Journal of the Industrial Technology Research Institute.
Korea Maritime University. Vol. 17. 1999.

확률 분포의 유사도와 클러스터링을 이용한 연속 음성의 음소경계 추출

유 강 주* · 신 옥 근**

A Speech Segmentation Method Based on Clustering and the Similarity of Probability Distribution

Gang-Ju You*, Ok-Keun Shin**

Abstract

A two stage real-time phoneme boundary detection method is proposed. The first stage, based on the Euclidean distances between adjacent feature vectors, extracts candidate phoneme boundaries by merging iteratively pairs of candidate segments whose inbetween distance is minimum. The iteration stops when the minimum distance found exceeds the threshold. The second stage, based on the probabilistic model representing the similarity of vectors between adjacent segments, refines the candidate phoneme boundaries found in the first stage.

The proposed method is applied to the test data set of TIMIT database and shows the real-time processing capability with a rather promising performance corresponding to 96.8% of phoneme boundary detection rate with 90.7% of insertion rate. In this experiment, an extracted boundary is considered to be correct, if it is located within 20ms from the manually segmented boundary.

Ⅰ. 서 론

최근에 많이 연구되고 있는 연속 음성의 인식(Speech Recognition) 방법에는 크게 나누어 프레임 단위(Frame Unit)로 인식하는 방법과 세그먼트 단위(Segment Unit)로 인식하는 방법이 있다.

전자의 경우 일상의 대화체 발화(Utterance)를 인식한다든지 인식해야 할 어휘가 많을 경우 어려움이 따른다. 특히 음성 모델(Speech Model)을 만들기 위해서는 많은 발화가 필요한데, 그것

* 한국해양대학교 대학원 컴퓨터공학과
** 한국해양대학교 컴퓨터공학과 조교수

들을 수집하는 것이 어렵다.[21]

후자의 경우에는 세그먼트를 음소(Phone), 음절(Syllable) 또는 단어(Word)등을 인식단위로 하여 인식하는 방법이 있고, 흔히 쓰이는 것이 음소단위 인식이다. 이 인식 방법은 대용량의 어휘 인식이나, 일상의 대화체 발화를 인식하는 자동 음성 인식(Automatic Speech Recognition: ASR) 시스템, 또는 음성을 합성하거나 만들어내는 시스템의 구현이 비교적 용이해서 많이 이용되고 있다. 이런 시스템에서는 음성 신호로부터 음소들을 정확히 추출하는 것(Speech Segmentation)이 결정적인 역할을 한다.

음성신호에서 음소단위의 경계를 추출하는 방법들에 대해 많은 연구가 진행되어 오고 있다. 이러한 방법들에는 HMM(Hidden Markov Model)을 이용해서 각 음소에 대한 모델을 만들고, 이를 이용해서 음소를 추출하는 방법[10-16,24-27], 각 음소의 레퍼런스(Reference)와 미지의 발화를 비교해서 음소를 추출하는 방법(Template Matching Method)[21], 벡터 양자화(Vector Quantization)와 클러스터링(Clustering)을 이용하는 방법[18], 음성의 변화(Speech Variation)를 이용해서 추출하는 방법[8,9,17,21], 확률 모델을 이용해서 추출하는 방법[19,20,22,23], 스펙트럼의 변화(Spectrum Variation)와 다이나믹 프로그램(Dynamic Program)의 일종인 레벨 빌딩 알고리즘(Level Building Algorithm)을 이용하는 방법[17] 등이 있다. 이들 중에서 확률모델을 이용하여 음소의 경계를 추출하는 방법의 성능이 우수하여 많이 연구되고 있으며, 특히 사전 지식 없이 자동으로 음소의 경계를 추출하는 방법에는 Andre-Obrecht가 제안한 방법[20,22]과 Goldental과 Eberman이 제안한 방법[23]등이 있다. 전자의 경우에는 음성샘플의 변화를 이용하여, 음소의 경계를 추출하며, 후자의 경우에는 연속된 두 세그먼트 사이의 특징벡터에 대한 분포와 클러스터링 방법을 이용하여 음소의 경계를 추출한다. 이 방법들은 성능은 우수하지만 계산량이 많아서 실시간 처리가 어렵다는 단점이 있다.

본 논문에서는 Goldental등의 방법을 기초로 하여 실시간으로 임의의 발화로부터 음소를 추출하는 알고리즘을 제안한다. 이 방법은 다음과 같이 전처리 과정(Preprocessing Unit)과 후처리 과정(Postprocessing Unit)으로 구성된다.

전처리 과정에서는 한 프레임을 10 ms로 하고, 연속된 두 세그먼트의 특징벡터에 대한 유클리디언 거리와 클러스터링 방법을 이용하여 음소의 후보경계를 추출한다.

후처리 과정에서는 한 프레임을 5ms로 하고, 전처리 과정에서 추출한 음소의 후보 경계를 토대로 하여 정밀하게 음소의 경계를 추출하며, Goldental등의 방법을 변형시켜 적용한다.

제안한 방법을 NIST(National Institute of Standards and Technology)에서 제작한 연속 음성 데이터베이스인 TIMIT(영어 음성 데이터베이스)에 적용하여, 음소의 경계를 추출하고, 그 결과를 검증하기 위하여 TIMIT 음성 데이터베이스에 있는 음소 경계와 비교하였다. 그리고 추출한 음소의 경계가 TIMIT 음성 데이터베이스에 있는 음소경계의 좌/우 20ms내에 있으면 올바르게 추출한 것으로 간주하였다.

본 논문의 II장에서는 제안한 음소경계 추출 방법을, III장에서는 제안한 방법의 검증을 위해서 TIMIT 음성 데이터베이스로 실험한 결과를 서술한 다음, IV장에서 결론을 맺는다.

Ⅱ. 음소의 경계 추출

본 논문에서 제안하는 방법은 그림 2.1과 같이 두개의 단계로 구성된다. 첫 번째는 전처리 단계로서, 연속된 두 세그먼트사이의 유클리디언 거리가 가장 가까운 세그먼트 쌍을 합하는 클러스터링 방법을 이용하여 음소의 후보 경계(Segment Boundary)를 추출한다. 두 번째는 후처리 단계로 전처리 과정에서 추출한 음소의 후보경계를 토대로 하여, 연속된 두 프레임사이의 확률 모델과 클러스터링을 이용하여 음소의 경계를 추출한다.

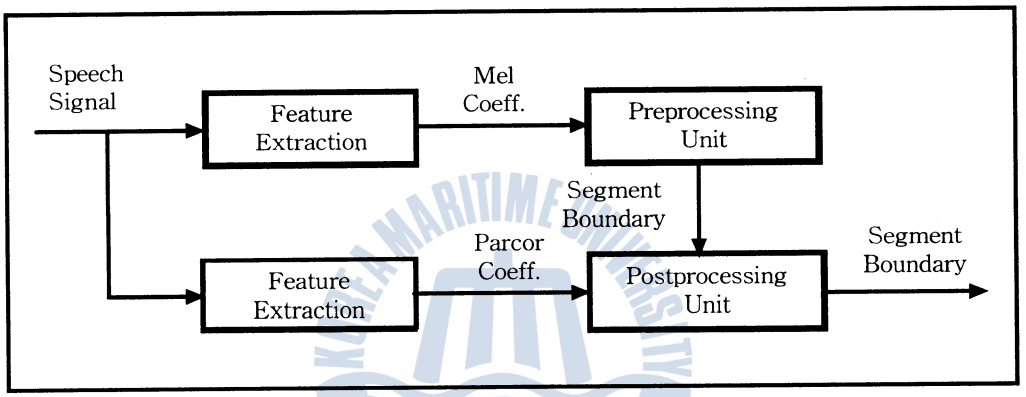

그림 2.1 제안한 방법의 구성도
Fig 2.1 Block diagram of the proposed method

2.1 전처리 단계

그림 2.2는 연속된 두 세그먼트 사이의 거리를 이용해서 클러스터링 하는 과정을 나타낸 것이다. 이 그림에서 S_i는 i번째 세그먼트를, D_{i-1}은 $(i-1)$번째 세그먼트와 i번째 세그먼트사이의 거리를 나타낸다. 그리고 이들을 이용해서 클러스터링 하는 과정은 다음과 같다.

1) 연속된 두 세그먼트 사이의 거리 $D_i(i=1,...,N-1)$를 모든 세그먼트 쌍에 대해서 구한다.
2) 1)에서 구한 거리들 중에서 가장 작은 거리와 그것에 해당하는 연속된 두 세그먼트를 구한다.
3) 2)에서 구한 거리가 임계치 보다 작다면 이 두 세그먼트를 합하여 하나의 세그먼트로 하고, 세그먼트의 개수를 하나 감소시킨 다음, 이 과정을 반복적으로 수행한다. 임계치 보다 작은 거리를 갖는 세그먼트 쌍이 없을 때 이 과정을 멈추고, 남아있는 세그먼트의 경계를 음소의 후보경계로 한다.

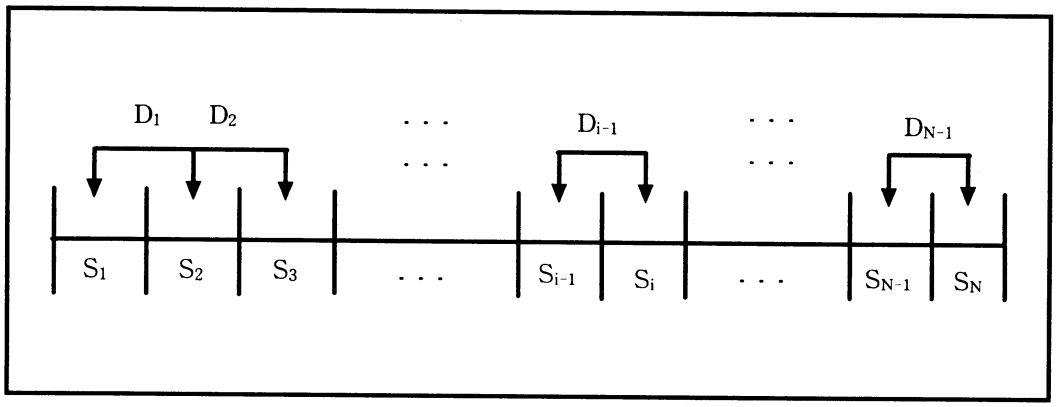

그림 2.2 연속된 두 세그먼트사이의 거리
Fig 2.2 The distance between two successive segments

아래에 전처리 단계에서 세그먼트의 경계를 추출하는 방법을 자세히 기술한다. 이 과정에서 사용되는 세그먼트의 집합(S)과 각 세그먼트에 대응하는 특징벡터의 집합(X)을 다음과 같이 정의한다.

$$S=(S_1, S_2, S_3, \ldots, S_i, \ldots, S_N) \quad \cdots\cdots (2\text{-}1)$$

$$X=(X_1, X_2, X_3, \ldots, X_i, \ldots, X_N) \quad \cdots\cdots (2\text{-}2)$$

여기서 N은 모든 세그먼트의 개수, X_i는 P차원의 특징 벡터이며, 연속된 두 세그먼트 사이의 거리 D_i는 식(2-3)과 같다.

$$D_i=\sqrt{\sum (X_i - X_{i+1})^T W (X_i - X_{i+1})}, i=1, \ldots, N-1 \quad \cdots\cdots (2\text{-}3)$$

$$W=\alpha I \quad \cdots\cdots (2\text{-}4)$$

여기서 W는 가중치 행렬이고, α는 i번째 세그먼트 와 $(i+1)$번째 세그먼트에 포함된 프레임의 개수이며, I는 항등행렬이다. 그리고 최소 거리 $mDist$와 이것에 해당하는 세그먼트의 인덱스 i^*는 각각 식(2-5)와 (2-6)와 같다.

$$mDist = \min_{i=1,\ldots,N-1} D_i \quad \cdots\cdots (2\text{-}5)$$

$$i^* = \arg\min_{i=1,\ldots,N-1} D_i \quad \cdots\cdots (2\text{-}6)$$

식(2-5)에서 $mDist$가 임계치 Thr보다 적으면, 식(2-7)에 의해서 i^*번째 세그먼트와 (i^*+1)번째 세그먼트를 합하여 하나의 세그먼트 특징벡터로 하고, N을 하나 감소시킨 다음 이 과정을 반복하며, 임계치 Thr보다 크다면 클러스터링을 멈춘다.

$$X'_{i^*} = \frac{\sum X_i}{SFN}, \ (X_i \in S_{i^*}, S_{i^*+1}) \quad \cdots\cdots (2\text{-}7)$$

여기서 X_i는 세그먼트 S_{i^*}과 S_{i^*+1}에 포함된 프레임들에 해당하는 특징벡터이고, SFN은 합해진 세그먼트에 포함된 프레임의 개수이며, X'_i 는 i^*번째 세그먼트의 새로운 특징벡터이다.

임계치 Thr은 모든 연속된 두 세그먼트 사이의 평균 거리보다 큰 거리들의 평균 OM과 이들의 표준 편차 $OSDiv$를 이용하여 구하며, 식(2-8)과 같다.

$$Thr = OM + \beta * OSDiv \quad \cdots\cdots (2\text{-}7)$$

여기서 β는 상수로 실험에 의해서 구한다.

2.2 후처리 단계

후처리 단계에서는 전처리 단계에서 추출한 음소의 후보경계에 대하여 Goldental과 Eberman이 제안한 방법[23]을 실시간 처리가 가능하도록 변형시켜 음소의 경계를 추출한다. 이 단계에서도 전처리와 비슷한 클러스터링 방법을 사용하였으며, 연속된 두 후보 세그먼트 사이의 거리는 특징벡터의 분포에 대한 유사도로 표현한다. 그림 2.3의 $W1$과 $W2$는 연속된 두 세그먼트 $S1$과 $S2$의 확률 모델을, $W0$는 두 세그먼트를 하나의 세그먼트 $S0$로 간주했을 때의 확률 모델을 나타낸다.

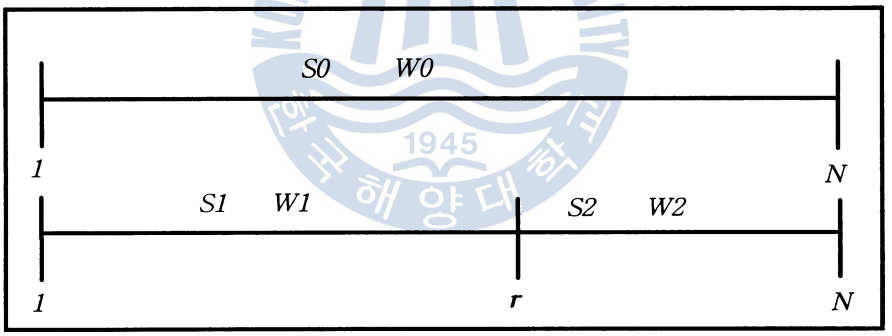

그림 2.3 세그먼트($S0$;$S1$,$S2$)와 확률 모델($W0$;$W1$,$W2$)
Fig 2.3 Segment($S0$;$S1$,$S2$) and their probabilistic models($W0$;$W1$,$W2$)

이 그림에서 N은 연속된 두 세그먼트 $S1$과 $S2$에 포함되어 있는 시간영역 음성샘플의 개수이며, r은 세그먼트 $S1$에 포함된 음성샘플 개수이다. 그리고 이 그림에서 언급한 모델은 일반적인 선형예측모델(Linear Predictive Model)[1,2,5-7]이며, 각 확률 모델은 선형예측 계수(Linear Predictive Coefficient)와 분산(Variance)으로 나타낼 수 있다.

일반적인 선형예측 모델은

$$y(n) = \sum_{i=1}^{p} a_i y(n-i) + e(n) \quad \cdots\cdots (2\text{-}9)$$

이다. 여기서 p는 선형 예측계수의 차수이며, $e(n)$은 예측오차(Prediction Error)로, 평균이 0이고,

분산이 V인 가우시안 프로세스(Gaussian Process)이다. 선형 예측계수 a_i와 분산 V로 이루어진 확률 모델 $W=(a_i, V)$이 주어졌을 때, 이 모델에서 음성샘플의 시퀀스 y_1^r이 나올 확률 $L(y_1^r|W)$은

$$L(y_1^r|W)=\prod_{m=1}^{r} p(e(m)|y_{m-p}^{m-1}, W) \quad \cdots\cdots (2\text{-}10)$$

이다. 여기서 y_1^r은 음성샘플의 시퀀스 $y(1), y(2), \ldots, y(r)$을 나타낸다. 그림 2.3에서 확률 모델 $W0$는 음성샘플 y_1^N을, $W1$은 음성샘플 y_1^r을, W2는 음성샘플 y_{r+1}^N을 가지며, 이것들에 대한 유사도 $D1$은 식(2-11)과 같다.[23]

$$D1=\underset{W0}{Max}\ \underset{W1}{Max}\ \underset{W2}{Max}\ \frac{L(y_1^N|W0)}{L(y_1^r|W1)\,L(y_{r+1}^N|W2)} \quad \cdots\cdots (2\text{-}11)$$

이 식에서 $L(y_1^r|W1)$은 확률 모델 W1에서 음성샘플 y_1^r이 출력될 확률이며, $L(y_{r+1}^N|W2)$은 확률 모델 $W2$에서 음성샘플 y_{r+1}^N이 출력될 확률이고, $L(y_1^N|W0)$는 확률 모델 $W0$에서 음성샘플 y_1^N이 출력될 확률이다. 그리고 이것들에 대한 유사도 $D1$은 음성샘플 y_1^r과 y_{r+1}^N의 특징벡터 분포가 유사하면 큰 값이 되고, 전혀 다르면 아주 작은 값이 된다.

본 논문에서는 식(2-11)의 유사도 $D1$을 구하기 위해 선형예측 계수 대신 Goldental과 Eberman이 제안한 Parcor 계수 $k(i)$사이의 거리 측정법[23]을 사용하였으며, 계수의 차수 p는 훈련에 의해서 구했다. 세그먼트 S의 Parcor 계수 $k(i)$, $i=1, \ldots, p$와 에너지 E가 주어졌을 때, 레지듀얼 에너지(Residual Energy) Ep는

$$E_p=E-\sum_{i=1}^{p} k(i)^2 \quad \cdots\cdots (2\text{-}12)$$

이다. 식(2-12)에서 Ep는 앞에서 언급한 가우시안 프로세스의 분산 V와 같다. 식(2-10)에서 선형예측 계수와 분산을 각각 Parcor 계수와 레지듀얼 에너지 Ep로 바꾸고, 이 식에 자연 대수를 취하면 식(2-13)과 같다.

$$l(y_1^r|W)=\frac{-r}{2}(\log(2\pi)+\log(E_p)+1) \quad \cdots\cdots (2\text{-}13)$$

그리고 식(2-13)에서 코딩 비용을 빼면, 확률 모델 W에서 y_1^r이 나올 확률의 자연 대수 값 $M(y_1^r)$이 되고, $M(y_1^r)$은

$$M(y_1^r)=l(y_1^r|W)-\frac{p}{2}\log(r) \quad \cdots\cdots (2\text{-}14)$$

이다. Goldental과 Eberman은 이 값을 구하기 위해, 식(2-14)를 Parcor 계수와 이 계수의 차수에 대해서 최대화시키는 MDL(Maximum Description Length)이라는 방법을 사용하였다. 이 방법은

모든 세그먼트에 대해서 각각을 최대화시켜야 하므로, 시간이 많이 걸리는 단점이 있다. 본 논문에서는 모든 세그먼트에 대해서 이 값을 최대화시킬 수 있는 Parcor 계수의 차수 p를 실험에 의해서 미리 정해놓음으로써, 실시간 세그먼테이션이 가능하도록 하였다.

세그먼트 $S1$과 $S2$가 주어졌을 때, 식(2-11)에 의해서 $S1$과 $S2$에 대한 유사도 $D1$은

$$D1 = M(y_1^r) + M(y_{r+1}^N) - M(y_1^N) \quad \cdots\cdots (2\text{-}15)$$

이 된다. 그리고 모든 세그먼트에 대한 거리 $D_i,\ i=1,\ldots,N-1$ 이 주어졌을 때, 임계치 Thr은

$$Thr = \beta \frac{\sum_{i=1}^{N-1} D_i}{N-1} \quad \cdots\cdots (2\text{-}16)$$

이다. 여기서 β는 상수이고, 훈련에 의해서 구해진다. 그리고 N은 모든 세그먼트의 개수를 나타낸다.

Ⅲ. 실험 및 고찰

본 논문에서 제안한 방법을 검증하기 위해 영어 음성 데이터베이스인 TIMIT 음성 데이터베이스의 테스트 데이터를 사용하였으며, 이들 중 절반은 임계치를 구하는데 필요한 β값과 Parcor 계수

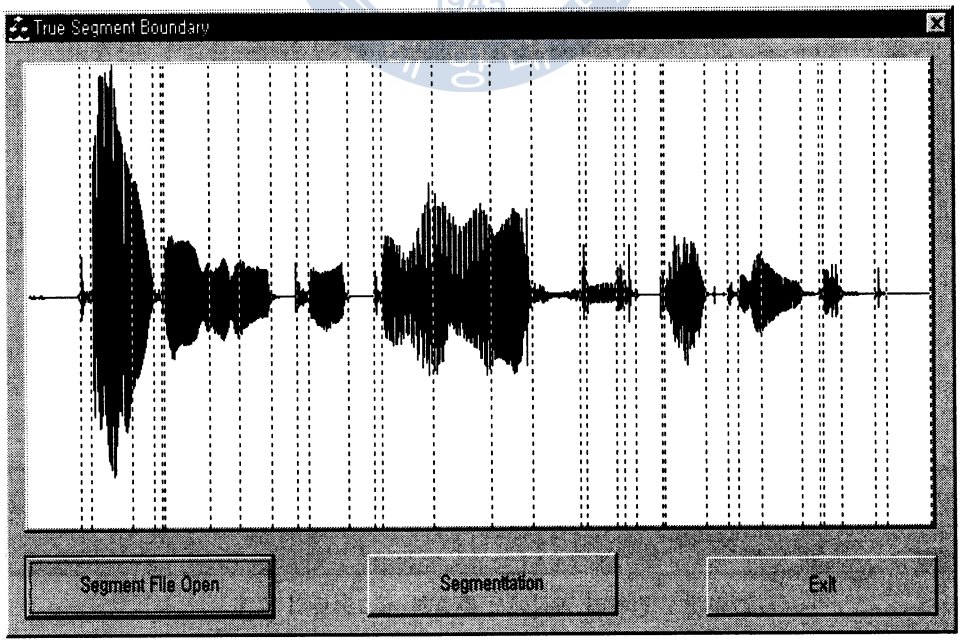

그림 3.1 TIMIT 음성 데이터베이스의 음소 경계

Fig 3.1 Manually segmented phoneme boundaries of the utterance "Elderly people are often excluded."

의 차수 p를 구하는데 사용하였고, 나머지 반은 음소의 경계를 추출하는데 사용하였다. 그리고 전처리 단계와 후처리 단계 모두, 음성샘플을 프레임단위로 나눌 때 오버랩(Overlap)과 프리엠퍼시스(Preemphasis)를 하지 않았다. 전처리 단계에서는 한 프레임을 10ms로 하여, 18차의 MFCC(Mel Frequency Cepstral Coefficient)[3,4]를 구하고, 거리 측정에는 가중치가 부여된 유클리디언 방법(Weighted Euclidean Method)을 사용하였다. 그리고 클러스터링을 하기 위해서 초기 세그먼트는 모든 프레임 각각을 하나의 세그먼트로 하였다. 후처리 단계에서는 한 프레임을 5ms로 하여, 16차의 Parcor 계수[1,2,6]를 구하고, 거리 측정에는 Goldental과 Eberman이 제안한 확률 모델에 기초한 LRM(Likelihood Ratio Method)방법[23]을 사용하였다.

그림 3.1은 TIMIT 음성 데이터베이스에 있는 "Elderly people are often excluded."라는 발화의 음소 경계를 나타낸 것이며, 수직으로 그린 점선이 음소의 경계이다.

그림 3.2는 전처리 과정을 통하여 추출한 음소의 후보경계를 나타낸 것이다. 그림 3.1과 그림 3.2를 비교해 볼 때, 전처리 단계에서는 TIMIT 음성 데이터베이스에 있는 음소의 경계를 대부분 포함하고, 그것보다는 많은 음소경계를 추출하였다는 것을 알 수 있다.

그림 3.3은 후처리 과정에서 추출한 음소경계이다. 그림 3.3과 그림 3.2를 비교해 볼 때, 후처리 과정을 통하여 전처리 과정에서 추출한 후보경계를 상당히 많이 줄였다는 것을 알 수 있다. 그리고 그림 3.3과 그림 3.1을 비교해 볼 때, 그림 3.1의 세그먼트 경계보다 90-100%정도 더 많은 음소경계를 추출하였다는 것을 알 수 있다.

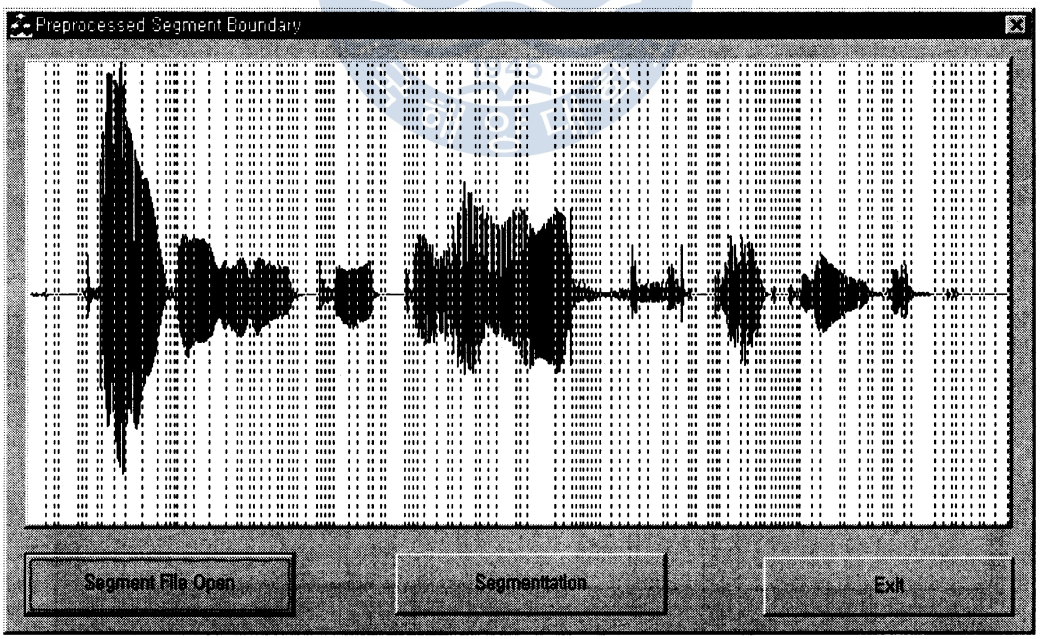

그림 3.2 전처리 과정에서 추출한 세그먼트의 경계

Fig 3.2 Segment boundaries after preprocessing

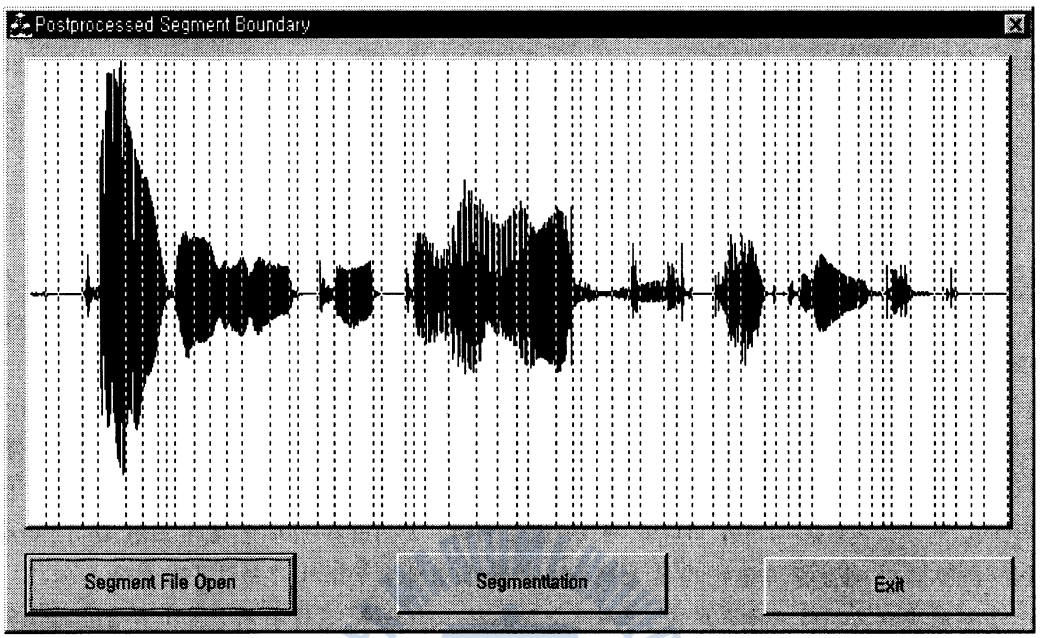

그림 3.3 후처리 과정에서 추출한 세그먼트의 경계
Fig 3.3 Segment boundaries after postprocessing

그림 3.4는 제안한 방법을 테스트 데이터에 적용해 본 결과를 나타낸 것이다. 이 그림에서 "Insertion"은 과추출 오류를 백분율로 나타낸 것이고, "Deletion"은 누락 오류를 백분율로 나타낸 것이다.

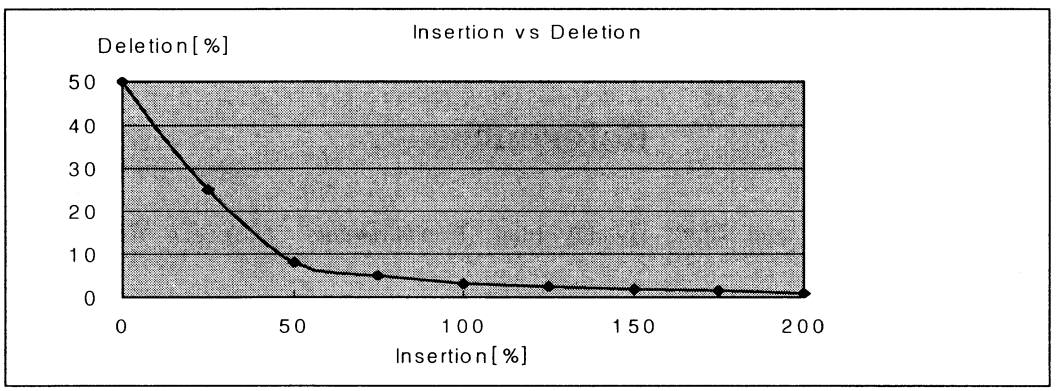

그림 3.4 Insertion과 Deletion비 사이의 관계
Fig 3.4 The relation between insertion and deletion ratio

제안한 방법으로 실험한 결과, 추출한 음소경계가 TIMIT 음성데이터베이스에 있는 음소경계의 좌/우 20ms내에 있을 경우 올바르게 음소의 경계를 추출한 것으로 간주하였을 때, Insertion은

90.743[%]였고, Deletion은 3.163[%]였으며, 올바르게 세그먼트의 경계를 추출한 것은 96.837[%]였다.

Ⅳ. 결 론

본 논문에서는 특징벡터들의 분포에 대한 확률 모델과 특징벡터들 사이의 거리, 그리고 클러스터링 방법을 이용하여, 실시간으로 임의의 발화로부터 음소의 경계를 추출하는 방법을 제안하였다. 이 방법은 다음과 같이 전처리 과정과 후처리 과정으로 구성된다.

첫 번째 단계는 전처리 과정으로 한 프레임을 10ms로 하여 MFCC를 추출한 다음, 연속된 두 세그먼트 사이의 특징벡터 변화에 기반을 둔 클러스터링 방법을 이용한다. 연속된 두 세그먼트사이의 유클리디언 거리를 모든 세그먼트에 대해서 각각 구하고, 이 거리들 중에서 가장 작은 값이 임계치 보다 작으면 두 세그먼트를 합하여 하나의 세그먼트로 하며, 그 거리가 임계치 보다 크면 클러스터링 과정을 멈춘다. 이 과정을 수행한 후에 남은 세그먼트의 경계를 음소의 후보경계로 한다.

두 번째 단계는 후처리 과정으로 한 프레임을 5ms로 하였으며, 전처리 과정에서 추출한 음소의 후보경계를 토대로 하여, 연속된 두 세그먼트 사이의 Parcor 계수의 MDL 확률 모델과 클러스터링 방법을 이용한다.

제안된 방법을 TIMIT 음성 데이터베이스의 테스트 세트에 적용해본 결과, 추출한 음소경계가 TIMIT에 있는 음소경계의 좌/우 20ms내에 있을 경우 올바르게 음소의 경계를 추출한 것으로 간주하였을 때, 음소경계가 올바르게 추출된 것은 96.837[%]이고, 실제 음소경계 보다 90.743[%]만큼 더 많이 추출하였으며, 음소경계를 추출하지 못한 것은 3.163[%]였다. 제안된 방법을 이용하여 미지의 발화로부터 비교적 정확한 음소의 경계 추출이 가능하였으며, 제안한 방법의 성능은 튜닝을 통해서 향상시킬 수 있을 것으로 기대된다.

References

[1] Lawrence Rabiner and Biing Hwang Jung, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] Claudio Becchetti and Lucio Prina Ricotti, *Speech Recognition Theory and C++ Implementation*, John Wiley and Sons Ltd, pp. 122-143, 1999.

[3] Steven B. Davis and Paul Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. ASSP, Vol. 28, No. 4, pp. 357-366, August, 1980.

[4] Rivarol Vergin, Douglas O'Shaughnessy and Vishwa Gupta, "Compensated Mel Frequency Cepstrum Coefficients," Proc. ICASSP-96, Vol. 1, pp. 323-326, 1996.

[5] Ronald W. Schafer and Lawrence Rabiner, "Digital Representations of Speech

Signals," Proc. IEEE, Vol. 63, No. 4, pp. 662-677, April, 1975.

[6] John Makhol, "Linear Prediction : A Tutorial Review," Proc. IEEE, Vol. 63, No. 4, April, 1975.

[7] Joseph W.Picon, "Signal Modeling Techniques in Speech Recognition," Proc. IEEE, Vol. 81, No. 9, September, 1993.

[8] Bojan Petek, Ove Andersen and Paul Dalsgaard, "On the Robust Automatic Segmentation of Spontaneous Speech," Proc. ICSLP-96, Vol. 2, October, 1996.

[9] Ivan Kopeck, "Automatic segmentation into Syllable Segments," Proc. Language Resources and Evaluation, pp. 1275-1279, May, 1998.

[10] Steven C. Lee and James R. Glass, "Real-Time Probabilistic Segmentation for Segment-Based Speech Recognition," Proc. ICSLP-98, November, 1998.

[11] Jane W. Chang and James R. Glass, "Segmentation and Modeling in Segment-Based Recognition," Proc. Eurospeech-97, pp.1199-1202, 1997.

[12] Bryan L. Pellom and John H.L. Hansen, "Automatic Segmentation and Labeling of Speech Recognition in Unknown Noisy Channel Environments," Proc. ESCA-NATO WORKSHOP, pp. 167-170, 1997.

[13] Andreas Kipp, Maria-Barbara Wesenick and Florian Schiel, "Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech," Proc. Eurospeech-97, pp. 1023-1026, 1997.

[14] Andreas Kipp, Maria-Barbara Wesenick and Florian Schiel, "Automatic Detection and Segmentation of Pronunciation Variation in German Speech Corpora," Proc. ICSLP-96, pp. 106-109, 1996.

[15] Antonio Bonafonte, Albino Nogueira and Antonio Rodriguez-Garrido, "Explicit Segmentation of Speech Using Gaussian Models," Proc. ICSLP-96, Vol. 2, October, 1996.

[16] James Glass, Jane Chang and Michael McCandless, "A Probabilistic Framework for Feature-Based Speech Recognition," Proc. ICSLP-96, pp. 2277-2280, 1996.

[17] Manish Sharma and Richard Mammone, " "BlIND" Speech Segmentation: Automatic Segmentation of Speech Without Linguistic Knowledge," Proc. ICSLP-96, Vol. 2, 1996.

[18] Yoshinao Shiraki and Masaaki Honda, "LPC Speech Coding Based on Variable-Length Segment Quantization," IEEE Trans. ASSP, Vol. 36, No. 9, September, 1988.

[19] Mari Ostendorf and Salim Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," IEEE Trans. ASSP, Vol. 37, No. 12, December, 1989.

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

[21] Torbjorn Svendsen and Frank K. Soong, "On the Automatic Segmentation of Speech

Signals," Proc. ICASSP-87, pp. 77-80, 1987.

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

[23] Brian Eberman and William Goldental, "Time-Based Clustering for Phonetic Segmentation," Proc. ICSLP-96, Vol. 2, 1996.

[24] Jane W. Chang, "Near-Miss Modeling: A Segment-Based Approach to Speech Recognition," PhD Thesis, Massachusetts Institute of Technology, 1998.

[25] Steven C. Lee, "Probabilistic Segmentation for Segment-Based Speech Recognition," PhD Thesis, Massachusetts Institute of Technology, 1998.

[26] Trym Holter, "Maximum Likelihood Modeling of Pronunciation in Automatic Speech Recognition," Norwegian University, 1997.

[27] Philipp Schmid, "Explicit N-Best Formant Features for Segment-Based Speech Recognition," Oregon Graduate Institute of Science & Technology, 1996.