유성음과 무성음의 경계를 이용한 연속 음성의 세그먼테이션

유 강 주† · 신 옥 근††

요 약

본 논문에서는 한국어의 음절 구조에 관한 지식과 유/무성음의 경계점을 음소 경계 추출에 이용하여 맹목 세그먼테이션의 성능을 향상 시킬 수 있음을 보인다. 본 고의 음소 경계 추출 방법은 후보 음소 경계를 추출하는 과정, 유/무성음의 경계점을 추출하는 과정, 그리고 최종 음소 경계를 선택하는 과정으로 구성된다. 후보 음소 경계는 이웃하는 두 클러스터의 유사도에 기초한 클러스터링 방법을 이용하여 추출하며, 유사도는 두 클러스터 사이의 특징 벡터 분포에 대한 확률 밀도의 비로 표현한다. 유/무성음의 경계점은 음성 신호의 스펙트럼으로부터 0~400Hz 사이의 전력 밀도를 구한다음, 이들의 변화량이 임계치를 초과하는 점들을 선택하여 추출한다. 최종 음소 경계는 후보 음소 경계와 유/무성음의 경계 점을 토대로 무성음 구간의 음소 수를 제한하여 추출한다. 본 고에서 제안한 방법을 이용하여 실험한 결과, 누락 오류가 약 10%일 때 과추출 오류가 맹목 세그먼테션에 비해 약 40%정도 감소되어 제안한 방법의 유용성을 확인할 수 있었다.

Segmentation of Continuous Korean Speech Based on Boundaries of Voiced and Unvoiced Sounds

Gang-Ju You† · Ok-Keun Shin††

ABSTRACT

In this paper, we show that one can enhance the performance of blind segmentation of phoneme boundaries by adopting the knowledge of Korean syllabic structure and the regions of voiced/unvoiced sounds. The proposed method consists of three processes : the process to extract candidate phoneme boundaries, the process to detect boundaries of voiced/unvoiced sounds, and the process to select final phoneme boundaries. The candidate phoneme boundaries are extracted by clustering method based on similarity between two adjacent clusters. The employed similarity measure in this process is the ratio of the probability density of adjacent clusters. To detect the boundaries of voiced/unvoiced sounds, we first compute the power density spectrum of speech signal in 0~400Hz frequency band. Then the points where this power density spectrum variation is greater than the threshold are chosen as the boundaries of voiced/unvoiced sounds. The final phoneme boundaries consist of all the candidate phoneme boundaries in voiced region and limited number of candidate phoneme boundaries in unvoiced region. The experimental result showed about 40% decrease of insertion rate compared to the blind segmentation method we adopted.

1. 서 론

최근에 많이 연구되고 있는 연속 음성의 인식(continuous speech recognition)에서는 세그먼트 단위(segment unit)의 인식 방법이 많이 사용되고 있으며, 인식 단위로는 음소(phone), 음절(syllable), 유사음절(semi-syllable) 또는 단어(word) 등이 있고, 이들 중 음소단위의 인식이 가장 많이 사용된다. 음소단위의 인식 방

† 정 회 원 : 한국해양대학교 대학원 제어계측공학과
†† 정 회 원 : 한국해양대학교 자동화정보공학부 교수
논문접수 : 2000년 4월 10일, 심사완료 : 2000년 6월 27일

법은 대용량의 어휘 인식, 특히 일상의 대화체 발화(utterance)를 인식하는 자동 음성 인식 시스템(automatic speech recognition system)의 구현이 비교적 용이해서 많이 이용되고 있으며, 이런 시스템에서는 음성 신호로부터 음소의 경계를 정확하게 추출하는 것(speech segmentation)이 아주 중요하다.

음성 신호로부터 음소의 경계를 추출하는 방법에는 발화의 전사(utterance transcription)등과 같이 발화에 대한 직접적인 지식을 이용하여 경계를 추출하는 방법, 음성에 대한 아무런 사전 지식 없이 경계를 추출하는 맹목 세그먼테이션(blind segmentation) 방법, 그리고 음소 경계에 대한 경험 데이터, 혹은 간접적인 지식을 이용하는 방법 등이 있다. 이들 방법 중Svendsen[1]등은 직접적인 지식인 발화의 전사와 벡터 양자화(vector quantization)기법을 이용하여 과추출 오류 75%, 누락 오류 1%의 결과를 얻었으며, Suh[2]등은 음소경계에 대한 경험 데이터를 NN(neural network)으로 구현함으로써 과추출 오류가 3.4%일 때 누락 오류 8%의 결과를 얻었고, Pellon[3-4]등은 음소경계에 대한 경험 데이터를 HMM(hidden markov model)으로 구현하여 과추출 오류가 27.2%일 때 누락 오류 1.6%의 결과를 얻었다. 맹목 세그먼테이션의 경우에는 Eberman등이 제안한 방법[5], Andre-Obrecht가 제안한 방법[6-7] 등이 있으며, Eberman등은 특징벡터의 분포를 나타내는 확률모델과 클러스터링 방법을 이용하여 과추출 오류 100%에 누락 오류 6.2%의 결과를, Andre-Obrecht는 음성샘플의 변화를 나타내는 확률모델을 이용하여 과추출 오류 120%에 누락 오류 2.8%의 결과를 얻었다. 이들 중에서 발화에 대한 직접적인 지식을 이용하는 음소 경계추출 방법은 성능이 우수하지만 미지의 발화에 대한 음성인식에는 사용할 수 없으며, 맹목 세그먼테이션은 상대적으로 성능이 낮은 단점이 있다.

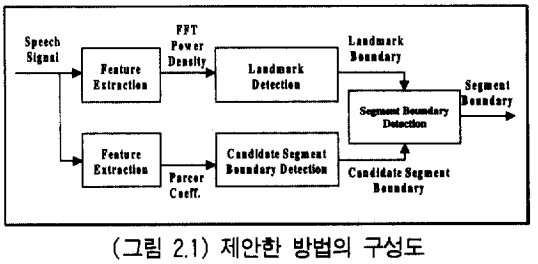
이런 음소단위의 경계추출 방법과는 달리 Liu는 음성 신호의 특이 점(landmark)을 추출하여 음성인식에 이용하는 방법을 제안하였으며, 음성신호를 여섯 개의 주파수 대역으로 분류한 다음 각 대역의 전력 밀도로부터 성문 진동(glottal vibration), 비강 진동(sonorant vibration), 파열음(burst) 등의 시작 점과 끝점을 추출한 결과, 약 90%정도의 특이 점을 올바르게 추출할 수 있었다[8].

본 논문에서는 Liu의 특이 점 추출 방법과 한국어의 음절 구조에 대한 지식을 이용하여 음소의 경계를 추출하는 방법을 제안한다. 한국어의 모든 음절은 (초성)-(중성)-(종성)으로 구성되며 영어, 불어 등의 유럽어에 비해 간단한 음절구조를 갖는다. 한국어의 음절구성 요소 중 중성인 모음은 모두 성문의 진동이 수반되는 유성음이므로, 모든 유성음 사이에는 최대 세 개까지의 음소(종성인 무성 자음-묵음 또는 잡음-초성인 무성 자음)가 나타날 수 있다. 본 고에서는 이러한 유성음과 무성음에 대한 지식을 음소 경계추출에 이용하여 맹목 세그먼테이션의 성능을 높일 수 있음을 보인다. 이를 위해 먼저 음성신호로부터 Eberman 등이 제안한 맹목 세그먼테이션 방법을 이용하여 음소의 후보 경계를 추출하고, 이와 동시에 유성음 구간과 무성음 구간을 구별할 수 있는 음성의 특이 점들을 추출한 다음, 특이 점과 음소의 후보 경계를 토대로 무성음 구간에서의 음소 수를 제한한다. 실험 결과 누락 오류가 약 10%일때 과추출 오류가 맹목 세그먼테이션에 비해 약 40%정도 감소됨을 확인할 수 있었다.

본 논문의 II장에서는 제안한 음소 경계 추출 방법을, III장에서는 제안한 방법의 검증을 위한 실험 결과를 서술한 다음, IV장에서 결론을 맺는다.

2. 특이 점을 이용한 음소 경계의 추출

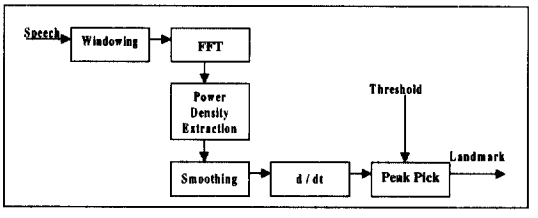
본 논문에서 제안하는 음소의 경계 추출방법은 (그림 2.1)과 같이 구성된다. 이 그림의 'Landmark Detection' 블록은 음성 신호의 전력 밀도(power density)를 토대로 유성음 구간과 무성음 구간을 추출하여, 유성음의 시작 점과 끝 점을 추출해 낸다. 'Candidate Segment Boundary Detection' 블록은 특징벡터의 분포에 대한 확률 모델과 클러스터링 방법[5]을 이용하여 음소의 후보 경계를 추출한다. 'Segment Boundary Detection' 블록은 한국어의 음절 구조에 관한 지식을 앞의 두 과정에서 추출한 특이 점과 음소의 후보 경계에 적용하여 최종적인 음소의 경계를 추출한다.



(그림 2.1) 제안한 방법의 구성도

2.1 특이 점 추출

본 논문에서는 Liu가 제안한 여섯 개의 주파수 대역[8] 중에서 첫 번째 주파수 대역(0~400Hz)의 전력 밀도를 이용하여 특이 점을 추출하였으며, 알고리즘 구성도는 (그림 2.2)와 같다.

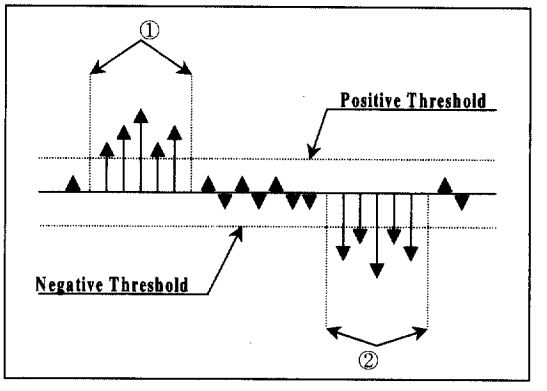


(그림 2.2) 특이 점 추출 방법의 구성도

이 그림의 'Windowing' 블록에서는 프레임 길이를 10ms로 하여 음성 신호를 프레임 단위로 분할한 다음, 각 프레임에 헤밍 윈도우를 적용하고, 'FFT' 블록에서는 윈도윙된 각 프레임에 대해 512-점 FFT를 수행한다. 'Power Density Extraction' 블록에서는 각 프레임에 대해 0~400Hz 사이의 전력 밀도를 구하고, 'Smoothing' 블록에서는 연속된 두 프레임(20ms) 사이의 평균 전력 밀도를 구한다. 'd/dt' 블록에서는 성문의 열림(release)과 막힘(closure)에 대한 변화를 나타내기 위해 차분 간격(dt)을 50ms(5 프레임)로 하여 평균 전력 밀도의 차분을 1프레임(10ms) 단위로 구한다음 데시벨(db) 단위로 변환하는데, 본 고에서는 Pellon등이 제안한 식 (2-1)과 같은 방법으로 평균 전력 밀도의 차분을 구하였다[4].

$$delta\_power_i = \frac{\sum_{j=1}^{\lfloor M/2 \rfloor} j(power_{i+j} - power_{i-j})}{2\sum_{j=1}^{\lfloor M/2 \rfloor} j^2} \quad (2\text{-}1)$$

식 (2-1)에서 $power_i$와 $delta_power_i$는 각각 i번째 프레임의 평균 전력 밀도와 평균 전력 밀도의 차분이고, $\lfloor M/2 \rfloor$은 $M/2$보다 크지않은 정수이며, M은 차분 간격을 프레임 수로 나타낸 것이다. 'Peak Pick' 블록에서는 'd/dt' 블록에서 구한 평균 전력 밀도의 차로부터 국부적인 최대값 또는 최소값을 선택함으로써 특이 점을 추출하는데, 음성이 무성음에서 유성음으로 천이하면 전력 밀도가 증가하므로 평균 전력 밀도차는 양의 값이 되고, 유성음에서 무성음으로 천이하면 전력 밀도가 감소하므로 평균 전력 밀도차는 음의 값이 된다. (그림 2.3)은 평균 전력 밀도차의 이러한 특징을 이용하여 'Peak Pick' 블록에서 특이 점을 추출하는 방법을 나타낸 것이다.



(그림 2.3) 피크 검출방법

(그림 2.3)에서 수직 화살표는 각 프레임의 평균 전력 밀도차를 나타내고 수평으로 그은 점선은 양의 임계치와 음의 임계치를 나타내며, 수직으로 그은 두 쌍의 점선 ①번과 ②번은 각각 양의 임계치와 음의 임계치를 초과하는 평균 전력 밀도차를 가지는 구간을 나타낸다. 이 그림에서 특이 점을 추출하는 과정은 다음과 같다.

1) 양의 임계치와 음의 임계치를 초과하는 구간 ①번과 ②번을 추출한다.
2) 1)에서 추출한 구간이 ①번 구간이면 가장 큰 평균 전력 밀도차를 갖는 프레임을 양의 특이 점으로 하고, ②번 구간이면 가장 작은 평균 전력 밀도차를 갖는 프레임을 음의 특이 점으로 한다.

이 두 과정을 통해서 추출한 특이 점 중 양의 특이 점은 성문이 진동을 시작하는 부분으로 유성음의 시작점에 해당하고, 음의 특이 점은 성문이 진동을 끝내는 부분으로 유성음의 끝점에 해당하며, 이들은 항상 쌍으로 존재해야 한다.

2.2 음소의 후보경계 추출

일련의 음성구간에서 각각의 음성 프레임을 초기 클러스터로 설정한 다음, 이웃하는 두 클러스터 사이의 유사도가 가장 큰 클러스터 쌍을 찾아 그 유사도가 임

계치 보다 크면 하나의 클러스터로 합하는 과정을 반복함으로써 음소의 후보경계를 추출하며, 유사도 측정 방법은 다음과 같다.

연속된 두 클러스터 S_1과 S_2가 주어졌을 때 M_1과 M_2를 각각 클러스터 S_1과 S_2의 확률 모델이라 하고, M_0는 두 클러스터를 하나의 클러스터 S_0로 간주했을 때의 확률 모델이라 하자. 이때 클러스터 S_0에는 N개의 시간영역 음성샘플이 포함되어 있고, 클러스터 S_1과 S_2에는 각각 r개와 $N-r$개의 샘플이 포함되어 있다고 가정한다. 또한 확률 모델 M_1, M_2, M_0가 일반적인 선형예측모델(linear predictive model)이라 하면, 음성신호 $y(n)$은 식 (2-2)과 같이 표현된다.

$$y(n) = \sum_{i=1}^{p} a_i y(n-i) + e(n) \qquad (2\text{-}2)$$

여기서 p는 선형예측 계수의 차수이며, $e(n)$은 예측오차(prediction error)로 평균이 0이고 분산이 v인 가우시안 프로세스(Gaussian process)이다. 선형예측 계수 a_i와 분산 v로 이루어진 확률 모델 $M=(a_i, v)$이 주어졌을 때, 이 모델에서 음성샘플의 시퀀스 $y_1^r=(y(1), y(2), \cdots, y(r))$이 발현할 확률 $L(y_1^r|M)$은

$$L(y_1^r|M) = \prod_{m=1}^{r} P(e(m)|y_{m-p}^{m-1}, M) \qquad (2\text{-}3)$$

로 표현될 수 있다. 앞에서 가정한 확률 모델 M_0, M_1, M_2는 각각 음성샘플 y_1^N, y_1^r, y_{r+1}^N에 대한 것이며, 시퀀스 y_1^r과 y_{r+1}^N 사이의 유사도 D는 식 (2-4)와 같이 정의한다.

$$D = \underset{M_1}{Max}\,\underset{M_2}{Max}\,\underset{M_0}{Max}\frac{L(y_1^N|M_0)}{L(y_1^r|M_1)L(y_{r+1}^N|M_2)} \qquad (2\text{-}4)$$

이 식에서 $L(y_1^r|M_1)$은 확률 모델 M_1에서 시퀀스 y_1^r이 발현할 확률이며, $L(y_{r+1}^N|M_2)$은 확률 모델 M_2에서 시퀀스 y_{r+1}^N이 발현할 확률이고, 그리고 $L(y_1^N|M_0)$은 확률 모델 M_0에서 시퀀스 y_1^N이 발현할 확률이다. 따라서 유사도 D는 음성샘플 y_1^r과 y_{r+1}^N의 특징벡터 분포가 유사하면 1에 가까운 값이 되고, 전혀 다르면 아주 작은 값이 된다.

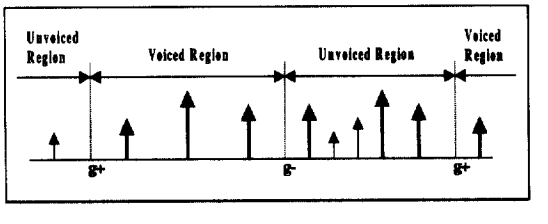
Eberman등은 식 (2-4)의 유사도 D를 Parcor 계수와 이 계수의 차수 p의 함수로 보고 각각의 발현확률 $L(y|M)$을 최대화 시키는 MDL(maximum description length)이라는 방법을 사용하였다[5]. 이 방법은 모든 클러스터와 계수의 차수 p 각각에 대해서 발현확률을 최대화 시켜야 하므로, 계산량이 많은 단점이 있다. 본 연구에서는 모든 클러스터에 대해서 이 값을 최대화 시키는 Parcor 계수의 차수를 실험에 의해서 미리 정해놓고, 음소의 후보경계를 추출하였다. 또한 클러스터링에 사용되는 임계치 th는

$$th = \sigma/\beta + m \qquad (2\text{-}5)$$

을 이용하여 구했다. 여기서 σ와 m은 각각 모든 클러스터 쌍들에 대한 유사도의 표준 편차와 평균이며, β는 가중치 상수로 실험을 통해서 구했다.

2.3 최종 음소 경계추출

본 고에서 제안하는 음소 경계 추출 방법에 사용되는 특이 점은 성문의 진동을 수반하는 유성음 구간과 성문의 진동을 수반하지 않는 무성음 구간을 나타내는 경계 점이며, 음소의 후보 경계는 특징 벡터의 확률밀도와 클러스터링 방법을 이용한 맹목 세그먼테이션의 결과이다. 최종적인 음소의 경계는 앞서 구한 특이점과 음소의 후보 경계에 한글의 음절 구조에 대한 지식을 적용, 무성음 구간의 음소 수를 제한하여 추출한다. 한글의 음절 구조는 (초성)-(중성)-(종성)으로 구성되며 이들 중 중성인 모음은 성문의 진동을 수반하는 유성음이므로, 유성음과 유성음 사이의 무성음 구간에서는 최대 세 개까지의 음소(종성인 무성 자음-잡음 또는 묵음-초성인 무성 자음)가 나타날 수 있다. (그림 2.4)는 후보 음소 경계로부터 특이 점을 이용하여 최종 음소 경계를 추출하는 방법을 나타낸 것이다.



(그림 2.4) 최종 음소 경계 추출

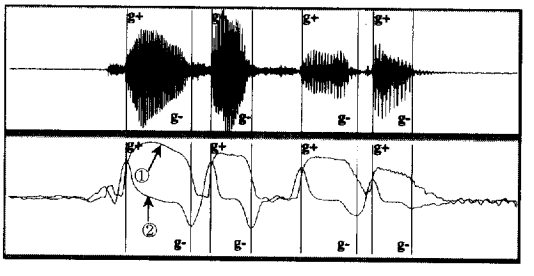
(그림 2.4)에서 수직으로 그린 점선은 특이 점을 나타내고, 수평 화살표는 유성음 구간과 무성음 구간을, "g+"는 유성음의 시작(무성음의 끝) 점을, 그리고 "g-"는 유성음의 끝(무성음의 시작) 점을 나타낸다. 수직 화살표는 이웃하는 두 클러스터 사이의 거리(유사도의 역수)를 나타내는데, 이 거리는 음소의 후보 경계에 의해서 나눠지는 음성신호의 각 구간들을 각각의 클러스터로 간주한 다음, 후보 음소 경계 추출과정에서 사용한 유사도 측정방법을 이용하여 구했다. 또한 수직 화살표 중에서 굵은 화살표는 최종 음소 경계로 선택된 후보 경계를 나타내고, 가는 화살표는 최종 음소 경계로 선택되지 않은 후보 경계를 나타낸다. 유성음 구간에 있는 모든 후보경계는 최종 음소 경계로 간주하고 무성음 구간에 있는 음소의 후보 경계는 연속된 두 클러스터 사이의 거리가 가장 큰 N개(무성음 구간의 음소 개수 제약 조건)를 최종 음소 경계로 간주 한다.

3. 실험 및 고찰

본 논문에서 제안한 음소 경계추출 방법을 검증하기 위해서 ETRI에서 제작한 한국어 음성 데이터 베이스인 POW로부터 음절 수가 6 이상인 240개의 데이터를 선택하여 사용하였으며, 이 데이터의 음소 경계는 스펙트로그램(spectrogram)과 음성 샘플의 변화를 관찰하여 수작업으로 추출하였다. 실험에 사용한 데이터는 주파수 16kHz, 16비트로 샘플링 되어 있으며, 240개의 데이터 중 절반은 음소의 후보경계 추출을 위한 Parcor 계수의 차수와 임계치 계산에 필요한 가중치 상수 β 를 구하는데 사용하였고, 나머지 반은 음소 경계추출에 사용하였다. 또한 음소의 후보경계 추출과정에서는 프리엠퍼시스(preemphasis)와 오버랩(overlap), 그리고 윈도윙(windowing)을 하지 않았고, 특이 점 추출과정에서는 윈도윙은 하고 프리엠퍼시스와 오버랩은 하지 않았으며, 이 두 과정에서 프레임 길이는 10ms로 하였다.

3.1 특이 점 추출

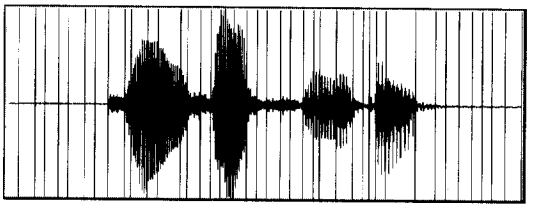
특이 점 추출 실험에서는 각 프레임의 음성 데이터에 대해 512-점 FFT를 수행한 다음 0~400Hz 사이의 전력 밀도를 이용하였다. 또한 이 과정에서는 특이 점을 추출하기 위해 양의 임계치와 음의 임계치를 각각 7db와 -7db로 설정하였으며, 이 임계치들은 식 (2-1)을 이용하여 실험으로 구했다. (그림 3.1)에 /비웃었습니다/ 라는 발화의 음성신호로부터 추출한 특이 점과 0~400Hz 사이의 전력 밀도와 전력 밀도차를 보였다. 이 그림에서 수직 실선은 특이 점을, "g+" 기호는 유성음의 시작점을, "g-" 기호는 유성음의 끝점을 나타낸다. 이 그림에서 ①번과 ②번 실선은 각각 0~400Hz 사이의 전력 밀도와 전력 밀도의 차분이며, 음성이 유성음에서 무성음으로 변할 때에는 전력 밀도가 감소하고, 무성음에서 유성음으로 변할 때에는 전력 밀도가 증가한다는 것을 알 수 있다.



(그림 3.1) 발화 / 비웃었습니다 /의 0~400Hz 대역의 에너지 밀도와 특이 점

3.2 음소의 후보경계 추출

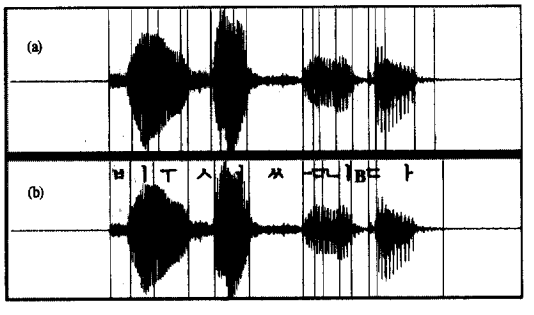
음소의 후보경계는 Parcor 계수를 이용하여 추출하였으며, Parcor 계수의 차수는 실험에 의해서 16차로 설정하였다. 클러스터링 과정에 사용되는 유사도 척도는 확률 밀도에 기초한 MLR(maximum likelihood ratio)방법을 이용하였고, 모든 프레임 각각을 하나의 클러스터로 초기화한 다음 유사한 인접 클러스터들을 합함으로써 클러스터링을 수행하였으며, 클러스터링의 임계치 계산에 필요한 가중치 상수 β는 실험을 통해서 3으로 설정하였다. (그림 3.2)는 / 비웃었습니다 / 라는 발화로부터 추출한 음소의 후보경계를 그린 것이며, 수직으로 그은 실선이 음소의 후보 경계이다.



(그림 3.2) 발화 /비웃었습니다/의 음소 후보 경계

3.3 최종 음소의 경계 추출

최종 음소 경계 추출에서는 무성음 구간을 나타내는 특이 점 사이의 음소 수를 제한하였다. 무성음 구간에서는 음소의 후보경계 추출에 사용된 MLR 방법을 이용하여 연속된 두 후보음소 사이의 거리(유사도의 역수)를 구한 후에, 거리가 가장 큰 N(무성음 구간의 음소 수)개의 후보경계를 최종 음소 경계로 간주한다. (그림 3.3)은 / 비웃었습니다 / 라는 발화에서 추출한 최종 음소 경계와 이 발화의 실제 음소 경계를 그린 것으로, 이 그림에서 (a)와 (b)의 수직 실선이 각각 최종 음소 경계와 실제 음소 경계를 나타낸다. (그림 3.3)과 (그림 3.2)를 비교해 보면 무성음 구간의 음소 수가 많이 줄어들었음을 알 수 있다.

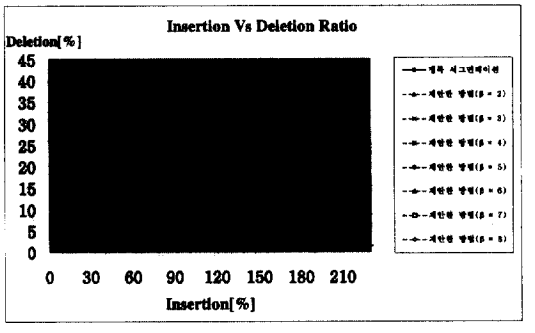


(그림 3.3) 발화 / 비웃었습니다 /의 최종 음소 경계와 실제 음소 경계

3.4 실험 결과

본 논문에서는 한국 전자통신 연구소의 한국어 음성 데이터베이스인 POW로부터 선별한 240개의 데이터(실험 데이터)중 절반을 음소 경계 추출 실험에 사용하였으며, 추출한 음소 경계가 실험 데이터에 있는 음소 경계의 좌/우 20ms내에 있을 경우 올바르게 음소 경계를 추출한 것으로 간주하고, 20ms내에 없을 경우 과추출(Insertion)오류로 간주하였다. 또한 실험 데이터에 있는 음소 경계 중에서 추출한 음소 경계에 포함되지 않은 것은 누락(Deletion) 오류로 간주하였다. 실험 데이터의 음소 경계는 스펙트로그램과 음성 신호의 변화를 관찰하여 수작업으로 추출하였기 때문에 약간의 오류를 내포하고 있을 것으로 추측된다.

(그림 3.4)는 맹목 세그먼테이션 방법과 본 고에서 제안한 방법을 실험 데이터에 적용하여 음소 경계를 추출한 결과를 나타낸 것이다. 이 그림의 x축과 y축은 각각 과추출 오류와 누락 오류이며, 실선은 후보 음소 경계 추출 과정의 임계치 계산에 필요한 가중치 상수 β를 2에서 8까지 1씩 증가시키면서 실험한 맹목 세그먼테이션의 결과이다. 각 점선은 각 β값에 해당하는 맹목 세그먼테이션의 결과와 유성음과 무성음을 구별할 수 있는 특이 점을 이용하여 무성음 구간의 음소 수를 최대 10개에서 1개까지 1씩 감소시키면서 실험한 결과를 나타낸 것이다. 각 점선의 오른쪽 끝점과 왼쪽 끝점은 각각 무성음 구간의 음소 수를 10개와 1개로 했을 때의 과추출 오류와 누락 오류이다.



(그림 3.4) Insertion과 Deletion비 사이의 관계

(그림 3.4)의 모든 점선과 실선을 비교해 보면 본 고에서 제안한 방법의 과추출 오류와 누락 오류가 맹목 세그먼테이션 방법에 비해 개선되었다는 것을 알 수 있으며, 특히 ①번 점선에서는 누락 오류가 약 10%일 때 과추출 오류는 맹목 세그먼테이션 방법에 비해 약 40%정도 감소함을 확인할 수 있다. 또 이 그림의 모든 점선에서는 무성음 구간의 제한 음소 수가 감소하면 누락 오류는 증가하고 과추출 오류는 감소한다는 것을 알 수 있다.

<표 3.1>은 (그림 3.4)에 나타낸 실험 결과 중 맹목 세그먼테이션의 결과와 무성음 구간의 허용 음소 수 N이 2와 3일 때의 결과를 요약한 것이다. 이 표의 β는 후보 음소 경계 추출 과정의 임계치 계산에 필요한 가중치 상수이며, 제안한 방법의 결과와 맹목 세그먼테이션 결과는 이 β값을 2에서 8까지 1씩 증가시키면서 실험한 결과이다. N = 2일 때 약 20%, N = 3일 때 약 40%정도의 과추출 오류가 감소됨을 확인할 수 있다.

〈표 3.1〉 실험 결과(N : 무성음 구간의 음소 수)

β	맹목 세그먼테이션		제안한 방법(N=2)		제안한 방법(N=3)	
	과추출 오류[%]	누락 오류[%]	과추출 오류[%]	누락 오류[%]	과추출 오류[%]	누락 오류[%]
2	210.511	1.736	120.119	7.919	137.527	5.85
3	156.124	6.397	88.918	11.439	103.9	9.37
4	118.383	12.39	68.823	16.361	81.831	14.649
5	93.08	18.454	55.172	21.712	66.968	20.262
6	76.742	23.448	45.803	26.207	56.908	24.875
7	64.067	26.801	38.787	29.061	49.465	27.99
8	54.649	30.535	33.769	32.438	43.829	31.272

(그림 3.4)와 <표 3.1>에서는 β값이 증가할수록 각 방법의 과추출 오류는 감소하고, 누락 오류는 증가하는데, 이는 β값이 증가하면 맹목 세그먼테이션(제안 방법의 후보 음소 경계 추출과정)의 클러스터링 임계치가 감소하므로, 맹목 세그먼테이션을 통해서 추출한 음소 경계 수(제안 방법의 후보 음소 경계 수)가 감소하고, β값이 감소하면 클러스터링 임계치가 증가하므로, 맹목 세그먼테이션을 통해서 추출한 음소 경계 수가 증가하기 때문이다. <표 3.1>과 (그림 3.4)의 결과를 종합해 볼 때 본 고에서 제안한 무성음 구간의 음소 수 제한 방법은 약간의 누락 오류를 발생시키지만, 과추출 오류를 40%정도까지 줄일 수 있어 맹목 세그먼테이션 방법의 성능 개선에 효과가 있음을 알 수 있다.

4. 결 론

본 논문에서는 한국어의 음절 구조에 관한 간단한 지식, 유성음 구간과 무성음 구간을 구분할 수 있는 특이 점, 그리고 특징 벡터의 확률 밀도와 클러스터링 방법을 이용하여 미지의 발화로부터 음소의 경계를 추출하는 방법을 제안하였다. 이 방법은 특이 점을 추출하는 과정, 음소의 후보경계를 추출하는 과정, 그리고 음소 경계를 추출하는 과정으로 구성된다.

특이 점을 추출하는 과정에서는 프레임 길이를 10ms로 하여 512-점FFT를 수행한 다음 0~400Hz 사이의 전력 밀도로부터 유성음 구간과 무성음 구간을 구별해 내며, 음소의 후보경계를 추출하는 과정에서는 연속된 두 클러스터 사이의 유사도를 이용하여 클러스터링 함으로써 후보 음소 경계를 추출한다. 최종 음소 경계를 추출하는 과정에서는 특이 점과 음소의 후보경계에 한글의 구조적 지식을 적용하여 무성음 구간의 음소 개수를 제한하였다.

제안된 방법의 검증을 위해서 한국어 음성 데이터베이스인 POW로부터 음절수가 6 이상인 240개의 발화 데이터를 선별하여 음소 경계추출 실험에 사용하였으며, 추출된 음소 경계가 기준 음소 경계의 좌/우 20ms 내에 있을 경우 올바르게 음소의 경계를 추출한 것으로 간주하였다. 그 결과 누락 오류가 약 10%일 때 과추출 오류가 맹목 세그먼테이션 방법에 비해 약 40%정도 감소되어 제안한 방법이 맹목 세그먼테이션 방법의 성능 개선에 효과가 있음을 알 수 있었다. 제안한 방법의 성능은 유/무성음 구간 뿐만 아니라 다른 특이 점을 추출하는 것과 최종 음소 경계의 튜닝을 통하여 향상시킬 수 있을 것으로 기대된다.

참 고 문 헌

[1] Torbjϕrn Svendsen and Frank K. Soong, "On the Automatic Segmentation of Speech Signals," Proc. ICASSP 87, pp.77-80, 1987.

[2] Youngjoo Suh and Youngjik Lee, "Phoneme Segmentation of Continuous Speech Using Multi-Layer Perceptron," Proc. ICSLP 96, Vol.3, pp.1297-1300, 1996.

[3] Bryan L. Pellon and John H. L. Hansen, "Automatic Segmentation of Speech Recorded in Unknown Noisy Channel Characteristics," Duke Univ., Technical Report RSPL-98-9, 1998.

[4] Bryan L. Pellon and John H. L. Hansen, "Automatic Segmentation and Labeling of Speech using the Duke University Speech Time-Aligner," Duke Univ., Technical Report RSPL-96-22, 1996.

[5] Brian Eberman and William Goldenthal, "Time-Based Clustering for Phonetic Segmentation," Proc. ICSLP 96, Vol.2, pp.1225-1228, 1996.

[6] Regine Andre-Obrecht, "Automatic Segmentation of Continuous Speech Signals," Proc. ICASSP 86, pp.2275-2278, 1986.

[7] Regine Andre-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals," IEEE Trans. ASSP, Vol. 36, No.1, pp.29-40, January, 1988.

[8] Sharlene A. Liu, "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am., Vol.100, No.5, pp.3417-3430, November, 1996.

유 강 주

e-mail : gjyu@ce.kmaritime.ac.kr

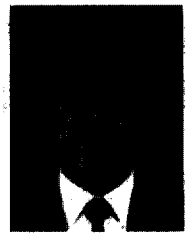
1996년 한국해양대학교 제어계측공학과 졸업(학사)

1998년 한국해양대학교 대학원 제어계측공학과(석사)

1998년~현재 한국해양대학교 시간강사

1999년~현재 한국해양대학교 제어계측공학과 대학원 (박사 과정)

관심분야 : 음성신호처리, 음성인식



신 옥 근

e-mail : okshin@hanara.kmaritime.ac.kr

1981년 서강대학교 전자공학과 졸업(학사)

1983년 부산대학교 전자공학과 (공학석사)

1989년 프랑스 Université_ de Franche-Comté_ (공학박사)

1983년~1995년 한국전자통신연구소 선임연구원

1995년~현재 한국해양대학교 자동화정보공학부 조교수

관심분야 : 신호처리, 음성신호처리, 음성인식