음절핵의 위치정보를 이용한 우리말의 음소경계 추출

Utilization of Syllabic Nuclei Location in Korean Speech Segmentation into Phonemic Units

신 옥 근*
(Ok-Keun Shin*)

요 약

음성신호의 음소경계 추출방법 중 음소에 대한 사전지식 없이 음성 데이타, 혹은 특징벡터의 변화를 감지하여 음소경계를 추출해 내는 맹목 세그먼테이션은 연속음성 인식시스템이나 코퍼스 제작에 중요한 역할을 하며 많은 연구가 진행되어 왔다. 이러한 맹목 세그먼테이션 방법은 사전지식을 필요로 하지 않아 비교적 쉽게 접근할 수 있으나 음운학적인 지식, 또는 음소나 음소경계에 대한 지식과 경험 데이타 등을 이용하는 지식 기반 세그먼테이션 방법에 비해 성능이 좋지 못한 단점이 있다. 본고에서는 우리말의 연속 음성을 맹목 세그먼테이션해서 후보 경계를 추출한 다음, 음절핵의 위치정보를 이용하여 후보 경계를 후처리함으로써 세그먼테이션 효율을 높이는 방법을 제안한다. 제안하는 방법의 전처리과정에서는 확률적인 거리 모델을 이용한 클러스터링 방법을 이용하였으며, 후처리과정에서는 음절의 핵 사이에 위치할 수 있는 음소의 수는 제한된다는 선험적인 지식을 이용하였다. 실험결과, 제안하는 방법을 이용했을 때의 삽입오류는 맹목 세그먼테이션에 비해 약 25% 감소하였다.

핵심용어: 음성인식, 음소경계추출, 음성 코퍼스, 음절핵, 후처리, 맹목 세그먼테이션, 음소
투고분야: 음성처리 분야(2.5)

ABSTRACT

The blind segmentation method, which segments input speech data into recognition unit without any prior knowledge, plays an important role in continuous speech recognition system and corpus generation. As no prior knowledge is required, this method is rather simple to implement, but in general, it suffers from bad performance when compared to the knowledge-based segmentation method. In this paper, we introduce a method to improve the performance of a blind segmentation of Korean continuous speech by postprocessing the segment boundaries obtained from the blind segmentation. In the preprocessing stage, the candidate boundaries are extracted by a clustering technique based on the GLR(generalized likelihood ratio) distance measure. In the postprocessing stage, the final phoneme boundaries are selected from the candidates by utilizing a simple a priori knowledge on the syllabic structure of Korean, i.e., the maximum number of phonemes between any consecutive nuclei is limited. The experimental result was rather promising : the proposed method yields 25% reduction of insertion error rate compared that of the blind segmentation alone.

Keyword: Speech recognition, Blind segmentation, Corpus, Nuclei, Distance measure, Phoneme

I. 서 론

세그먼트 모델을 이용한 음성인식에 대한 연구에서 음성신호의 기본 인식 단위인 음절, 혹은 음소사이의 경계를 추출해 내는 세그먼테이션은 음성정보의 사전 지식 활용 여부에 따라 맹목 세그먼테이션(blind segmentation)과 지식기반 세그먼테이션으로 크게 나눌 수 있다. 전자의 맹목 세그먼테이션은 발화된 음성에 대한 아무런 사전 지식 없이 음성신호, 또는 스펙트럼의 불연속점을 찾아 인식 단위의 경계를 찾는 방법이고, 후자의 지식 기반 세그먼테이션에는 발화의 전사(transcription), 음소의 수 등과 같은 발화에 대한 직접적인 지식, 음운학적인 특징 등의 선험적인 지식, 그리고 확률모델, HMM 또는 신경망 등의 경우와 같이 훈련된 데이터를 이용하는 방법이 있다.

지식기반 세그먼테이션들 중 발화에 대한 직접적인 지식을 이용하는 방법은 주로 코퍼스 제작 등에 유용하며, 가장 효율이 높은 것으로 알려진 훈련된 데이터를 사용하는 방법은 공개된 음성 코퍼스가 거의 없는 우리말의

* 한국해양대학교 자동화정보공학부 조교수
접수일자 : 2000년 2월 14일

경우, 이용에 많은 제약이 따른다. 음운학적인 특징과 같은 선험적인 지식을 이용하는 방법은 음성인식, 코퍼스 제작 등에 다양하게 이용될 수 있으며, 이를 이용한 연속음성의 음절, 혹은 음소 단위 경계추출에 대한 연구가 Mermelstein[4]의 1975년의 연구 이후 근래에 다시 활발해지고 있다[1, 5, 13, 14]. 특히 Wu[13] 등은 영어의 연속 음성신호의 에너지 밀도 스펙트럼으로부터 음절의 시작점(onset)을 추출함으로써 단어의 오인식율을 10%정도 감소 시킬 수 있었으며, Lee [1] 등은 통계적인 세그먼테이션 모델에 음향학적인 특이점(landmark)에 대한 지식을 이용하여 연속영어음성에서 후보 음소경계의 수를 30% 감소시킬 수 있었다.

이들이 대상으로 하는 영어는 매우 복잡한 음절구조를 가지고 있어 음절단위의 세그먼테이션이나 인식이 용이하지 않음을 고려한다면 영어보다 훨씬 간단한 음절구조를 갖는 우리말의 음운학적인 특징을 이용한다면 더 좋은 성과를 얻을 수 있을 것이다.

본고에서는 우리말의 연속음성에서 음절핵(syllabic nucleus, 모음)의 위치를 찾은 다음, 인접한 음절핵 사이의 음소의 수는 대부분의 경우 최대 3개 (종성-휴지부/폐쇄-초성)이내로 제한 된다는 선험적인 지식을 음소경계 추출에 이용하는 방법에 대해 기술한다. 이를 위해 통계적인 모델을 이용한 방법으로 연속음성으로부터 후보 음소경계와 음절핵의 위치를 각각 추출한 다음, 이웃하는 음절핵 사이에 올 수 있는 음소경계의 수에 제한을 두어서 최종적인 음소경계를 추출해 낸다. 본고에서는 편의상 음성신호로부터 후보 경계를 추출해 내는 과정을 전처리, 후보 경계로부터 음절핵의 위치를 이용하여 최종적인 음소경계를 추출해 내는 과정을 후처리라 부르기로 한다(그림 1). 전처리 과정은 Goldenthal 등[11]이 제안한 맹목 세그먼테이션방법을 변형시켜 사용하고, 후처리 과정에서는 F1(first formant)의 에너지 밀도 스펙트럼에 Mermelstein이 제안한 소위 컨벡스 홀(convex-hull) 알고리즘[4]을 적용하여 음절핵을 추출한 다음 이것을 이용하여 최종적인 음소경계를 선택한다.

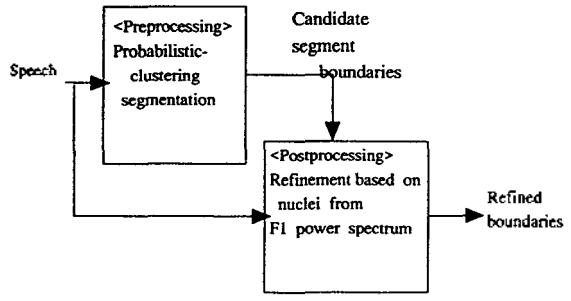

그림 1. 음절핵의 위치정보를 이용한 음소경계 추출
Fig. 1. Phonetic segmentation based on syllabic nuclei extraction.

본고의 구성은 다음과 같다. II장에서는 통계적인 거리척도와 클러스터링 방법을 이용한 전처리 과정에 대해 설명한 다음, III장에서는 음절핵을 추출하고 이를 이용하여 최종적인 음소경계를 선택하는 방법에 대해 기술한다. 이어서 IV장에서는 제안하는 방법의 유용성을 보이기 위한 실험결과에 대해 기술한 다음 V장의 결론으로 끝맺는다.

II. 전처리 : 클러스터링을 이용한 후보 음소경계 추출

Eberman과 Goldenthal[11]은 Andre-Obrecht[12]가 제안한 시간 영역에서의 통계적인 세그먼테이션 방법의 단점을 보완하여 LPC영역에서 GLR(generalized likelihood ratio)을 거리척도로 하는 클러스터링 방법을 제안하였으며, 비교적 우수한 성능(TIMIT 코퍼스 이용, 25msec 오차범위내에서 누락오류 6.2%, 삽입오류 100%)을 가짐을 보였다. 본고에서는 이들이 제안한 방법을 바탕으로 음소경계 추출의 전처리를 수행한다. 먼저 클러스터간의 거리 척도에 대해 기술하고 나서 클러스터링에 대하여 설명한다.

2.1. 클러스터간의 통계적인 거리 척도

일반적인 LPC 모델은 식 (1)과 같이 나타낼 수 있다.

$$s[n] = \sum_{i=1}^{Q} a_i\, s[n-i] + \nu[n] \qquad (1)$$

여기서 s[n]은 음성샘플, Q는 LPC 예측계수의 차원이며, $\nu[n]$은 평균이 0, 분산이 V인 백색 가우스 프로세스로 모델링될 수 있는 LPC 예측오차이다. 이 모델에서 LPC 파라미터 $\theta = (\{a_i\}, V)$가 주어졌을 때, 일련의 음성 시퀀스 s_1^r (= s[1]...,s[r])의 발현확률 $L(s_1^r \mid \theta)$은 식 (2)와 같이 표현될 수 있다.

$$L(s_1^r \mid \theta) = \prod_{t=1}^{t=r} p(\nu[t] \mid s_{t-Q}^{t-1},\ \theta) \qquad (2)$$

주어진 음성구간에서 식 (2)의 $L(s_1^r \mid \theta)$은 LPC 모델 θ의 변수이며, $L(s_1^r \mid \theta)$을 최대화시키는 θ를 θ^M이라 하고, 이 때의 $L(s_1^r \mid \theta)$을 식 (3)과 같이 $L^M(s_1^r \mid \theta)$이라 하자.

$$L^M(s_1^r \mid \theta) = \max_{\theta} \prod_{t=1}^{t=r} p(\nu[t] \mid s_{t-Q}^{t-1},\ \theta) = \prod_{t=1}^{t=r} p(\nu[t] \mid s_{t-Q}^{t-1},\ \theta^M) \qquad (3)$$

이제 연속적인 두 개의 클러스터 C1 = s_1^r, C2 = s_{r+1}^n가 주어졌을 때 이들 사이의 거리 d(C1,C2)는 식 (3)을 이용하여 식 (4)와 같이 GLR (generalized likelihood ratio)로 표현할 수 있다.

$$d(C1, C2) = \frac{L^M(s_1^n|\theta')}{L^M(s_1^r|\theta_1)\ L^M(s_{r+1}^n|\theta_2)} = \frac{L(s_1^n|\theta'^M)}{L(s_1^r|\theta_1^M)\ L(s_{r+1}^n|\theta_2^M)} \quad (4)$$

이 식에서 θ_1, θ_2, θ'은 각각 클러스터 C1, C2, 그리고 C1과 C2를 하나의 클러스터로 간주했을 때의 LPC 모델이다.

식 (4)의 d(C1, C2)를 계산하기 위해서는 최적의 θ_1^M, θ_2^M, θ'^M의 파라미터들, 즉 각 LPC모델들의 차원 Q1, Q2, Q3, 그리고 분산 V1, V2, V3을 구해야 한다. Goldenthal 등은 분산을 추정하기 위해 LPC 계수대신 Pacor계수를 사용하였으며, 최적의 Q를 찾아내기 위해서는 Q의 값을 변화시켜 가면서 여러 번에 걸쳐 반복 계산해서 구하였다. 일반적으로 LPC계수의 차원 Q가 커질수록 예측오차는 작아지므로 최적의 Q는 허용된 범위내에서 최대값을 가져야 하고, 따라서 최적의 Q를 구한다는 것은 의미가 없어진다. 이 점을 보완하기 위해 이들은 코딩 비용(coding cost)을 도입하고 그 값을 추정하여 Q의 값이 커질수록 비용도 증가하게 함으로써 Q의 값이 허용된 최대값으로 커지는 것을 막았다.

본 연구에서는 식 (4)의 거리척도를 이용하되, 분산은 Goldenthal 등과 같이 Parcor를 이용하여 구하고 LPC 계수의 차원 Q는 실험을 통해 추정하여 고정시킴으로써 계산량을 감소시킨다. Q의 값을 고정시킴으로써 발생하는 오류는 큰 영향을 미치지 않을 곳으로 예상하여 본 연구에서는 무시하였다.

2.2. 클러스터링

2.1절에서 기술한 거리척도를 이용하여 주어진 음성구간의 모든 인접한 클러스터들 사이의 거리를 구한 다음, 거리가 가장 가까운 한쌍의 클러스터를 찾아서 그 거리가 임계치보다 작으면 하나의 클러스터로 합친다. 새로 만들어진 클러스터에 대하여 이것과 인접한 클러스터들 사이의 거리를 구함으로써 이웃하는 클러스터들 사이의 거리를 갱신한 다음, 임계치보다 작은 거리를 갖는 클러스터 쌍이 없어질 때까지 이 과정을 반복한다. 이 때 임계치는 실험을 통하여 구하였으며, 이 과정을 통해서 얻어지는 음소의 경계를 후보 음소경계로 한다.

III. 후처리 : 음절핵의 추출 및 음소경계의 결정

3.1. 포만트의 검출

여러 가지 방법의 포만트 추출 알고리즘들이 개발되어 있으며 [7,8], LPC (linear predictive coding)를 이용한 방법도 많이 연구되고 있다. 특히 Welling[9, 10] 등이 제안하는 LPC와 다이나믹 프로그래밍을 이용한 포만트 추출방법은 다른 방법에서 필요로 하는 정점 검출(peak-picking), 혹은 다항식의 해를 찾아내는 과정 등을 필요로 하지 않는 효율적인 방법이다. 본고에서는 이들이 제안한 모델을 이용하여 포만트 및 포만트별 에너지 밀도를 추정한다. 이 모델에서 Welling 등은 한 프레임의 음성신호에 대한 모든 주파수 영역을 미리 정해진 K개의 주파수 영역으로 나눈 다음, 각각의 영역을 식 (5)와 같이 2개의 LPC 계수 α_k와 β_k를 갖는 2차원 디지털 공명기로 모델링한다 (k = 1..,K).

$$H_k(e^{j\omega}) = \frac{1}{A_k(e^{j\omega})} = \frac{1}{(1 - \alpha_k e^{j\omega} - \beta_k e^{j2\omega})} \quad (5)$$

이 때 이 공명기의 k번째 주파수 영역의 에너지 밀도 스펙트럼 $|H_k(e^{j\omega})|^2$는 식 (6)과 같이 구해지는 포만트 주파수 F_k에서 최대값을 갖게 된다.

$$F_k = \arccos\left(-\frac{\alpha_k(1-\beta_k)}{4\beta_k}\right) \quad (6)$$

따라서 한 음성 프레임의 포만트들은 각각의 주파수 영역에 대한 공명기의 오차 E_k의 합을 최소화하는 F_k들로 구성이 되는데, 다이나믹 프로그래밍을 이용하여 F_k를 추출해 낼 수 있다. 여기서 E_k는 k번째 주파수 대역에서의 음성신호 s[n]과 LPC의 출력 (α_k s[n-1] + β_k s[n-2]) 사이의 자승오차로, Parseval의 관계식을 이용하여 식 (7)과 같이 정의된다.

$$E_k = \frac{1}{\pi}\int_{F_{k-1}}^{F_k} |S(e^{j\omega})|^2\ |A(e^{j\omega})|^2 d\omega \quad (7)$$

우리말에서 각 음절은 초성(onset), 중성(nucleus) 및 종성(coda)로 구성되며, 중성을 음절핵 (syllabic nucleus)이라고도 부르는데, 중성의 자리에는 모음 하나만(이중모음 포함)이 나올 수 있다[2]. 이 음절의 핵에서 음향 에너지는 국부적으로 최대인 경우가 많으므로[3], Mermelstein은 음성의 에너지 밀도 스펙트럼에 주파수 대역별로 가중치를 준 에너지 밀도 분포를 이용하여 영어음성을 음절 단위로 나누었다[4]. Liu는 에너지 밀도 스펙트럼을 모두 6개의 주파수 영역으로 나누어 음향학적인 특이점(landmark)을 추출한 다음, 성문의 진동 유무를 검출하기 위해 0 ~ 400Hz사이의 에너지 밀도 스펙트럼을 이용하였다[5]. 한편 Howiit[6]는 Mermelstein의 음절 경계 추출 알고리즘의 성능을 향상시키기 위한 몇 가지 실험을 하였는데, Mermelstein이 이용한 광대역 에너지 밀도 스펙트럼 대신에 F1 (1st Formant)주위의 에너지 스펙트럼을 이용함으로써 우수한 성능을 얻을 수 있음을 확인하였다. 본고에서는 Howiit의 연구결과를 토대로 F1까지의 에너지를 이용하여 음절핵을 추출한다. 이를 위해 먼저 프레임별로 포만트를 검출한 다음, F1까지의 에너지 밀도로부터 음절핵을 추출해 내고 추출된 음절핵을 기준으로 후보 경계 중에서 음소경계를 선택한다.

3.2. F1 에너지 밀도 컨벡스 홀

F1의 에너지 밀도 분포는 각 프레임의 푸리에 에너지 밀도 스펙트럼(Fourier power spectrum)으로부터 0Hz ~ F_1 Hz 까지의 에너지 밀도를 합하여 구한다. 이 에너지 밀도 윤곽선에 Mermelstein이 제안한 컨벡스 홀(convex-hull) 알고리즘을 이용하여 에너지 밀도 분포의 국부적인 최대점과 최소점을 찾아낸 다음, 국부적인 최대점을 음절의 핵으로 간주한다. 컨벡스 홀은 그림 2에 보이는 것과 같이 에너지 밀도 분포를 둘러싸는 선으로 국부적인 최소점에서 최대점까지는 단조증가하며, 국부적인 최대점에서 최소점까지는 단조 감소하는 선으로, 두개의 국부적인 최소점 사이가 하나의 (음절)영역이 된다. 이때 이웃하는 정점과 골 사이의 에너지 밀도의 차가 임계치보다 큰 곳을 국부적인 최소점으로 간주하는데, 이 에너지 밀도의 임계치는 실험적으로 구해진다. 그림 2에서 H1과 H3는 임계치보다 크고 H2는 임계치보다 적은 경우이다.

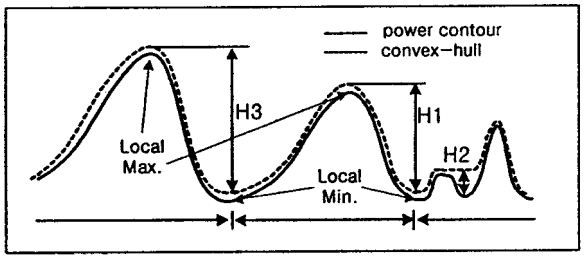

그림 2. 에너지 밀도의 컨벡스 홀과 국부적인 최대 및 최소점
Fig. 2. Convex-hull of F1 energy density function and it's local maxima and minima.

3.3. 최종 음소경계의 선택

전처리과정에서 구한 후보 음소경계로부터 최종적인 음소경계를 선택하는 방법의 예를 그림 3에 보인다. 음절핵 사이에 위치할 수 있는 최대 음소의 수가 3이라고 할 때 최대 음소경계의 수는 4개이다. 그림 3의 영역 R1에서 추출된 후보 경계는 모두 6개인데 이들 중 거리가 가장 큰 것부터 4개가 선택되어 최종 음소경계가 된다. 또 영역 R2에서는 2개의 후보 경계가 추출되었으며 이들은 최대 음소경계의 수보다 작으므로 2개 모두가 최종 음소경계로 간주된다.

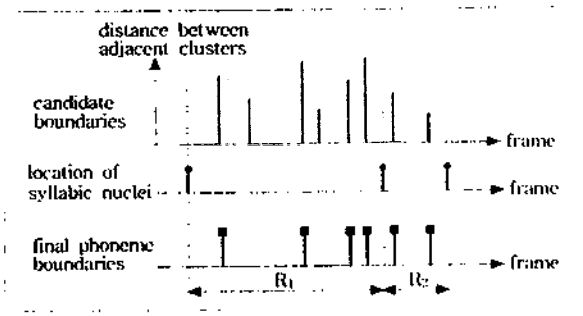

그림 3. 후보 경계로부터 최종 음소경계의 선택
Fig. 3. Selection of final phoneme boundaries from candidates.

IV. 실험결과

4.1. 음성 데이타

본 연구에서는 한국전자통신연구원의 POW 음성 데이터 중 음절의 수가 6개 이상인 음성샘플 240개를 무작위로 선택하여 이용하였다. 이 음성 샘플들은 16KHz, 16-비트 PCM으로 샘플링되어 있고 기준 음소경계는 본 연구실에서 수작업을 통하여 마련하였으며, 음소의 수는 모두 4240개이다. 음절핵의 위치를 찾아내는 실험에서는 240개의 음성 샘플중 100개는 임계치를 구하기 위한 학습용으로, 나머지 140개는 평가용으로 구별하여 사용하였다.

4.2. F1의 스펙트럼 에너지 밀도

음절핵의 위치를 구하기 위해 먼저 입력된 음성 데이터에 H(z) = 1-0.95z-1의 High Emphasis 필터를 적용한 다음, 18.75msec 길이의 해밍윈도우를 6.25msec 간격으로 적용하였다. 이 결과를 512-point FFT를 적용하여 에너지 밀도 스펙트럼을 구한 다음 3.1절에 기술한 방법으로 각 프레임마다 4개씩의 포만트 주파수를 구하였다. 아래 그림 4의 (a)와 (b)에 /가위 바위 보로/의 어절에 대한 음성 파형과 포만트를 보인다.

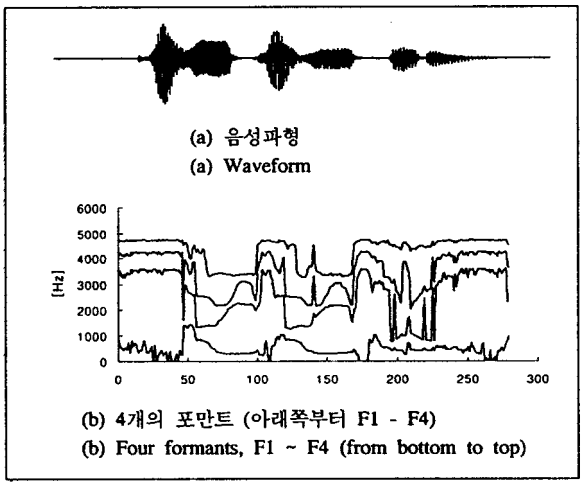

그림 4. /가위 바위 보로/의 포만트 추출
Fig. 4. Formant extraction example for /gaybayoro/.

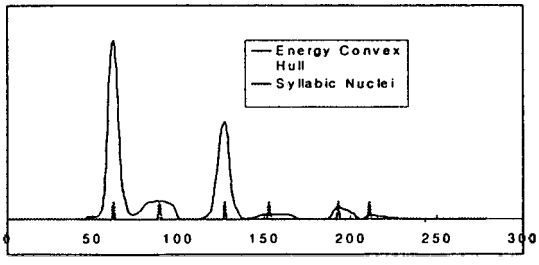

그림 5. F1 에너지 밀도의 컨벡스 홀과 음절의 핵
Fig. 5. Convex hull of F1 power and the syllabic nuclei.

4.3. 음절의 핵 추출

그림 4(b)처럼 구하여진 포만트들 중, 직류성분을 제외한 31Hz에서 F1까지의 에너지밀도 스펙트럼에 3.2절에서 기술한 방법을 적용하여 F1 에너지 밀도의 컨벡스 훌을 구하였다. 그런 다음에 컨벡스 훌의 국부적인 최대점을 찾아 음절의 핵으로 간주하였으며 그림 4에서 사용한 음성샘플에 적용한 결과를 그림 5에 보인다.

F1 에너지 밀도의 컨벡스 훌로부터 음절의 핵을 추출할 때 에너지 밀도의 임계치가 너무 크면 누락되는 음절의 핵이 많아지고, 너무 작아지면 삽입오류가 늘어난다. 본 고에서는 100개의 학습용 음성샘플을 정하여 임계치를 추출해 내었으며, 임계치가 정규화된 에너지 밀도의 단위로 0.02일 때 가장 좋은 결과를 얻을 수 있었다. 이 임계치를 구하기 위해 사용한 척도는 다음과 같다 :

1. 모든 모음 음소 각각에 대하여,
   a. 추출해낸 음절핵의 수가 0 이면 가중치가 1인 누락오류,
   b. 추출해낸 음절핵의 수가 1 이면 누락 및 삽입오류는 0,
   c. 추출해낸 음절핵의 수가 2, 3, 4 이면 가중치가 각각 1, 2, 3인 삽입오류,
2. 무성자음 및 휴지부(pause, stop)에 대하여,
   a. 추출해낸 음절핵의 수가 0 이면 누락 및 삽입오류는 0,
   b. 추출해낸 음절핵의 수가 1, 2, 3 이면 가중치가 각각 1, 2, 3인 삽입오류,
3. 유성자음에 대하여,
   a. 유성자음은 모음과 비슷한 F1 에너지 밀도를 갖는 경우가 있으므로 유성자음은 임계치 설정에서 제외. 즉, 삽입 및 누락 오류의 가중치를 0으로 하여 성능평가에 고려하지 않음.

이상의 척도를 적용하여 얻어진 삽입오류 및 누락오류의 합을 전체오류로 하여 이 전체오류를 최소화하는 값을 임계치로 정하였다.

아래의 표 1에 평가용 음성 샘플 140개에 대해 이 임계치를 적용했을 때의 결과를 보인다. 이 표에서 알 수 있듯이 F1의 에너지 밀도 컨벡스 훌을 이용한 음절핵의 추출률은 약 68.3%로, 더 정확한 방법의 개발이 필요하다. 반면에 무성자음 혹은 휴지부는 비교적 잘 제거됨을 볼 수 있다.

표 1. F1 컨벡스 훌을 이용한 음절핵의 추출 결과
Table 1. Results of syllabic nuclei extraction using F1 convex-hull.

핵의 수/음소 / 음소의 종류	0	1	2	3	4	삽입오류	누락오류
모음	254	680	55	5	2	71	254
무성자음, 휴지, 폐쇄	908	75	0	0	0	75	0
유성 자음	386	88	2	0	0	-	-

4.4. 음소경계의 추출

4.4.1. 전처리에 의한 음소경계 추출

2장에서 기술한 전처리에 의해 얻어진 음소경계 추출의 결과를 그림 6의 'Baseline'곡선으로 표시하였다. 한 프레임의 길이는 16.25ms로 하였으며, 간단한 실험을 통해 시험해 본 결과 Goldenthal[11]의 경험처럼 Parcor계수를 추출하기 전에 *High Emphasis*, 해밍윈도우 등의 처리는 하지 않는 것이 좋은 성능을 얻을 수 있었으므로 본 연구에서도 이 과정을 생략하였다. LPC계수 (Pacor계수)의 차원 Q는 16으로 고정하였으며, 클러스터링의 임계치를 변화시키면서 삽입 및 누락오류를 측정하였다. 삽입오류가 100%일 때 누락오류는 약 17%였으며 이 결과는 Goldenthal의 영어음성을 대상으로 한 실험결과 (삽입오류 100%, 누락오류 6.2%)와 비교할 때 누락오류가 약 2.7배가 되는데, 그 원인으로 허용 오차의 범위(25ms vs. 20ms), 우리말과 영어의 차이, 기준음소경계 추출을 위한 수작업 세그먼테이션의 정확성의 부족, 알고리즘의 변형 등을 생각할 수 있다. 이들 중 가장 중요한 원인으로 생각되는 것은 우리말과 영어의 차이로, 영어의 경우 스펙트로그램이나 음성파형에서 음소의 경계가 분명히 구별되는 경우가 많았으나 우리말의 경우는 그 경계가 분명치 않은 경우가 더 많았으며, 수작업을 통한 세그먼테이션 과정에서도 상당한 어려움을 겪었다. 이것은 수작업 세그먼테이션을 통한 기준 음소경계 설정의 부정확성으로도 해석될 수 있을 것이다.

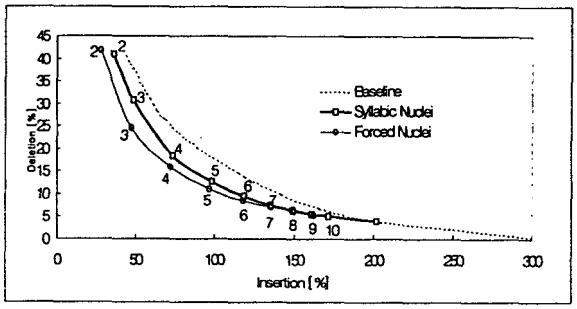

그림 6. 음소경계 추출 실험 결과
Fig. 6. Performance of phoneme boundary extraction.

4.4.2. 정확한 음절핵의 위치를 이용한 후처리

음절핵의 위치 정보가 어느 정도 효용성이 있는지를 시험하기 위해 '정확한' 음절핵의 위치를 제공한 다음, 3.3절에서 기술한 최종 음소경계의 선택 과정을 수행해 보았다. 수작업으로 구한 기준 음소경계와 전사(transcription) 정보를 이용하여 모음으로 구분된 기준음소경계 내에서 에너지 밀도가 최대가 되는 점을 찾아 '정확한' 음절의 핵으로 삼았다. 이 때 이용한 전처리 과정의 출력은 삽입오류 202%, 누락오류 3.9%일 때의 후보 경계이며, 이보다 낮은 누락 오류(높은 삽입오류)를 갖는 후보 경계를 이용할 때에는 좋은 결과를 얻지 못하였다. 이 이유는 전

처리과정에서 얻어지는 후보 경계에 너무 많은 삽입오류가 있는 반면에, 후처리 과정에서는 허용된 최대 음소의 수를 초과하는 음소경계만 제거하기 때문으로 분석된다. 정확한 음절핵을 이용한 실험 결과는 그림 6의 'Forced Nuclei'로 표시한 곡선이며 곡선 옆의 숫자는 음절핵 사이에서 허용한 최대 음소의 수이다. 클러스터링의 임계치를 파라미터로 하는 전처리과정의 결과인 'Baseline'의 곡선과는 파라미터가 서로 다르지만 삽입 대 누락오류의 성능을 비교하기 위해 하나의 도표에 표시하였다. 이 'Forced Nuclei'곡선에서 삽입오류가 96.5%일 때 누락오류는 약 11%로, 'Baseline'과 비교하면 약 35%정도의 삽입오류를 감소시킬 수 있었으며 음절핵의 위치를 이용하여 세그먼테이션의 성능을 향상시킬 수 있음을 확인할 수 있었다.

4.4.3. 추출된 음절핵의 위치를 이용한 후처리

4.4.2와 같은 조건에서, F1 에너지 밀도 분포에 컨벡스 훌을 적용하여 찾아 낸 음절핵의 위치를 후처리에 이용하였을 때의 결과를 그림 6의 'Syllabic Nuclei' 곡선으로 표시하였다. 우리말의 경우 연속하는 음절핵 사이에 위치할 수 있는 음소의 수는 0개 (모음-모음)에서 3개(모음-종성-휴지부/폐쇄-초성-모음)까지이므로 최대 음소의 수는 3개이다. 그러나 최대 음소의 수를 3개로 했을 때 그림 6에서 볼 수 있는 것처럼 누락오류가 약 30%나 된다. 과추출(over-segmentation)을 허용하여 음절핵 사이의 최대 음소 수를 5로 제한했을 때 삽입오류 98.2%, 누락오류 12.7%의 결과를 얻었으며, 'Baseline'의 성능과 비교할 때 약 25%정도의 삽입오류 감소 효과를 얻을 수 있었다. 3.3절에서 언급한 것처럼 음절핵 사이에 위치할 수 있는 최대 음소의 수인 4를 기준으로 했을 때 삽입오류, 누락오류는 각각 73.9%, 18.4%이며, 이 경우에도 'Baseline'의 성능 보다 약 25%정도의 삽입오류 감소효과를 얻을 수 있었다.

아래의 표 2에 음소경계추출 실험결과를 요약하여 나타낸다.

표 2. 후처리 방법에 따른 성능 비교
Table 2. Performance comparison of postprocessing methods.

Postprocessing Error	Baseline (no Postproc.)	Forced Nuclei	Syllabic Nuclei
Insertion Error	100	96.5	98.2
Deletion Error	17.1	11.0	12.7

V. 결 론

우리말은 음절 구조의 특성상 연속적인 음절핵 사이에 위치할 수 있는 음소의 수가 대부분의 경우 3개 이하로 제한된다. 본고에서는 이러한 우리말의 특성을 이용하여 연속음성의 맹목 세그먼테이션의 성능을 향상시킬 수 있는 후처리 방법을 제안하였다. 본 연구의 전처리 과정에서는 확률 통계적인 거리척도를 이용한 클러스터링 방법으로 후보 경계를 추출하였다. 후처리 과정에서는 입력 음성신호로부터 음절핵의 위치를 찾은 다음, 전처리 과정에서 추출된 후보 경계로부터 음절핵 사이에 위치할 수 있는 최대의 음소 수만큼 선택함으로써 삽입오류를 약 25% 감소시킬 수 있었다. 음절핵은 F1 에너지 밀도 분포의 국부적인 최대점을 찾음으로써 추출하였는데, 본고에서 사용한 척도로 약 68.3%의 정확성을 보였다. 정확한 음절핵의 위치를 제공한 실험에서는 약 35%정도의 음소경계의 삽입 오류가 감소됨을 볼 때 삽입오류의 감소율은 음절핵 위치의 정확성과 밀접한 상관관계가 있는 것으로 추측된다. 따라서 더 정확한 음절핵 추출방법의 개발이 필요하며, 단순한 음절핵의 위치 뿐 아니라 근래에 많이 연구되고 있는 음성의 특이점(landmark)이나 음절경계 등의 정보를 같이 사용하면 더 좋은 결과를 얻을 수 있을 것으로 기대된다.

참 고 문 헌

1. S. C. Lee et al., "Real-Time Probabilistic Segmentation for Segment-Based Speech Recognition," Proc. of ICSLP 98, Vol.5, pp.1803-1806, Sydney, Australia, Nov. 1998.
2. 이호영, 국어 음성학, (태학사, 1996), pp. 128-129.
3. 구희산 공역, 음성학과 음운론, (한신출판사, 1995), pp. 88.
4. Mermelstein, "Automatic Segmentation of Speech into Syllabic Units," Journal of Acoustic Society of America," Vol.58, N0.4, pp.880-883, Oct. 1975.
5. S. Liu, "Landmark detection for distinctive feature-based speech recognition," Journal of Acoustic Society of America," Vol.100, N0.5, pp.3417-3430, Nov. 1996.
6. A. W. Howitt, "Vowel landmark detection," Proc. of Eurospeech '99, vol. 6, pp.2777-2780, Budapest, Hungary, Sept. 5-9, '99.
7. G. E. Kopec, "Formant tracking using hidden Markov models and vector quantization," IEEE Trans. on Acoust., Speech, Sig. Proc., Vol. ASSP-34, No.4, pp.709-729, Aug. 1986.
8. R. C. Snell, et al. "Formant location from LPC analysis data," IEEE Trans on speech and audio processing, Vol. 1, No. 2, pp.129-134. Apr. 1993.
9. L. Welling, et al. "A Model for Efficient Formant Estimation," Proc. of ICASSP '96, pp.797-800, May, '96.
10. L. Welling and H. Ney, "Formant Estimation for Speech Recognition," IEEE Trans. on Speech and Audio Proc., Vol. 06, No. 1, pp. 36-48, Jan. '98.
11. B. Eberman et al. "Time-Based Clustering for Phonetic Segmentation," Proc. of ICSLP'96, pp.1225-1228, Philadelphia, Oct. 3 - 6, 1996.
12. Andre-Obrecht "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 36, No.1, pp. 29 - 40, Jan., 1988.
13. S. L. Wu et al. "Integrating Syllable Boundary Information into Speech Recognition," Proc. of ICASSP-97, vol.2,

pp.987-990, Munich.
14. "Syllable Onset Detection from Acoustics," Master Thesis, UC Berkeley, May 1997.

▲ 신 옥 근(Ok-Keun Shin)

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
※ 주관심분야 : 신호처리, 음성신호처리, 음성인식