에너지 threshold를 구하는 데 사용한 방법은 다음과 같다.

Step 1. Window 내에 평균 log에너지를 구한다. 이 때 window의 크기는 어절의 시작부분에서 10 ms, 그리고, 어절의 끝 부분에서 20ms를 사용한다.

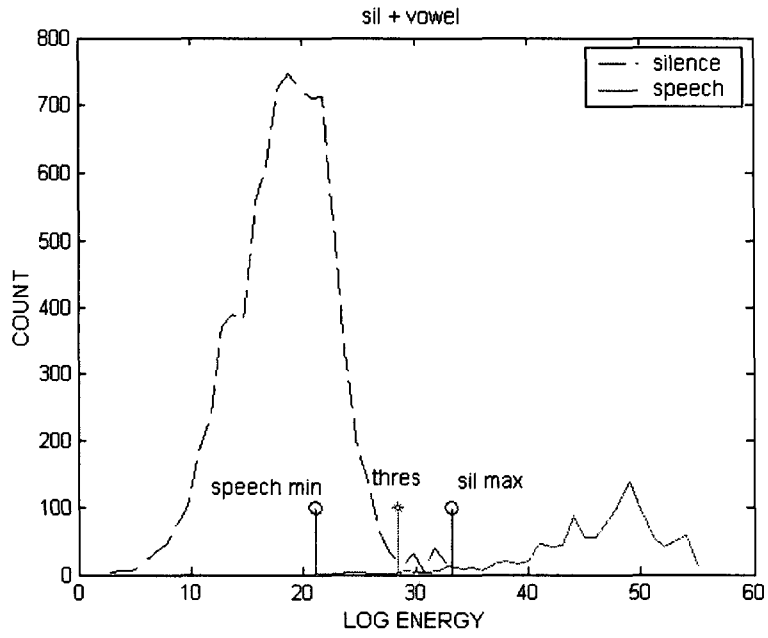
Step 2. 음성 부분의 최대값(speech max)과 묵음 부분의 최소값(sil min)을 구한다.

Step 3. speech max에서 sil min까지 threshold 값을 변화시키면서 다음 식을 처음 만족하는 threshold를 찾는다. (이 때 const. 값은 경험적으로 결정하며 본 논문에서는 0.2로 정하였다.)

$$\frac{\sum_{speech\_max}^{thres} \#\ speech\ \ frame}{\sum_{sil\_min}^{thres} \#\ sil\ \ frame} \geq const.$$

4. 실험 결과 및 분석

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



<그림 4> 프레임별 log 에너지의 분포 및 threshold 선정

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. 음소 분할 기준은 원광대학교 음성정보기술산업지원센터(SiTEC)의 기준안을 토대로 하였으며, 수작업에 의한 음소 분할 작업은 전문 레이블링 교육을 이수한 두 명의 전자공학과 학생이 수행하였다.

4.2. 실험 결과

표 2는 HMM을 triphone으로 모델링 할 때 tied state 수를 변화시킴에 따른 음소 분할 성능을 나타낸 것이다. 음성 인식의 경우 tied state수를 조절함에 따라 성능이 상당히 달라지지만, 표 2의 결과를 보면 tied state수에 따른 음소 분할 성능의 변화는 거의 없었다.

표 3은 triphone과 monphone 모델에 대해 mixture 수를 1개에서 5개까지 증가시키면서 음성분할을 한 실험 결과이다. 이 때 triphone은 표 2에서 성능이 가장 우수한 경우에 대해서 실험하였다. 음성인식의 경우와는 달리 monophone의 음소 분할 성능이 triphone보다 우수하게 나타났다.

<표 2> Tied state 수에 따른 성능
(총 state 수 : 41982)

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 3> 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 4> Monophone 모델에서 모델링 방법에 따른 성능
(a) 후처리를 하지 않은 경우

Modeling		BW	BW-SKM	SKM
오차범위	≤10ms	59.6	61.7	69.1
	≤20ms	83.6	84.2	82.8

(b) 후처리를 하지 않은 경우

Modeling		BW	BW-SKM	SKM
오차범위	≤10ms	62.4	64.2	71.3
	≤20ms	85.3	86.3	84.2

표 4의 (a)는 3.2절에서 언급한 세 가지 HMM 모델링 방법에 따른 성능을 나타내었고, (b)는 각각의 모델링 방법으로 얻어진 음소 경계 정보를 토대로 에너지 기반 후처리를 한 결과를 나타내었다. 3.3절에서 언급한 바와 같이 에너지에 의한 후처리는 자동 음소 분할에서 묵음 구간이 검출된 부분에서만 적용하였다.

표 4를 보면 단지 Baum-Welch reestimation 모델링에 의해 음소 분할한 경우보다 Baum-Welch reestimation 모델링에 의해 얻어진 음소 경계 정보를 이용해서 segmental K-means 알고리즘으로 다시 모델링 한 경우(BW-SKM 및 SKM)의 성능이 더 우수하다. 그리고 전체 DB에 대한 평균과 공분산에 의해 초기 모델을 재설정한 경우(SKM)가 Baum-Welch reestimation의 최종 모델을 초기 모델로 사용하는 경우(BW-SKM)보다 성능이 더 우수한데, 이는 BW-SKM의 경우 segmental K-means 알고리즘을 적용하더라도 Baum-Welch reestimation의 최종 모델에서 크게 변화되지 않기 때문으로 판단된다. 또한 에너지 기반의 후처리를 한 경우 모델링 방법과 관계없이 전체적으로 성능이 개선됨을 알 수 있다.

표 5는 표 4(a)에서 가장 좋은 성능을 나타낸 모델(SKM)에 대해 후처리 후 음소 그룹 pair별 자동 음소 분할 성능을 나타낸 것이다. 여기서 음소 그룹으로는 파열음, 불파음, 마찰음, 파찰음, 유음, 이중모음, 단모음의 8 가지를 선정하였다.

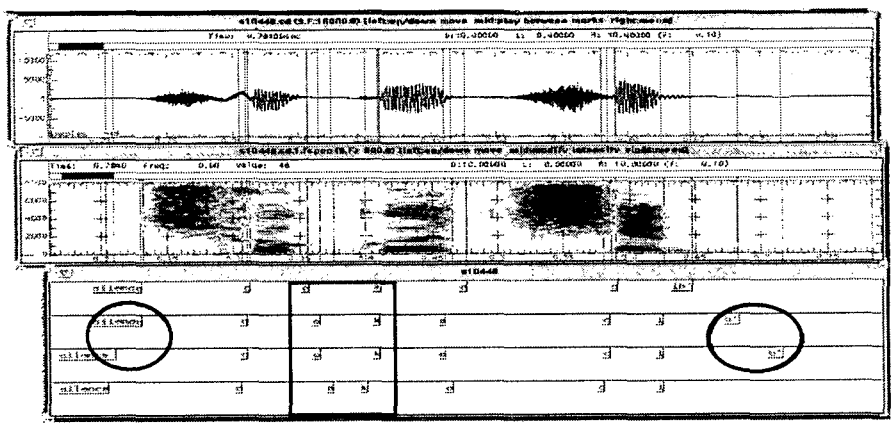
<표 5> SKM 모델링 방법에 의한 음소그룹 pair별 분할성능
(각 pair별로 오차범위 10ms 이내에 들어오는 비율)

선행 \ 후행	파열음	불파음	마찰음	파찰음	비음	유음	이중모음	단모음
파열음	-	-	-	-	-	-	97.4	91.9
불파음	12.5	-	37.5	0.0	0.0	-	-	100
마찰음	-	-	-	-	-	-	83.3	85.6
파찰음	-	-	-	-	-	-	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

표 5에서 보면 음소분할 성능의 저하가 크게 일어나는 경우를 두 가지로 나누

어 볼 수 있다. 첫 번째는 불파음 다음에 파열음, 마찰음, 파찰음이 오는 경우인데, 이 경우에 불파음의 폐쇄 구간과 파열음, 마찰음, 파찰음의 폐쇄구간이 구별되지 않기 때문에, 수작업으로 음소 분할 시 두 음소 사이의 묵음 구간의 절반 위치를 인위적인 경계로 정한 것과 관련이 있다고 보여진다. 두 번째는 비음과 비음, 유음과 유음, 모음과 모음 혹은 모음과 유음이 연속해서 오는 경우이다. 이 경우는 두 음소 사이의 경계를 사람의 눈과 귀로도 사실 구별하기 어려우므로 수동 음소 분할의 정확성에도 문제가 있으며, 수동 음소 분할이 어려운 부분은 자동 음소 분할도 어렵다는 것을 나타낼 수도 있다.

그림 5는 본 논문에서 검토된 방법으로 자동 음소 분할한 예를 보여 준다. 전반적으로 BW-SKM 방법보다 SKM 방법이 수작업에 의한 음소 분할 경계에 근접하는 것을 알 수 있다(그림에서 사각형 표시부분 참조). 또한, 자동 음소 분할한 후 에너지 기반의 후처리 한 경우, 묵음과 음성의 시작 부분인 ㅊ(c) 사이의 경계와 불파음화 ㅂ(b')과 묵음 사이의 경계가 수작업의 의한 경계에 더 다가감을 확인할 수 있다(그림에서 원으로 표시된 부분 참조).



<그림 5> 검토된 방법으로 자동 음소 분할된 결과. ('초코칩')

5. 결 론

본 논문에서는 코퍼스 기반의 음성 합성기를 개발하기 위한 준비 작업으로서, 합성 단위를 자동으로 분할해주는 방법에 대해서 논의하였다. 이를 위하여 HMM을 이용한 여러 가지 음향 모델링 방법과 에너지 기반의 후처리 방법이 검토되었다. 음소 분할 실험 결과 음성 인식의 경우와는 달리 triphone 모델에서 tied state 수에 따른 영향은 미미했으며, triphone 모델보다는 monophone 모델 이용 시 더 우

수한 성능을 나타내었다. 본 논문에서 구현한 시스템의 자동 음소 분할 성능은 후처리를 하지 않은 경우, 오차범위가 10ms인 경우에 69.1%이며, 오차범위가 20ms인 경우에 82.8%로 나타났다. 그리고, 에너지 기반의 후처리를 한 경우 동일한 오차범위에 대해 각각 71.3% 및 84.2%의 성능을 얻었다.

모음의 경우 자음에 비해 평균적인 음소 지속 시간이 길기 때문에 모음에 대한 HMM은 상태의 수 등 topology를 자음의 HMM과 달리하는 것이 바람직하다. 앞으로 음소 그룹별로 HMM topology를 다르게 구성하는 방안과 음소 그룹 pair 특성에 따라 각각 다른 후처리 방법을 적용하는 방안에 대해 연구가 진행될 예정이다.

참 고 문 헌

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

[2] Svendsen, T. and K. Kvale (1990), Automatic alignment of phonetic labels with continuous speech, in *Proc. ICSLP* Nov., pp.429~432.

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

[4] Ljolie, A. and M. D. Riley (1991), Automatic segmentation and labeling of speech, in *Proc. ICASSP* Apr., pp.473~476.

[5] Brugnara, F., D. Falavigna and M. Omologo (1993), Automatic segmentation and labeling of speech based on hidden Markov models, *Speech Communication.* vol.12. no.4. Aug., pp.357~370.

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

[7] 홍성태, 김제우, 김형순(1998), 자동 음성분할 및 레이블링 시스템의 성능향상, 「말소리」 35-36. 12월, 대한음성학회, pp.175~189.

접수일자: 2002년 4월 30일
게재결정: 2002년 5월 24일

▶ 박혜영(Haeyoung Park)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-1704
Fax: 051) 515-5190
E-mail: phyoe@pusan.ac.kr

▶ 김형순(Hyungsoon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
Fax: 051) 515-5190
E-mail: kimhs@pusan.ac.kr