자동 음성분할 및 레이블링 시스템의 성능향상

홍 성태, 김 제우, 김 형순(부산대 전자공학과)

<차 례>

<Abstract>

Performance Improvement of Automatic Speech Segmentation and Labeling System

Seong Tae Hong

Database segmented and labeled up to phoneme level plays an important role in phonetic research and speech engineering. However, it usually requires manual segmentation and labeling, which is time-consuming and may also lead to inconsistent consequences. Automatic segmentation and labeling can be introduced to solve these problems. In this paper, we investigate a method to improve the performance of automatic segmentation and labeling system, where Spectral Variation Function(SVF), modification of silence model, and use of energy variations in postprocessing stage are considered.

In this paper, SVF is applied in three ways: (1) addition to feature parameters, (2) postprocessing of phoneme boundaries, (3) restricting the Viterbi path so that the resulting phoneme boundaries may be located in frames around SVF peaks. In the postprocessing stage, positions with greatest energy variation during transitional period between silence and other phonemes were used to modify boundaries.

In order to evaluate the performance of the system, we used 452 phonetically balanced word(PBW) database for training phoneme models and phonetically balanced sentence(PBS) database for testing. According to our experiments, 83.1% (6.2% improved) and 95.8% (0.9% improved) of phoneme boundaries were within 20ms and 40ms of the manually segmented boundaries, respectively.

1. 서론

연속음성을 음소 단위로 분할하고 레이블링하는 일은 음성처리 연구의 여러 분야에서 매우 유용한 일이다. 음소를 기본단위로 하는 음성인식 시스템의 훈련과정에서 음소단위로 정확히 분할된 음성 데이터베이스가 충분히 제공된다면 인식성능을 크게 향상시킬 수 있다. 또한, 음성합성에서도 각 음소에 해당하는 음성부분을 추출하고 음소지속기간 정보를 추출하는데 음소 분할된 음성 데이터베이스가 필요하며, 음성학 연구에도 중요한 정보로 활용될 수 있다.

특히 최근의 음성인식 기술은 주로 코퍼스(corpus) 기반의 접근방식, 즉 대용량의 음성 데이터를 분석하여 얻어진 통계적 모델에 의거하여 인식을 수행하는 방식에 기초를 두고 있다. 이러한 접근방식이 성공을 거두기 위해서는 많은 양의 음성 데이터를 수집하는 것만으로는 충분하지 않으며, 음성 데이터를 음소와 같은 음성의 기본단위들로 분할하고 레이블링하여 그 다음 단계의 통계적 처리가 가능하도록 가공하는 작업이 필수적으로 요청된다. 음성신호의 음소분할 및 레이블링 작업은 사람이 직접 수행할 수 있지만, 수작업에 의한 음소분할 및 레이블 링은 다음과 같은 문제점을 지닌다(Leung & Zue, 1984). 첫째로 이 과정은 스펙트로그램 판독 및 반복되는 듣기평가를 통해 이루어지므로 매우 지루한 작업일 뿐만 아니라 많은 시간이 소요되게 된다. 둘째로 수작업에 의한 음소분할은 높은 수준의 음성학적 지식을 요하며, 소수의 음성학 전문가에 의존할 수밖에 없다. 세째로 음소경계 선정을 위한 구체적인 판단기준을 미리 정해놓더라도 상당 부분의 경우 주관적인 판단을 피할 수 없으며, 이에 따라 음소경계 선정과정에서의 일관성이 보장되지 못한다(Eisen et al., 1992). 그밖에도 지루한 작업이 계속됨에 따른 판단오류도 문제가 될 수 있다.

음소분할 및 레이블링 작업이 자동으로 수행될 수 있다면 위에서 언급한 문제들이 해결될 수 있으며, 부분적으로나마 자동적인 음소분할 및 레이블링이 가능해지면 수작업에 의한 업무량이 크게 절감될 수 있다. 이에 따라 자동 음소분할 및 레이블링을 위한 여러 가지 방법들이 개발되어 왔으며(Ljolie & Riley, 1991)(Wheatley et al., 1992)(Brugnara et al., 1993), 특정언어에 국한된 것이지만 상품화된 음성 레이블링 시스템이 나오기도 했다(The Aligner, 1994). 우리말에 대해서도 지금까지 연구 개발되어 온 자동 음소분할 및 레이블링 기술들을 이용해서 자동 음소분할 및 레이블링 시스템이 구현되었다(성종모 등, 1997). 본 논문에서는 음성의 변화특성을 잘 표현하는 Spectral Variation Function(SVF)를 이용하여 한국어 자동 음성분할 및 레이블링 시스템의 성능 향상을 위한 방법을 고려하였다. 또한 후처리에서 에너지 변화량을

이용하여 추가적인 성능 향상을 유도하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 자동 음소분할 및 레이블링 기술에 대해서 서술하고, 3장에서는 음소분할과 레이블링 시스템 구성에 대해서 다루며, 4장에서는 성능향상을 위해 SVF를 시스템에 적용한 방법을 검토한다. 또한 후처리에서 에너지 변화량을 이용하는 방법을 설명하였다. 그리고 5장에서는 음소분할 및 레이블링 시스템을 이용해서 음성 분할한 실험 결과를 보이고, 마지막으로 6장에서 결론을 맺는다.

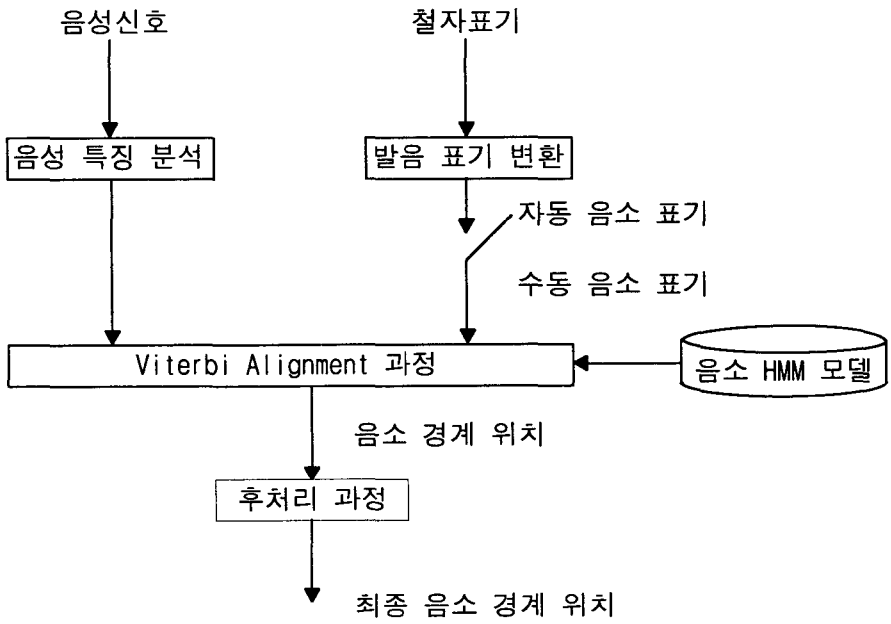
2. 자동 음소분할 및 레이블링 기술

자동 음소분할 및 레이블링은 일반적으로 음소표기(phonetic transcription) 등의 언어학적 정보가 제공되는 상황에서 이루어지며, 그 과정은 다음과 같다. 먼저 훈련과정에서 각각의 음소들에 대한 대표 패턴 또는 통계적 모델들을 미리 구성한다. 그리고, 입력음성과 이에 해당하는 음소 시퀀스가 입력되면, 구성 가능한 음소 시퀀스에 의해 각 음소들의 대표 패턴 또는 모델들을 연결해서 입력음성과 매칭시키는 과정에서 음소경계 정보가 얻어진다. 이러한 패턴 혹은 모델 매칭 방법으로는 Dynamic Time Warping(DTW) 방법과 Hidden Markov Model((HMM) 방법이 사용될 수 있다. 이 두 방법 중에서 DTW 방법에서는 복수 개의 대표 패턴들만으로 음성신호에 내재된 변화요인들을 대처하는 데에 한계가 있으므로, 이러한 변화요인들을 확률모델 형태로 다루는 HMM 방법이 널리 사용되고 있다(Ljolie & Riley, 1991)(Wheatley et al., 1992)(Brugnara et al., 1993).

자동 음소분할 및 레이블링을 위해 제공되는 언어학적 정보로는 음소표기(phonetic transcription)와 철자표기(orthographic transcription)를 들 수 있다. 그 중에서 음소표기가 주어지는 경우에는 음소분할 및 레이블링은 이미 이루어진 상황이므로 음소들 간의 묵음구간의 검출과 각각의 음소들의 경계정보만을 찾아내면 된다. 이에 반해서 철자표기가 정보로 주어지는 경우에는 철자표기가 발음되는 형태의 음소표기로 바꾸어 주는 과정이 필요하며, 동일한 단어라도 다양한 형태로 발음될 수 있으므로 복수개의 발음사전을 필요로 하기도 한다(Wheatley et al., 1992).

<그림 1>에 언어학적 정보가 제공되고, 매칭방법으로 HMM을 사용할 때의 자동 음소분할 및 레이블링 시스템의 구성도가 나타나 있다. 음성 신호가 입력되면 음성 특징 분석 과정에서 음성특징계수들을 추출하고 미리 구성된 음소 HMM 모델들을 음

소표기 정보에 따라 연결하여 추출된 음성특징계수들의 시퀀스와 Viterbi 알고리즘에 의해 최적 time alignment를 수행하는 과정에서 각각의 음소의 경계위치가 얻어진다. 이 때, 음소표기 정보는 사람의 수작업에 의해 직접 제공되거나, 철자표기 텍스트 정보를 발음표기 변환 프로그램에 의해 자동 변화된 결과가 사용된다. 이 시스템의 출력은 각각의 음소 레이블과 그 경계위치들이며, 후처리 과정에서 음성학 전문가에 의한 수동 보정 작업이 이루어진다.



<그림 1> HMM을 이용한 자동 음성분할 및 레이블링 시스템의 구성도

3. 자동 음소분할 및 레이블링 시스템의 구성

3.1 음성신호 전처리 과정

입력 음성에 대해 16kHz로 샘플링된 10ms마다 20ms 크기의 분석창을 씌워 음성 특징을 추출하였다. 음성분석 시간단위로는 일반적으로 음성인식에서 널리 사용되는 10ms의 정밀도를 사용하였으나, 정교한 음소분할 및 레이블링을 위해서는 앞으로 보다 미세한 음성분석 시간단위가 필요하다고 판단된다. 참고로 TIMIT 음성 데이터베이스 구축 시 사용된 자동 음소분할에서는 2.5ms의 시간단위가 사용된 것으로 알려

져 있다(Ljolie & Riley, 1991).

음성인식을 위한 음성특징분석 파라메터는 음소변별력이 뛰어나면서 음성학적으로 중요하지 않은 변화요인에 둔감한 특성을 가지는 것이 요구된다. 지금까지 음성인식에 효과적으로 사용되어 온 음성특징 파라메터로는 LPC cepstrum 또는 Mel-frequency cepstrum과 이들의 시간 축 미분 값들을 들 수 있으며, 단구간 에너지와 그 미분치도 중요한 정보로 활용된다. 본 시스템에서는 청각기관의 주파수 선택특성을 고려한 Mel-frequency cepstrum계수(MFCC)와 delta MFCC, 그리고 정규화된 단구간 에너지와 그 미분치를 음성특징 파라미터로 사용하였다.

3.2 레이블링 단위

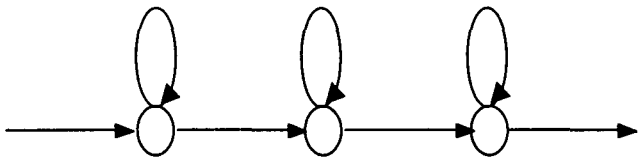
음성을 분할하고 레이블링하기 위한 기본단위로는 음소, 유사음소(phoneme-like unit), 변이음 등이 사용될 수 있다. 본 연구에서는 유사음소를 기본단위로 선정하였다. 유성음화와 불파음화, 그리고 [r]의 [l]되기 등 3가지 규칙을 적용하고 'ㅖ'와 'ㅒ', 그리고 '괴', '궤', '괘'를 각각 하나로 묶어서 45개의 유사음소set을 정하고 묵음을 포함하여 총 46개의 레이블링 단위를 선정하였다. 본 논문에서 구현한 자동 음성분할 및 레이블링 시스템에 사용되는 유사음소 목록과 기호는 표 1에 나타나 있다.

<표 1> 유사음소 및 묵음 기호 목록

번 호	기 호	설 명	번 호	기 호	설 명
1	g	ㄱ	24	m	ㅁ
2	gg	유성음화 ㄱ	25	n	ㄴ
3	g'	불파음화 ㄱ	26	N	ㅇ
4	G	ㄲ	27	r	ㄹ
5	d	ㄷ	28	l	[r]의 [l]되기
6	dd	유성음화 ㄷ	29	a	ㅏ
7	d'	불파음화 ㄷ	30	ja	ㅑ
8	D	ㄸ	31	v	ㅓ
9	b	ㅂ	32	jv	ㅕ
10	bb	유성음화 ㅂ	33	o	ㅗ
11	b'	불파음화 ㅂ	34	jo	ㅛ
12	B	ㅃ	35	u	ㅜ
13	s	ㅅ	36	ju	ㅠ
14	S	ㅆ	37	U	ㅡ
15	z	ㅈ	38	i	ㅣ
16	zz	유성음화 ㅈ	39	E	ㅐ + ㅔ
17	Z	ㅉ	40	je	ㅒ + ㅖ
18	c	ㅊ	41	wE	ㅚ + ㅞ + ㅙ
19	k	ㅋ	42	wa	ㅘ
20	t	ㅌ	43	wi	ㅟ
21	p	ㅍ	44	wv	ㅝ
22	h	ㅎ	45	Wi	ㅢ
23	hh	유성음화 ㅎ	46	SIL	묵음

3.3 음소 모델 구성

HMM을 이용하여 유사음소모델을 구성하기 위해서는 관찰확률분포를 이산분포, 연속분포 또는 준 연속분포 중에서 선정해야 하며, 상태 수와 천이방식 등 HMM topology와 더불어 일부 파라메터의 tying 여부를 정해야 한다. 이러한 사항의 결정은 음소분할 실험을 통한 성능평가에 따라 이루어져야 할 것이다. 본 논문에서는 그림 2에 나타낸 형태의 모델을 사용하였다. 관찰확률분포로는 각 상태 속에서 관찰될 연속확률분포를 사용하였다.



<그림 2> 유사음소 모델링을 위한 HMM 구조

4. 자동 음성 분할 시스템의 성능 향상을 위한 방법 고려

4.1 Spectral Variation Function(SVF)

SVF는 음성신호의 연속적인 관찰벡터들간의 상호 유사성을 측정하는 파라메터로 사용된다(Brugnara et al., 1993),(Flammia et al., 1995). SVF의 값은 0과 1사이에서 정규화 되어 있으며 그 값이 작으면 그 프레임의 주변에서는 음향학적인 정보가 서로 비슷하다는 것을 표현하고, 반대로 SVF값이 크면 그 프레임에서 음향학적인 정보의 변화가 큼을 나타낸다. SVF의 peak값은 그 순간에 음향학적 정보가 가장 많이 변화하고 있음을 의미하고, 따라서 음소의 경계를 표현하는 경향이 있다.

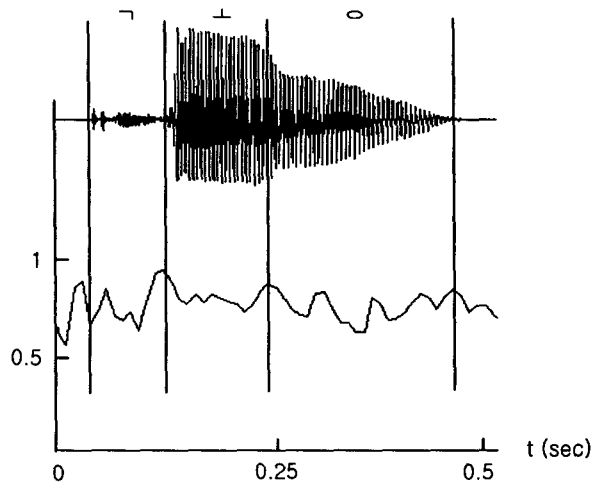
SVF는 음성특징 벡터열 $O_1 O_2 \cdots O_t \cdots O_T$ 에 대해, 다음과 같은 수식으로 전개된다.

$$P_t = O_t - A_t \quad \text{and} \quad A_t = \frac{1}{2L+1} \sum_{k=-L}^{L} O_{t+k} \tag{1}$$

$$\cos(i,j) = \frac{P_i \cdot P_j}{\| P_i \| \, \| P_j \|}, \quad t-L \le i < t, \; t < j \le t+L \tag{2}$$

$$SVF(t) = \frac{1}{2}\left[1 - \frac{1}{L^2} \sum_{i,j} \cos(i,j)\right] \tag{3}$$

이 때, L은 좌우로 고려한 윈도우 프레임의 개수를 나타낸다. 본 논문에서는 좌우로 3프레임씩 고려하여 SVF값을 구했다. SVF는 12차 MFCC와 에너지에 가중치를 주어 구했다. 그림 3은 실제로 수작업으로 분할한 경계들과 SVF로 구한 값들과의 비교이다. 그림에서 보는 바와 같이 SVF peak가 수작업으로 분할한 경계와 일치 또는 근접하는 경우가 많이 있으나, 경계위치가 아닌 곳에서의 추가적인 peak도 상당히 있는 것으로 관찰되었다.



<그림 3> 단어 "공"에 대한 수작업으로 분할한 경계와 SVF의 비교

4.2 SVF를 이용한 성능개선 방법 검토

먼저 SVF를 음성특징 파라메터의 일부로 사용하였다. Baseline 시스템에서는 청각기관의 주파수 선택특성을 고려한 mel-frequency cepstral coefficients (MFCC)와 에너지, 그리고 그 미분치를 포함하여 모두 26차의 특징 파라메터를 사용하는데, 여기에 SVF를 특징 파라메터로 첨가시켜 27차로 음성을 훈련시킨 모델을 사용하여 음소의 경계를 구했다. 그리고 SVF를 구하는 과정에서, 정규화 된 에너지부분을 고려할 때 가중치를 줌으로써 에너지 분포가 SVF값에 더 많은 영향을 주도록 하였다. 두 번째로 baseline 시스템을 사용하여 음성을 훈련시킨 모델을 사용하여 음소의 경계를 자른 후, 가장 가까운 SVF의 peak가 있는 곳으로 경계를 이동시키는 후처리 방법을 고려하였다. 마지막 방법으로는 처음에 고려한 것과 같은 방법으로 SVF를 특징 파라메터에 포함시켜 유사음소 모델을 훈련하고, 그리고, Viterbi alignment과정에서

SVF peak가 발생하는 순간의 주변 프레임에서 다음 음소로 전이할 확률을 계산하는 부분에서 적절한 보상을 가함으로 더 정확한 경계를 찾도록 유도하였다.

4.3 묵음모델의 수정을 통한 성능개선 방법 검토

묵음모델은 짧은 구간의 묵음을 모델링하기 위해 1번째 상태에서 3번째 상태로의 천이와 3번째 상태에서 1번째 상태로의 천이가 허용되도록 모델을 구성하였다. 하지만 음소의 순서를 이미 알고 Viterbi decoding 하는 경우, skip을 허용하면 묵음구간에서 성능저하가 발생하는 것으로 판단된다. 따라서 묵음모델에 대해서도 그림 2와 같이 모델링하는 방법을 적용하였다.

4.4 에너지 변화량을 이용한 성능개선 방법 검토

묵음에서 다음 음소로 전이하는 구간의 경우, 사람이 직접 수작업으로 이 구간의 음소경계를 분할할 때는 비교적 정확한 경계분할이 가능하다. 왜냐하면, 이 구간에서는 에너지의 변화가 확실하게 관찰되기 때문이다. 따라서 묵음에서 다음 음소로 전이하는 경계 주변에서는 후처리 방법을 통해 에너지 변화량이 가장 큰 위치로 음소경계를 수정함으로 더욱 정확한 경계를 찾을 것으로 판단된다.

5. 실험 및 결과

본 논문에서 구현한 음소분할 및 레이블링 시스템의 성능향상 평가를 위한 실험이 수행되었다. 본 논문에서는 각각의 유사음소 모델의 훈련을 위해 국어공학센터의 음소적으로 균형이 잡힌 452개 단어(Phonetically Balanced Word (PBW)) 데이터베이스에서 남성화자 38명분의 음성을 사용하였다. 그리고 음소분할 실험은 국어공학센터의 음소적으로 균형을 이루는 (Phonetically Balanced Sentence(PBS))중 10명의 남성화자로부터 5문장씩 선택하여 모두 50문장을 가지고 성능향상에 관한 실험을 하였다(김봉완, 이용주 외, 1996). 국어공학센터의 PBS음성 데이터베이스는 방음 부스에서 Senheizer HMD224X를 사용하여 녹음되었으며, 디지털 오디오 테이프에 저장되었다. A/D변환은 Sun Workstation Sparc 20환경에서 실시되었고, DAT-Link+가 사용되었다. 음성 데이터는 16 kHz로 sampling되고 16 Bit로 양자화 되었다. 이 실험을 위해 음성학에 대한 지식을 가진 사람이 일정한 훈련을 거친 다음, 음성 데이터를 분할 및 레이블링 하였다. 성능비교에는 수작업으로 분할한 음소와 자동음성분할 및

레이블링 시스템이 분할한 음소가 일치하는 경우의 음소경계를 비교하였다. 이 때 일치하는 유사음소의 개수는 총 4153개이다. 각 분할 실험 결과에서 음소경계 검출 결과는 50개의 문장에서 수작업으로 분할한 음소와 자동음성분할 및 레이블링 시스템이 분할한 음소가 일치하는 4153개의 유사음소 개수 중에서 경계 오차가 주어진 오차 범위 안에 해당하는 개수의 비를 백분율로 표현한다. State 3개에 대하여 mixture 3개의 경우에 대해서 실험하였다.

<표 2>에서는 SVF를 이용한 성능향상을 나타낸다. SVF를 특징 파라메터에 포함시켜 훈련한 경우 성능향상이 이루어졌다. 또한, SVF 값들이 음향학적 변화에 대한 정보를 가지고 있기 때문에 Viterbi alignment 과정에서 다음 음소로 전이할 확률을 구하는 과정에 고려될 때 약간의 추가적인 성능향상이 이루어졌다.

<표 2> SVF를 이용한 음소경계 검출 성능

경계오차 범위	26차 특징파라메터	27차 특징파라메터 (SVF 포함)	27차 특징파라메터 +Viterbi(SVF)
≤ 10ms	47.5% (1973)	50.0% (2075)	51.5% (2139)
≤ 20ms	76.9% (3195)	77.9% (3237)	79.1% (3283)
≤ 30ms	90.2% (3746)	90.7% (3767)	91.6% (3803)
≤ 40ms	94.9% (3939)	95.6% (3968)	95.6% (3971)
≤ 50ms	97.0% (4030)	97.2% (4036)	97.5% (4050)

( ) 안은 개수, 총 유사음소 개수 = 4153 개

후처리 과정에 SVF를 고려한 경우에는 성능향상에 별로 영향을 미치지 못했다. 일반적으로는 자동으로 자른 경계보다는 SVF peak가 평균적으로 정확한 경계에 가깝게 위치하지만, 후처리 과정이 효과적인 경우와 효과적이지 못한 경우가 모두 발생하여 결국에는 성능향상에 별로 기여하지 못한 것으로 판단된다.

<표 3>은 묵음모델을 수정하고 후처리 단계에 에너지 변화량을 이용하여 추가적인 성능향상을 나타낸 결과를 보여준다. 묵음에서 다음 음소로 전이하는 경계의 경우, 에너지량이 뚜렷하게 증가하는 것이 관찰되므로 후처리 과정을 통해 이러한 경계 주변에서 에너지 변화량이 가장 큰 위치로 경계를 수정함으로써 상당한 성능향상을 얻을 수 있었다.

<표 3> 후처리 단계에 에너지 변화량을 이용한 음소경계 검출 성능

경계오차 범위	27차 특징파라메터 +Viterbi(SVF)	묵음모델수정	후처리 단계에 에너지변화량이용
≤ 10ms	51.5% (2139)	52.0% (2160)	58.9% (2444)
≤ 20ms	79.1% (3283)	79.9% (3320)	83.1% (3453)
≤ 30ms	91.6% (3803)	91.9% (3818)	92.0% (3820)
≤ 40ms	95.6% (3971)	95.8% (3980)	95.8% (3979)
≤ 50ms	97.5% (4050)	97.7% (4057)	97.7% (4056)

( ) 안은 개수, 총 유사음소 개수 = 4153 개

실험결과를 정리해 보면, SVF를 적용한 방법 중에서는 SVF를 특징 파라메터에 추가한 후 음소 경계 검출과정에서 제약조건으로 사용한 경우의 성능이 가장 우수했다. 그리고 후처리 과정을 통해 에너지 변화량을 고려했을 때 추가적인 성능향상이 이루어졌다. 실험내용 중 가장 우수한 결과는 표 3에서 보는 바와 같이 20ms이내에서는 83.1%가 포함되었으며, 40ms이내에는 95.8%가 포함되었다.

그 다음으로 음소그룹별 전이에서의 성능을 비교하기 위하여, 본 논문에서는 음소 레이블들을 표 4과 같이 4개의 그룹으로 묶어서 각 음소그룹별 경계추정 성능을 분석해 보았다(허웅, 1983)(Ladefoged, 1993). 여기서 공명음은 모음과 유성자음들로 구성된다.

<표 4> 음소 그룹목록

그룹	그룹 식별자	음소표기
공명음	0	ㄴ, ㅁ, ㅇ, ㄹ, ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ, ㅐ, ㅔ, ㅢ, ㅒ, ㅖ, ㅚ, ㅙ, ㅞ, ㅘ, ㅟ, ㅝ
파열음	1	ㄱ, ㄲ, ㅋ, ㄷ, ㄸ, ㅌ, ㅂ, ㅃ, ㅍ
마찰음 파찰음	2	ㅈ, ㅉ, ㅊ, ㅅ, ㅆ, ㅎ
묵음	3	묵음

<표 5>에 각 음소그룹별 전이에서의 성능결과를 나타내었다. 각 음소그룹 전이별로 발생한 음소의 개수, 평균과 표준편차를 나타내었으며, 오차범위는 10ms이내, 20ms이내, 그리고 40 ms이내로 정리하였다. 이들 표에서 예를 들어 음소경계 종류가

12라는 것은 그룹 1의 음소군으로부터 그룹 2의 음소군으로 전이되는 경우를 의미한다. 이 때, 오차는 자동음성분할 및 레이블링 시스템의 음소 분할한 경계를 수작업으로 음소 분할한 경계로 뺀 것이다. 따라서 평균이 음의 값이면 수작업으로 음소 분할한 경계보다 자동음성분할 및 레이블링 시스템은 평균적으로 왼쪽에서 음소의 경계를 분할한 것을 나타내고, 반대로 평균이 양의 값이면 수작업으로 음소 분할한 경계보다 자동음성분할 및 레이블링 시스템은 평균적으로 오른쪽에서 음소 경계를 분할한 것을 나타낸다.

<표 5>로부터 여러 가지 음소그룹 경계 중에서 파열음에서 공명음으로 전이할 때에 상대적으로 높은 성능을 나타내는 것을 볼 수 있다. 그리고, 파열음에서 파열음, 마찰음, 파찰음 및 묵음으로 전이할 때에 음소분할 성능이 크게 떨어지는 결과를 보이고 있다. 이 때, 앞쪽의 파열음은 불파음화 된 종성으로서 입술이 닫히면서 음의 자질이 표현된다. 그리고 나서, 다음 음성을 발음하는 사이에 숨소리나 입술소리 등의 화자잡음들이 끼어 드는 경우가 빈번히 발생되었다. 수작업으로 음소분할을 하는 경우에는 입술이 닫히는 순간을 경계로 정확하게 자를 수 있다. 이에 반하여 자동음성분할 및 레이블링의 경우에는 잡음구간까지 음성으로 잘못 알고 경계를 자르기 때문에 이와 같은 성능저하가 생기는 것으로 관찰되었다. 그러나, 이러한 경우가 전체 음성 데이터베이스에서 차지하는 비중은 크지 않아서, 전체적인 성능에는 큰 영향을 미치지 않았다.

<표 5> 음소그룹 전이시의 음소분할 성능
(27차 특징파라메터+Viterbi(SVF)+SIL 모델수정의 경우)

음소경계 종류	개수	평균 (ms)	표준편차 (ms)	10ms이내 (%)	20ms이내 (%)	40ms이내 (%)
00	1437	3.00	29.24	53.8	79.1	94.0
01	424	-3.95	16.31	46.0	79.0	97.9
02	302	-7.87	19.96	46.4	80.1	94.7
03	358	3.69	14.43	55.9	86.0	99.2
10	619	1.06	10.33	72.9	95.2	99.4
11	14	65.53	70.80	0.0	0.0	21.4
12	21	46.87	51.93	9.5	19.1	28.6
13	54	29.46	38.83	9.3	33.3	81.5
20	439	6.78	16.17	60.1	84.7	96.6
30	59	-1.09	10.82	71.2	94.9	100.0
31	306	-15.15	18.60	18.6	59.2	100.0
32	120	-11.41	18.88	25.8	65.8	97.5
전체	4153	0.63	22.39	52.0	79.9	95.8

또한 묵음에서 파열음, 마찰음 및 파찰음 등으로 전이하는 과정에서도 상당한 오류가 발생했는데, 이러한 음소경계의 경우 에너지 변화량이 뚜렷이 관찰된다. 따라서

표 3에서 이미 언급한 바와 같이 이러한 경우에는 후처리 과정을 통해 자동분할 된 경계 주변에서 에너지 변화량이 가장 큰 위치로 경계를 수정해 줌으로써 성능향상을 꾀할 수 있었다. 묵음에서 다른 음소그룹(공명음, 파열음, 마찰음 및 파찰음)으로 전이하는 경우에 대해 에너지 변화량을 이용한 후처리를 적용한 결과가 표 6에 나타나 있다. 표 5와 표 6을 비교해 보면 이들 경우에 대해 성능이 향상된 것을 볼 수 있다. 에너지 변화량을 이용한 후처리를 여타의 음소그룹 전이에도 적용해본 결과 일부는 개선되고 일부는 더 악화되는 등 일관성 있는 성능향상이 이루어지지 않아서, 본 논문에서는 묵음에서 다른 음소그룹으로 전이하는 경우에만 후처리를 적용하였다.

<표 6> 음소그룹 전이시의 음소분할 성능
(27차 특징파라메터+Viterbi(SVF)+에너지를 이용한 후처리를 한 경우)

음소경계 종류	개수	평균 (ms)	표준편차 (ms)	10ms이내 (%)	20ms이내 (%)	40ms이내 (%)
00	1437	3.00	29.24	53.8	79.1	94.0
01	424	-3.95	16.31	46.0	79.0	97.9
02	302	-7.87	19.96	46.4	80.1	94.7
03	358	3.69	14.43	55.9	86.0	99.2
10	619	1.06	10.33	72.9	95.2	99.4
11	14	65.53	70.80	0.0	0.0	21.4
12	21	46.87	51.93	9.5	19.1	28.6
13	54	29.46	38.83	9.3	33.3	81.5
20	439	6.78	16.17	60.1	84.7	96.6
30	59	-0.25	7.99	**83.1**	**96.6**	100.0
31	306	-3.16	8.40	**87.9**	**94.1**	100.0
32	120	-2.07	13.76	**80.0**	**86.7**	96.7
전체	4153	1.80	21.80	**58.9**	**83.1**	95.8

6. 결론

본 논문에서는 음성의 변화특성의 정보를 잘 표현하는 SVF를 이용하여 음성 특징 파라메터에 첨가하고, SVF의 peak값을 구하여 Viterbi alignment 과정에 제약을 가

함으로 기존의 음성분할 시스템의 성능을 향상시켰다. 또한 후처리 과정에 에너지 변화량을 이용함으로 추가적인 성능향상을 얻을 수 있었다.

본 시스템은 음소단위로 분할된 방대한 양의 데이터베이스를 구축하는데 기존의 수작업에 의한 방식에 비해서 많은 시간과 비용을 절감할 수 있으며, 이러한 환경을 통해서 앞으로 방대한 양의 음성 데이터베이스가 구축된다면 정교한 음소 모델 구현에 큰 역할을 할 것으로 판단된다.

<참고문헌>

김봉완, 이용주 외(1996), 공동이용을 위한 단어음성DB의 구축 및 PBS설계에 관한 검토, 「제 13회 음성통신 및 신호처리 워크샵 논문집」, 256-261.

성종모, 김형순(1997), 자동 음소분할 및 레이블링 시스템의 구현, 「한국음향학회지」, 제16권 5호, 50-59.

허 웅(1983), 「국어학 : 우리말의 오늘 어제」, 샘 문화사.

A. Ljolie and M. D. Riley(1991), "Automatic segmentation and labeling of speech", in *Proc. ICASSP*, 473-476.

B. Eisen, H. Tillmann and C. Draxler(1992), Consistency of judgements in manual labeling of phonetic segments: the distribution between clear and unclear cases, in *Proc. ICSLP*, 871-874.

B. Wheatley, G. Doddington, C. Hemphill and J. Godfrey(1992), Robust automatic time alignment of orthographic transcriptions with unconstrained speech, in *Proc. ICASSP*, I-553-556.

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

G. Flammia, P. Dalsgarrd, O. Andersen and B. Lindberg(1995), Segment based variable frame rate speech analysis and recognition using a spectral variation function, in *Proc. ESCA Eurospeech*, 511-514.

H. C. Leung and V. Zue(1984), A procedure for automatic alignment of phonetic transcriptions with continuous speech, in *Proc. ICASSP*, 429-432.

P. Ladefoged(1993), *A Course In Phonetics,* 3rd Edition, Harcourt Brace College Publishers.

The Aligner(1994), *A system for automatic time alignment of English text and speech,* Entropic Research Laboratory, Inc.

접수일자: 1998년 10월 14일
게재결정: 1998년 10월 30일

▶ 홍성태(Seong Tae Hong)
주소: 부산광역시 금정구 장전동 산 30
소속: 부산대학교 공과대학 전자공학과

▶ 김제우(Je Woo Kim)
주소: 부산광역시 금정구 장전동 산 30
소속: 부산대학교 공과대학 전자공학과

▶ 김형순
주소: 부산광역시 금정구 장전동 산 30
소속: 부산대학교 공과대학 전자공학과
전화: 051) 510-2452
e-mail: kimhs@hyowon.cc.pusan.ac.kr