유/무성/묵음 정보를 이용한 TTS용 자동음소분할기 성능향상*

김민제(울산대), 이정철(울산대), 김종진(ETRI)

<차 례>

<Abstract>

Improvement of an Automatic Segmentation for TTS Using Voiced/Unvoiced/Silence Information

Min-Je Kim, Jung-Chul Lee, Jong-jin Kim

For a large corpus of time-aligned data, HMM based approaches are most widely used for automatic segmentation, providing a consistent and accurate phone labeling scheme. There are two methods for training in HMM. Flat starting method has a property that human interference is minimized but it has low accuracy. Bootstrap method has a high accuracy, but it has a defect that manual segmentation is required

In this paper, a new algorithm is proposed to minimize manual work and to improve the performance of automatic segmentation. At first phase, voiced, unvoiced and silence classification is performed for each speech data frame. At second phase, the phoneme sequence is aligned dynamically to the voiced/unvoiced/silence sequence according to the acoustic phonetic rules. Finally, using these segmented speech data as a bootstrap, phoneme model parameters based on HMM are trained.

For the performance test, hand labeled ETRI speech DB was used. The experiment results showed that our algorithm achieved 10% improvement of segmentation accuracy within 20 ms tolerable error range. Especially for the unvoiced consonants, it showed 30% improvement.

* Keywords: Automatic segmentation, HMM, Voiced/unvoiced/silence information.

* 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다.

1. 서 론

최근의 TTS 합성 방식에서는 음성 데이터 용량뿐만 아니라 음성 데이터의 분절 표기 정확도가 합성음의 음질을 결정짓는 중요한 요소이다. 따라서 정확한 대용량의 음소분할 및 레이블링이 이루어진 음성 데이터베이스의 구축은 필수적이다. 하지만 음성 데이터베이스 구축 과정은 대부분 수작업에 의해 수행되며 이로 인한 많은 문제점들이 발생하게 되어 자동음소분할 기술개발의 필요성이 대두되었다[1].

자동음소분할은 높은 일관성과 정확성을 확보하고 있는 모델 매칭 방법의 HMM 방식이 널리 사용되고 있으며, 자동음소분할 결과의 성능을 향상시키기 위하여 음성·음향학적 지식을 결합한 HMM기반 자동 음소 분할기, 혹은 후처리기 개발이 진행되고 있다[2][3][4].

후처리 기술에는 신경망을 이용하는 방법, 스펙트럼의 기울기를 이용하여 경계를 수정하는 방법, 음성파형의 에너지를 이용하여 묵음과 음성의 경계를 수정하는 방법등이 사용되고 있다[5][6][7][8][9]. 하지만 이런 후처리 기술들은 특정 화자에서는 잘 적용되나, 다른 화자에 대해서는 잘 적용되지 않는 문제점을 가지고 있다.

또한, HMM기반 자동음소분할기는 bootstrap을 사용하지 않을 경우(flat) 훈련문장들에 대한 global mean, variance 구하여 그것으로 초기 각 음소에 대해 동일한 HMM 모델을 구성한다. 즉, 초기에 각 문장들을 음소열의 수만큼 균등하게 나누어 훈련시킨다는 의미이다. 그 결과 초성의 경우 전체 자동음소분할 정확률에 비해 10%이상 낮게 나타남을 보였고, 특히 무성음에서 낮은 정확률을 보였다.

따라서 본 논문에서는 flat start로 자동음소분할시 오류가 가장 큰 무성음의 정확률을 향상시킬 수 있으며, 다른 화자에 대해 공통적으로 적용할 수 있는 방법 제시를 목적으로 하였다. 이를 위해 음성파형에서 유/무성/묵음 구간을 검출하고, 검출된 결과를 음소열과 유/무성/묵음열을 음소의 속성을 이용하여 대응시킴으로써 각 음소에 시간정보를 부여하여 HMM기반 자동음소분할기의 bootstrap으로 이용함으로써, 자동음소분할기의 성능을 향상시키기 위한 방안을 제시하였다.

본 논문의 구성은 2장에서는 HMM 기반 자동음소분할 Baseline 시스템 구성에 대해서 서술하고, 3장에서는 유/무성음 구간에 대한 정보를 사용하여 자동 음소분할 성능 향상 방법을 제안하였다. 4장에서는 실험결과를 보이고, 마지막으로 결론을 맺도록 한다.

<표 2> 특징 파라미터에 따른 자동음소분할결과

	Set 1	Set 2	Set 3
정확률(%)	67.45	64.32	61.67

본 논문에서는 위의 결과를 바탕으로 MFCC를 음성 특징 파라미터로 사용하였다.

2.3 음성 분석창 크기 선정

음소자동분할 시스템에서는 정교한 음소분할을 위해서 음소경계 검출의 정밀도가 10ms 수준인 것은 바람직하지 않으며 정밀도가 5ms를 넘지 않아야 좋을 것으로 판단되어져 왔으며, 미국의 TIMIT 음성 데이터베이스의 경우에는 2.5ms 시간단위를 사용한 것으로 알려져 있다[3]. 따라서 본 논문에서는 일반적으로 사용되고 있는 5ms을 선정하였다.

2.4 가우시안 밀도 함수

확률밀도 모델링 방법 중의 하나인 GMM은 밀도함수의 개수를 늘리면 밀도함수는 근사화 할 수 있으며 특히 EM 알고리즘에 의하여 빠르고 간편하게 파라미터 추정할 수 있다는 장점이 있다. 하지만 그 만큼의 모델을 훈련시키는 시간이 늘어나지만 시간의 증가에 비하여 정확률이 크게 증가 되지 않는 단점이 있어 가우시안 밀도 함수의 수는 실험에 의해 결정하는 것이 옳다. <표 3>은 가우시안 밀도 함수 수에 따른 자동음소 분할 실험결과를 보여주고 있다. 실험결과 가우시안 밀도 함수 수를 늘여도 정확률은 크게 향상 않는 것을 확인 하였다.

따라서 본 논문에서는 밀도 함수 수는 실험결과 최상의 정확률을 보이는 3개로 선정 하였다.

<표 3> 가우시안 밀도 함수 수에 따른 자동음소분할결과

	1	2	3	4	5	6
정확률(%)	67.34	67.35	67.45	67.41	67.41	67.32

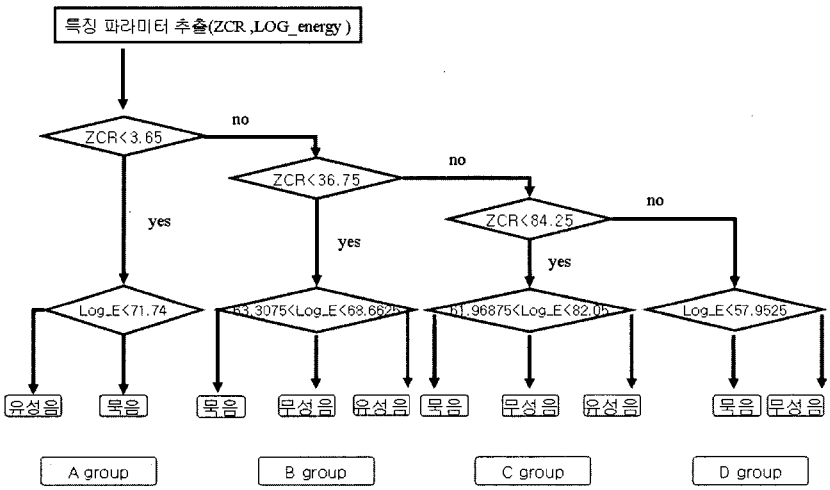
3. 유성/무성/묵음 정보를 이용한 자동음소분할 성능 향상

본 논문에서의 성능개선 방법은 음성파형에서 유/무성음, 묵음 구간을 검출한 후, 음소열을 유/무성음, 묵음 구간에 음소의 속성을 이용하여 대응시켜 각 음소에 시간정보를 할당하여 자동음소분할기의 bootstrap으로 이용하였다.

3.1 유성/무성/묵음 구간 검출

본 논문에서는 유/무성/묵음 검출 방법 중 가장 간단한 에너지와 영교차율을 이용하였다. 에너지가 상대적으로 크게 나타나면 유성음이라 간주하고 상대적으로 작게 나타나면 무성음이라 간주하면 된다. 판별의 정확성을 위하여 영교차율을 함께 사용하였다. 영교차율이 상대적으로 높으면 무성음으로 간주하고 상대적으로 낮으면 유성음으로 간주하면 된다[11].

유/무성음, 묵음 구간에 대한 각각의 에너지와 영교차율을 구하고, 각 구간에서의 히스토그램을 이용하여 각각에 대한 경계값을 정의하였다. 유/무성 구간 정의는 입력 파형에서 프레임별로 로그에너지와 영교차율을 구하고 정의된 경계값을 이용하여 <그림 2>와 같은 분류과정을 통하여 구간 정보를 구하였다. <그림 3>은 음성파일의 유/무성음 판정 결과를 보여주고 있다.



<그림 2> 유성음/무성음/묵음 결정 방법

<표 4>는 중성 'ㅏ'에 대한 속성의 예를 보여준다.

<표 4> 중성'ㅏ'에 대한 속성의 예

	유/무성/묵음열	확률값
단어의 중간	유성음	0.95
	묵음+유성음	0.05
단어의 끝	유성음	0.4
	유성음+무성음	0.55
	유성음+무성음+유성음+무성음	0.05

유성/무성/묵음 구간 검출결과와 수작업으로 음소분할된 결과를 이용하여 각 음소가 가질 수 있는 속성들은 정의되었으며, 확률값이 너무 낮은 속성에 대해서는 검색 시간을 고려하여 제외하였다.

여기서 속성은 50문장의 수작업으로 음소분할된 결과를 사용하여 정의하였다. 하지만 다른 DB에 대해서 음소분할시 수작업으로 음소분할된 결과를 이용한 추가적인 속성의 추출 없이 정의된 속성을 이용하여 자동음소분할이 가능하게 된다.

어떤 음소의 속성은 단어의 끝일 경우와 그렇지 않을 경우, 앞에 음소가 단어의 끝일 경우로 나누어 규칙을 추출하였다. 이는 동일한 음소라도 다른 유성/무성/묵음열을 가지기 때문이다. 예를 들어 음소 'ㅎ'의 경우 단어의 시작에서는 무성음과 묵음으로 표현 될 수 있지만 단어의 중간에서는 무성음의 유성화로 인하여 유성음으로 나타나기 때문에 앞의 음소를 고려하여야 하며, 'ㅏ'의 경우는 단어의 중간에서는 유성음으로 나타나지만 단어의 끝에서는 단어와 단어 사이의 짧은 휴지(short pause)로 인하여 유성음과 무성/묵음이 연결되어 존재하게 된다.

본 논문에서 음소열을 유/무성/묵음 구간에 할당을 위한 경로 탐색 방법은 다음과 같다. 여기서 N은 음소의 수, M은 유/무성/묵음 열의 수를 의미한다.

* 현재음소 sim[i], 현재 유/무성/묵음 UV[j]

$0 \le i \le N,\ 0 \le j \le M, 1 \le k \le N$

step 1 : i, j 값을 0으로 k는 1로 초기화

step 2 : sim[i] 음소의 속성은 유/무성/묵음열 고려하여 이전에 적용된 속성을 제외하고 선택 가능한 속성 중에서 확률이 가장 높은 속성을 적용하고, i를 1증가 시키고, j를 속성의 유/무성/묵음열에 맞게 증가.

step 3 : 만약 i<N이고 j<M이면 step 2 수행.

만약 i=N, j=M이면 현재 경로의 확률값이 이전의 후보 경로의 확률값보다 크면 현재의 경로를 후보 경로로 선택하고 경로와 확률값을 저장.

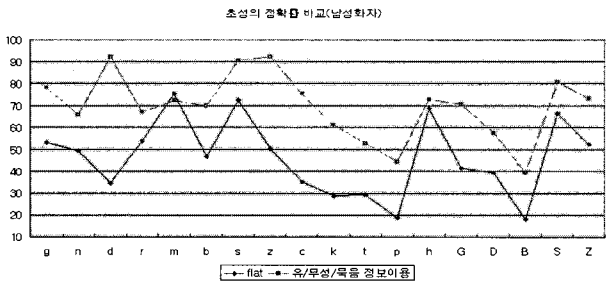
만약 i=N이고 j<M 또는 i<N 이고 j=M 이면 경우에는 현재의 경로는

<표 9> 자동음소분할결과 (정확률%)

		전체	초성	초성 (무성음)	중성	종성	묵음
여성 화자	(가)	67.65	52.37	46.44	83.67	66.68	44.75
	(나)	78.43	78.1	78.48	79.74	86.74	46.98
	(다)	90.81	90.65	90.83	93.85	90.01	69.35
남성 화자	(가)	63.25	51.54	49.81	74.26	67.43	41.33
	(나)	74.25	75.95	79.00	74.37	81.65	44.83
	(다)	89.56	88.47	90.83	92.84	92.27	59.50

자동음소분할시 유/무성/묵음정보를 사용한 경우 flat start한 경우보다 전체정확률이 10%이상 향상되었다. 그리고 초성의 경우 25%정도의 정확률이 향상되었으며, 초성에서 무성음소의 경우 30%이상 정확률이 향상됨을 알 수 있었다. 그리고 종성의 경우에도 15%이상의 성능이 개선되었다. 하지만 중성의 경우 정확률의 향상이 없었다. 이는 중성의 경우 대부분 유성음으로 나타나게 되고 한 유성음 구간에는 여러 음소들이 존재하기 때문에 환경에 따라 각 음소들의 시간할당을 다르게 해야 하지만, 본 논문에서는 이에 대한 문제는 고려하지 않았기 때문이다.

<그림 6>은 남성화자의 경우 초성 각각에 대한 정확률을 나타낸다. 초성의 경우 25%정도의 정확률이 향상되고, 초성의 'ㅅ', 'ㅈ'과 같이 무성음의 특징이 크게 나타나는 음소의 경우 90%이상의 높은 정확률을 보였다. 그리고 다른 무성음소의 경우에도 flat start한 경우 낮은 정확률을 보였지만, 본 논문에서 제시한 방법을 사용한 결과 30~40%이상의 정확률이 향상되어 초성 무성음의 경우 flat start한 경우한 경우보다 30%이상 성능이 향상됨을 보였다. 하지만 초성의 'ㅁ'의 경우에는 flat start한 경우보다 다소 낮게 나왔는데 이는 초성 'ㅁ'의 경우 유성음의 특징이 크기 때문이다.



<그림 6> 초성의 정확률 비교(남성화자)

결과에서 보듯이 본 논문에서 제시한 방법은 자음의 자동음소분할의 성능향상에 효과적이었으며, 특히 무성음의 특징을 가지는 음소에 대해 더 효과적이라 할 수 있다.

5. 결 론

본 논문에서는 수작업으로 음소분할된 데이터 없이 flat start로 자동음소분한 경우 무성음소의 분할 정확률이 낮은 문제점을 해결하기 위하여 다른 DB에 대해서도 적용이 가능한 유/무성음 정보와 음소의 속성을 매칭시킨 결과를 자동음소분할기의 boostrap으로 사용하는 방법을 이용하여 HMM기반 자동음소분할기의 성능을 향상시키는 방안을 제시하였다. 그 결과 자음(무성음)에 대해 30%이상의 성능이 향상됨을 확인할 수 있었고, 전체적으로 10%이상 정확률이 향상되었다. 이는 무성음의 정보가 무성음소의 자동음소분할의 성능향상에 효과적이라 할 수 있다.

본 논문에서의 유/무성음 구간 검출은 수작업으로 분할된 음소의 경계와는 차이가 있으며, 유/무성음 두가지 분류로 나누는 것은 매우 기본적이다. 따라서 성능을 더욱 향상시키기 위해서는 분류의 종류를 증가시켜 수작업된 결과와 유사한 bootstrap의 생성을 위한 연구가 필요하다.

또한, 유성음소의 성능 향상을 위하여 유성음 구간내에 여러 음소들이 존재하여 음소별로 시간을 할당하지만, 실제 음소의 지속시간 길이는 음소 자체의 성질뿐 만 아니라 주변의 음소환경, 한 단어내의 음소 개수, 단어 내에서의 음소의 위치, 강세 여부 등 다양한 요소에 의해 영향을 받기 때문이다. 따라서, 환경에 따라 음소의 지속시간을 결정할 수 있는 방법이 필요하다.

참 고 문 헌

[1] 박순철, 김봉완, 이용주, "문맥종속 반음소단위에 의한 음운 자동 레이블링 시스템의 성능 개선", *말소리*, 제37호, pp. 23-48, 1999.

[2] 이기승, 김정수, "문자-음성 합성기의 데이터 베이스를 위한 문맥 적응 음소 분할", *한국음향학회지*, 제22권, 제2호, pp. 135-144, 2003.

[3] 박창목, 왕지남, "음소 음향학적 변화 정보를 이용한 한국어 음성신호의 자동 음소 분할", *한국음향학회지*, 제20권, 제8호, pp. 24-30, 2001.

[4] F. Brugnara, D. Falavigna and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication*, vol. 12, no. 4, pp. 357-31, 1993.

[5] 박혜영, 김형순, "자동 음성 분할을 위한 음향 모델 에너지 기반 후처리", *말소리*, 제43