論文96-33B-4-19

연속음에서의 각 음소의 대표구간 추출에 관한 연구

(A Study on Extraction of the Frames Representing Each Phoneme in Continuous Speech)

朴 贊 應 *, 李 |夬 熙 *

(Chan Eung Park and Kwae Hi Lee)

요 약

연속음성인식에서 음소를 인식대상으로 하였을 경우, 매우 적은 인식대상만으로도 무한단어인식시스템의 구현이 가능하다. 또한 연속음성인식 시스템에서 인식에 앞선 음소구간의 구분은 시스템의 복잡성을 대폭 줄일 수 있다. 그러나 인접 음소에 따른 상호 조음 현상 때문에 음소간에 정확한 경계를 추출하는 것은 매우 어렵다. 본 논문에서는 이러한 경계를 추출하는 대신 그 음소를 대표할 수 있는 대표구간을 검출하는 알고리듬을 제안한다. 음성 특징의 단기간 변화특성을 이용하여 특성최소변화구간과 특성최대변화구간을 추출한 후, 특성최소변화구간을 기준으로 한 장기간 변화특성들로부터 유성음간의 음소 변화를 추출하고, 특성최대변화구간은 음소의 천이구간 혹은 짧은 음소구간으로 간주하였고, 이들 짧은 음소구간의 중심구간과 특성최소변화구간을 각 음소의 대표구간으로 추출하였다. 본 연구에서는 비교적 적은 계산량을 갖는 켑스트랄 계수와 가중 켑스트랄 거리를 음성특징과 변화특성으로 사용하여 실험을 수행하였고, 실험용 데이터는 자체적으로 실내 환경에서 녹음한 음성 데이터를 사용하였다. 대표음소구간 추출 실험결과를 통하여 기존의 음소구분 알고리듬들에서 보다 적은 계산량으로 각 음소의 대표구간들이 높은 성능을 갖고 추출되는 것을 확인할 수 있었고, 이렇게 추출된 음소의 대표 구간들은 음소단위의 연속음성인식을 위하여 유용하게 사용될 수 있다.

Abstract

In continuous speech recognition system, it is possible to implement the system which can handle unlimited number of words by using limited number of phonetic units such as phonemes. Dividing continuous speech into the string of terms of phonemes prior to recognition process can lower the complexity of the system. But because of the coarticulations between neiboring phonemes, it is very difficult to extract exactly their boundaries. In this paper, we propose the algorithm to extract short terms which can represent each phonemes instead of extracting their boundaries. The short terms of lower spectral change and higher spectral change are detected. Then phoneme changes are detected using distance measure with this lower spectral change terms, and higher spectral change terms are regarded as transition terms or short phoneme terms. Finally lower spectral change terms and the mid-term of higher spectral change terms are regarded as to represent each phonemes. The cepstral coefficients and weighted cepstral distance are used for speech feature and measuring the distance because of less computational complexity, and the speech data used in this experiment was recorded at silent and ordinary in-door environment. Through the experimental results, the proposed algorithm showed higher performance with less computational complexity comparing with the conventional segmentation algorithms and it can be applied usefully in phoneme-based continuous speech recognition.

* 正會員, 西江大學校 電子工學科

(Electronic Engineering Dept., Sogang University)

接受日字:1996年2月26日, 수정완료일:1996年4月2日

Ⅰ. 서 론

지난 10여년 동안 음성인식 분야에서는 통계적인 패턴인식 접근방법, 혹은 신경망에 의한 방법 등에 의하여 괄목할만한 성과를 거두어 왔다. 그러나 아직까지도 대용량 어휘 연속음성 인식 시스템이나 음성-문자 변환기 등에 있어서는 풀어야할 많은 과제들을 갖고 있다. 그중 해결되어야 할 가장 큰 과제는 단어를 인식단위로 하는 인식 시스템은 하드웨어의 제한된 메모리와 계산속도로 인하여 이들 시스템의 실시간 처리 시스템의 구현을 어렵게 하고 있다는 것이다. 이에 대한 해결책으로 단어보다 더 기본적인 음성단위들, 다시 말하면, 음소 혹은 음절 등을 인식단위로하여 인식을 수행하는 연구들[1][2][3]이 활발하게 이루어지고 있다. 특히 40여개의 음소[4][5]를 갖는 한글에서는 이들 음소들을 인식단위로 사용함으로써 이러한 한계들을 극복할 수 있을 것이다.

이와 같은 음소단위의 인식방법에는 다음의 두 가지 접근방법이 있다. 그 첫째는 그림 1-(a)에서와 같이 음소구분과 인식을 동시에 수행하는 방법이다.

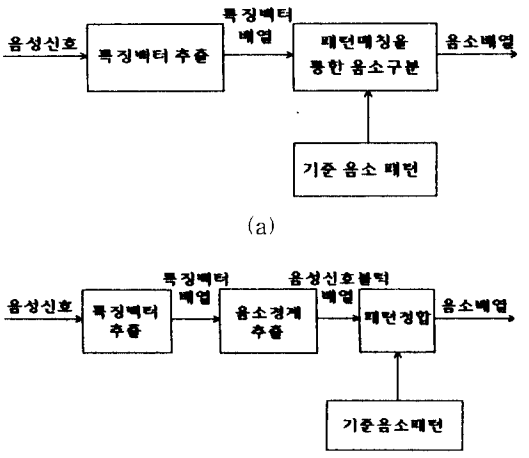


그림 1. 음소 인식 방법
(a) 음소구분과 인식의 동시 수행 (b) 음소구분 후 인식

Fig. 1. Phoneme recognition approach.
(a) Joint segmentation and recognition (b) Segmentation prior to recognition

그림 1-(a)에서와 같이 추출된 음성의 특징벡터의 배열을 사전 정보와의 정합을 통하여 음소의 배열로 변환하는 방법으로 최근에는 HMM과 신경회로망을 결합한 형태의 연구[6]가 활발히 이루어지고 있다.

두 번째 방법으로는 음소경계추출 후 음소 인식방법이 있다. 이 방법은 그림 1-(b)에서와 같이 우선 음성신호의 변화특성을 이용하여 음소의 경계를 추출한 후에 추출된 각각의 음성신호 블록들로부터 특징 벡터를 추출하고 기준 패턴과의 정합을 통하여 인식을 수행하는 방법이다. 본 논문은 두 번째 접근 방법에서 인식의 전 단계에서 수행하게 될 음소의 경계 추출에 관한 연구이다.

음소의 경계 추출은 Basseville, Andre-Obrecht들에 의하여 여러 가지 알고리듬들이 연구되었다. 주로 통계적인 접근방법을 사용한 이들 경계추출 방법은 Basseville와 Benveniste의 divergence test 방법[7]과 divergence test 방법의 문제점을 보완한 Andre-Obrecht의 forward-backward 방법[8]등이 대표적이라 할 수 있다. 이들 방법에서는 보다 정확한 음소의 경계점을 추출하기 위하여 음성신호의 블록단위처리보다는 샘플단위의 처리를 하고있으나 계산량이 과다하고 일부 음소들에 대하여는 음소의 경계를 잘 찾지 못하는 단점들을 갖고 있다. 이들 방법 이외에도 조정호들이 제안한 방법[9]도 있다.

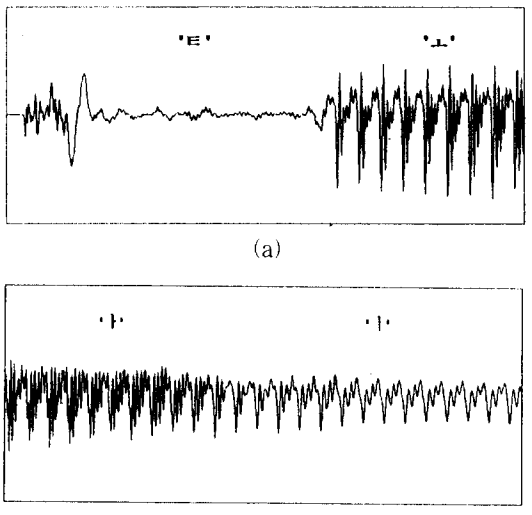


그림 2. 음소 경계의 예
(a) 무성자음과 모음 (b) 모음과 모음

Fig. 2. Examples of phoneme boundary.
(a) Unvoiced consonant and vowel (b) Vowel and vowel

그러나 이러한 경계추출은 그림 2-(a)에서의 예와 같

이 무성음과 유성음의 음소 사이에서는 뚜렷하게 나타나는 경우가 있지만, 그림 2 (b)에서의 예와 같이 유성음과 유성음 혹은 유성음과 무성음들 사이에서 상호간의 조음현상 때문에 정확한 경계를 찾기가 어려울 뿐만 아니라 이러한 경계부근의 천이구간에서의 음성샘플들은 음소인식을 위하여 별다른 의미를 갖지 못한다. 왜냐하면 이러한 천이구간의 음성샘플은 조음현상에 의하여 인접음소의 영향을 받으므로 인접음소에 따라 특성이 달라지기 때문이다. 따라서 이러한 음소의 경계를 정확히 찾는 것보다는 특성의 변화가 적은 안정된 대표구간을 찾는 것이 더 효율적이라 할 수 있다.

본 논문에서는 각 음소의 대표구간을 찾는 알고리듬을 제안한다. 우선 단기간 음성특징 벡터의 변화 특성을 이용하여 하나의 대표 구간을 찾은 다음, 이를 기준구간으로 하여, 다음 음소의 대표구간을 찾는다. 이 과정과 병행하여 단기간 음성특징 벡터가 급격히 변화하는 구간을 추출하여 이를 음소간의 천이구간 혹은 짧은 음소구간으로 간주하게 된다. 음성 특징 벡터로는 비교적 계산이 간단한 켑스트럴 계수를 사용하였고, 특징의 변화는 기준구간과의 가중 켑스트랄 거리를 이용하여 구하였다. 제안된 알고리듬을 일반 실내 환경에서 취득한 4명의 음성데이터를 대상으로 하여 실험을 수행하였다.

Ⅱ. 통계적인 방법을 이용한 음소의 구분

음성신호 y_n이 (1)식과 같은 형태의 통계적인 모델로 특징 지워진다고 가정한다.

$$y_n = \sum_{i=1}^{p} a_i y_{n-i} + e_n \tag{1}$$

여기서 e_n은 acoustic channel의 excitation이고, $var(e_n) = \sigma_n^2$ 인 0의 평균값을 갖는 가우시안 시퀀스라고 하면, 이 모델은 벡터 Θ로 표현되며 (2)식과 같이 정의된다.

$$\Theta^T = (\theta^T, \phi^T), \quad 단 \quad \theta^T = (a_1, a_2, \cdots, a_p) \tag{2}$$

여기서 ϕ는 σ_n 시퀀스에 의하여 결정되는 파라메터 벡터이다.

1. Divergence test

Divergence test는 (2)식에서 정의된 모델 Θ_0와 Θ_1의 거리를 그림 3에서와 같이 적절한 방법으로 측정함으로써 이루어진다. 음성 샘플의 시퀀스를 $Y_k^T = (y_1, y_2, \cdots, y_k)$라 하고, 그림 4에서 두 모델의 조건부 밀도를 $g_0(y_k|Y_{k-1})$와 $g_1(y_k|Y_{k-1})$이라 하면, g_0, g_1사이의 conditional cross entropy는 (3)식과 같으며 이것으로 두 모델간의 거리를 측정하였다.

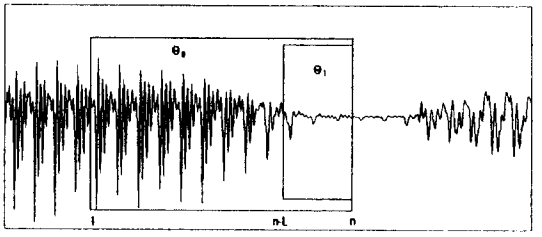


그림 3. Divergence test 방법
Fig. 3. Divergence test method.

$$w_k = \int g_0(y|Y_{k-1}) \log \frac{g_1(y|Y_{k-1})}{g_0(y|Y_{k-1})}\, dy - \log \frac{g_1(y|Y_{k-1})}{g_0(y|Y_{k-1})} \tag{3}$$

또한 가우시안의 경우는 (4)식과 같이 되며, 이것들의 누적합은 (5)식과 같이 W_n의 변화 특성에 따라 음소 경계를 추출한다.

$$w_k = \frac{1}{2}\left(2\,\frac{e_{0,k}e_{1,k}}{\sigma_1^2} - \left[1 + \frac{\sigma_1^2}{\sigma_0^2}\right]\frac{e_{0,k}^2}{\sigma_0^2} + \left[1 - \frac{\sigma_1^2}{\sigma_0^2}\right]\right)$$

$$단, \quad e_{q,k} = y_k - \sum_{i=1}^{p} a_{q,i} y_{k-i}, \quad q = 0,1 \tag{4}$$

$$W_n = \sum_{k=1}^{n} w_k \tag{5}$$

실제 수행에 있어서는 Page Hinkley rule[10]을 적용하여 (6)식의 조건을 만족하는 r을 음소의 경계로 추출하게 된다.

$$\hat{W}_n = \sum_{k=1}^{n} (w_k + \delta)$$
$$\max_{r:1 \le r \le n} (\hat{W}_r - \hat{W}_n) > \lambda \tag{6}$$

이때 유성음에 대하여는 $(\delta_v, \lambda_v) = (0.2, 40)$을 적용하고, 무성음에 대하여는 $(\delta_u, \lambda_u) = (0.4, 80)$을 적용한다. 그림 4에서는 divergence test의 예를 보여준다.

2. Forward-Backward divergence 방법

Divergence test에서 '모음 + 공명음'의 경계를 잘 찾지못하는 대신 '공명음 + 모음'의 음소경계는 비교적 잘 찾는 특성을 이용하여, 모음의 길이가 사전에 결정된 길이를 초과할 경우, 처리방향을 역으로 하여 di-

vergence test를 수행한다. 그림 5의 예에서와 같이 이전의 경계점까지 시험을 하여 새로운 경계점이 찾아지면 그것을 음소경계로 추가하고, 다시 정의 방향으로 시험을 계속하게 된다.

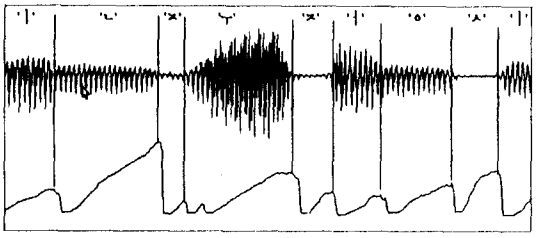


그림 4. Divergence test의 예
Fig. 4. Example of divergence test.

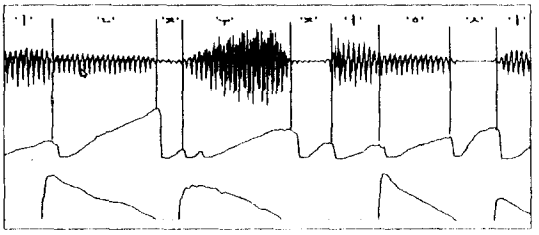


그림 5. Forward-backward divergence test
Fig. 5. Forward-backward divergence test.

3. '조정호'들의 방법

1989년 조정호들에 의하여 제안된 음소분할 알고리듬[19]에서는 그림 6에서와 같이 우선 quasi-stationary한 특성을 갖는 구간인 모델 M_0와 M_1을 찾은 후에 또 다른 모델 M_s를 모델 M_0에서 M_1까지 이동시키면서 모델 M_0와 M_s, 모델 M_1과 M_s의 주파수 특성 변화를 측정하여 음소의 경계를 추출하였다. 이 방법에서 기본적으로 모든 음소에는 quasi-stationary한 구간이 존재한다는 가정을 전제조건으로 하고 있다.

즉, 주파수 특성 변화 $SC(n)$은 (7)식으로부터 구해지며 여기서 d_{GN}은 이득 정규화 Itakura Saito 왜곡(gain normalized Itakura-Saito distortion)이다.

$$SC(n)=\frac{1}{2}\left[\ d_{GN}(\ |X(z)|^2;|G_0(z)|^2)\ -\ d_{GN}(\ |X(z)|^2;|G_1(z)|^2)\right] \qquad (7)$$

이때 주파수 특성 변화점은 모델 M_s와 모델 M_0, M_1의 거리가 같아지는 점, 즉 $SC(n)$이 최소가 되는 점을 (8)식에 의하여 구하게 된다.

$$\theta=\arg\min_{n}\ |SC(n)| \qquad (8)$$

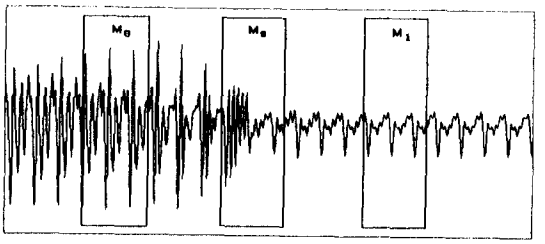


그림 6. 조정호들의 방법
Fig. 6. Cho's method.

III. 대표 음소 구간의 추출

그림 2-(b)의 예에서와 같이 유성음에 해당하는 음소들의 경계는 대부분의 경우 그 경계를 명확히 구분한다는 것은 대단히 어려운 일이다. 그리고 정확히 구분한다고 해도 경계점 부근의 천이구간에서의 음성신호의 주파수 특성은 상호 조음현상 때문에 인접 음소에 따라 달라지게 되고, 이러한 천이구간의 음성신호는 오히려 음소단위의 음성인식에서 인식율을 떨어뜨릴 우려가 있다. 따라서 이러한 인식율 저하 요인을 제거하고, 음소를 대표할 수 있는 음소구간을 추출하는 새로운 알고리듬을 제안한다. 물론 대부분의 무성자음들은 그림 7의 예에서와 같이, 특히 초성이나 종성에서, 이러한 안정된 구간을 갖지 못하여 그 주파수 특성의 변화가 천이구간과 유사하다. 그러나 이러한 주파수 특성의 변화구간의 위치, 길이들을 감안하여 주파수 특성의 변화가 큰 구간으로부터 무성음의 대표 음소구간의 추출이 가능하다.

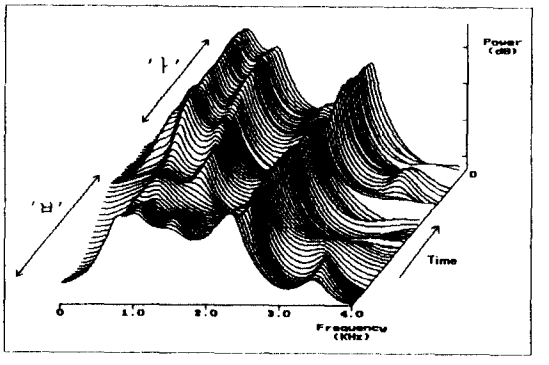


그림 7. 주파수 전력밀도 변화의 예
Fig. 7. Example of the change in power spectral density.

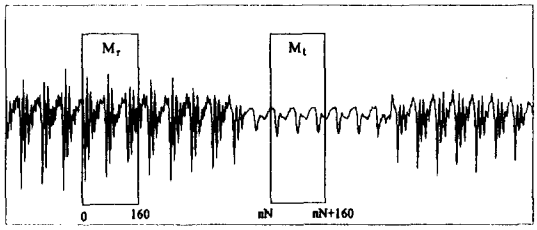
1. 대표음소구간 추출 알고리듬

제안한 방법에서는, 처리의 실시간성을 위하여, 가급적 적은 계산량으로의 구현과 역방향(backward)처리 배제 등이 고려되었다. 한 음소에서 다른 음소로 천이되는 구간에서는 주파수 특성의 변화가 크게 두 가지 다른 형태로 나타난다. 즉 모음과 모음 혹은 모음과 유성자음 사이에서는 변화가 서서히 이루어지며, 무성자음과 모음 사이에서는 급격히 변화가 이루어진다.

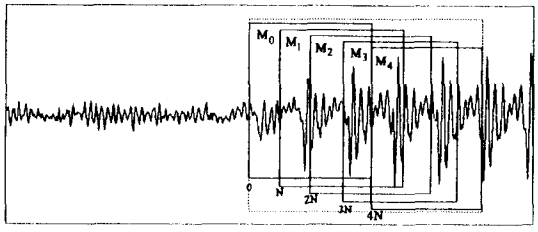
프레임 단위의 음성 신호 모델 M은 다음 식에 의하여 정의되는 특징벡터 Φ로 표현된다.

$$\Phi=(c_1, c_2, \cdots, c_n) \tag{9}$$

여기서 c_n은 켑스트랄 계수이다.



(a)



(b)

그림 8. 제안된 방법에서의 모델 위치
(a) 장기간시험 (b) 단기간시험
Fig. 8. Model position in proposed method.
(a) Long term test (b) Short term test

그림 8-(a)에서와 같이 모델 M_r의 위치를 결정하고, 모델 M_t를 N 샘플만큼 씩 이동시켜 가면서 모델 M_t와 M_r의 거리를 측정한다.

모델 M_r과 M_t의 거리를 $d_{long\ term}=d(M_r, M_t)$ 라 하고, $d_{long\ term} > \lambda_{ll}$이면 다른 음소로 바뀐 것으로 한다. 이 장기간시험은 주파수 특성의 변화가 매우 천천히 일어나는 모음과 모음 혹은 모음과 유성자음들 사이에서의 특성변화를 찾기 위한 것이다. 그림 9에 장기간시험 결과를 $d_{long\ term}$의 변화로 보였다.

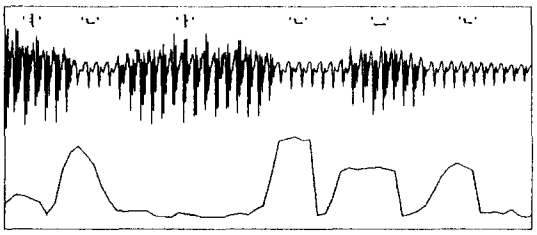


그림 9. 장기간시험
Fig. 9. Long term test.

또한 그림 10에서와 같이 짧은 구간을 N 샘플씩 이동하면서 구간 내에서 모델 M_0와 nN 샘플 거리의 모델 M_n들과의 거리를 각각 구한다. 그리고 이들의 평균거리와 최대거리 $d_{mean(st)}$, $d_{\max(st)}$는 각각 (10), (11)식으로 표현된다. 여기서 $d_{mean(st)} < \lambda_1$이면, 그 구간을

$$d_{mean(st)}=\frac{1}{n}\sum_{k=1}^{n} d(M_0, M_k) \tag{10}$$

$$d_{\max(st)}=\max_{k} d(M_0, M_k) \tag{11}$$

안정된 주파수 특성을 갖는 음소의 대표구간으로 간주하고, $d_{\max(st)} > \lambda_2$ 이면, 주파수 특성의 변화가 심한 음소의 천이구간 혹은 무성자음 구간으로 간주하게 된다. 이 단기간시험은 그림 10에서와 같이 주파수 특성의 변화가 급격히 일어나는 음소 천이구간 혹은 초성, 종성에서의 폐쇄음구간들을 찾고, 긴 구간동안 비교적 동일한 주파수 특성이 계속되는 모음, 유성자음들에서의 대표구간을 찾기 위한 것이다. 그림 10에 음소 변화에 따른 $d_{mean(st)}$, $d_{\max(st)}$의 변화도 보였다.

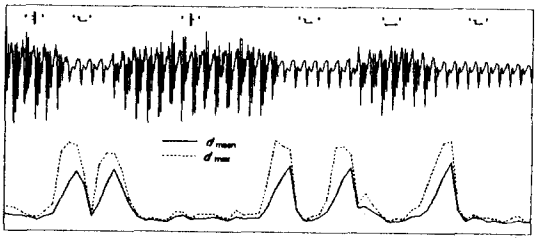


그림 10. 단기간 시험
Fig. 10. Short term test.

모델간의 거리의 측정은 켑스트랄 계수가 갖고 있는 무의미한 정보에 의한 variability를 조절하여 식별 신뢰성을 높일 수 있는 가중켑스트랄 거리측정법[11]을

사용하였다. 켑스트랄 계수는 (12)식과 같이 LPC 계수로부터 계산된 LPC 켑스트랄 계수를 사용하였다.

$$c(1) = -a(1)$$
$$c(i) = -a(i) - \sum_{k=1}^{i-1}(1-k/i)a(k)c(i-k) \qquad (12)$$
단, $1 < i \le p$

가중 켑스트랄 거리는 (13)식으로부터 얻는다. 여기서 h는 보통 $\frac{p}{2}$ 값으로 선택된다.

$$d_{wcep} = \sum_{n=1}^{p}[\ w(n)c_t - w(n)c_r\ ]^2$$

단, $w(n) = \begin{cases} 1+h\sin(\frac{n\pi}{p}) & \text{for } n=1,2,\ldots,p \\ 0 & \text{for } n\le 0,\ n>p \end{cases}$ (13)

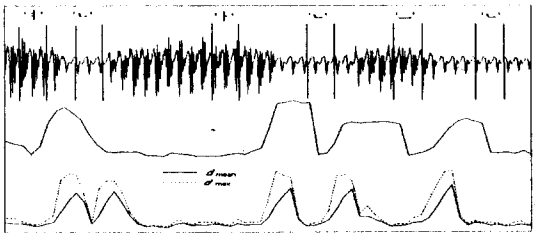


그림 11. 제안된 방법에 의한 음소 대표구간 검출
Fig. 11. Detection of representing period by proposed method.

2. 구현 방법

실제 구현을 위하여 10차의 LPC계수로부터 추출된 10차의 켑스트랄 계수를 모델의 특징벡터로 사용하였다. LPC계수는 8kHz로 표본 추출된 16비트 음성신호의 160샘플(20 ms)을 해밍 창틀(Hamming window)을 거친 후 Durbins recursive procedure[12]에 의하여 구하였다. 또한 장기간시험에서는 1프레임(160샘플)의 기준 모델로부터 5 ms(40샘플)씩 이동하며 모델간의 거리 $d_{long\ term}$을 측정하였고, 단기간시험에서는 짧은 구간(320샘플, 40ms)내에서 1프레임을 5ms씩 이동하면서 첫 프레임에 대한 거리를 측정하여 $d_{mean(st)}$와 $d_{max(st)}$를 구하고, 이 짧은 구간을 다시 5ms씩 이동하면서 반복 시행한다.

λ_0, λ_1, λ_2의 문턱값들은 실험에 의하여 다음과 같이 선택하였다.

$$\lambda_0 = 20.0,\quad \lambda_1 = 1.0,\quad \lambda_2 = 15.0,$$

대표구간 추출 절차는 다음과 같다.

i) $d_{mean(st)} < \lambda_1$인 짧은 구간을 찾고 이 짧은 구간의 중간 프레임을 해당 음소의 대표구간으로 하고 등록한다.

ii) i)에서 구한 대표구간을 기준구간으로 하여 40샘플(5ms.)씩 이동하면서, $d_{long\ term} > \lambda_0$인 프레임을 찾으면 음소가 바뀐 것으로 간주한다. 동시에 단기간시험을 통하여 $d_{max(st)} > \lambda_2$인 구간을 찾아 그 구간을 구간길이에 따라 음소의 천이구간 혹은 주파수 특성 변화가 큰 음소구간으로 한다. 음소구간으로 간주될 경우 그 구간의 중앙 프레임을 해당음소의 대표구간으로 등록한다.

iii) 단기간시험과 장기간시험을 통하여 음소가 바뀌거나, 천이구간이 나타나기 전에 현재 $d_{mean(st)}$ <기준 $d_{mean(st)}$, $d_{long\ term} < \lambda_0$인 구간을 찾으면 기준구간을 현재구간으로 대치하고, 대표구간 등록을 갱신한다.

iv) i)에서 iii)의 과정을 반복한다.

Ⅳ. 실험 결과

1. 실험 환경과 음성 데이터

음소 대표구간 추출 실험은 IBM PC PENTIUM 120 Mhz를 사용하였고, 전체 시스템의 구성도는 그림 12와 같다. 그림에서와 같이 입력된 음성 신호는 8kHz로 샘플링되어 16비트의 선형 PCM 데이터로 디지틀화 된다. 이 디지틀 음성 데이터는 에너지와 영교차율을 이용한 끝점검출[13] 부에서 음성구간을 찾은 후, 그 음성구간에 해당하는 데이터는 $H(z) = 1 - 0.95z^{-1}$의 프리엠파시스를 거쳐 10차의 LPC 분석부에 가해진다. 여기서 구해진 LPC계수로부터 켑스트럴 계수를 구한 후 이것들을 이용하여 음소 대표구간을 찾게된다.

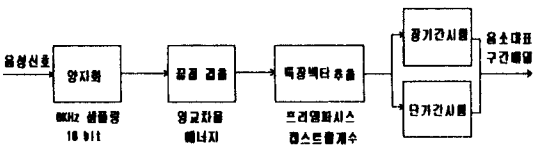


그림 12. 전체 시스템 구성도
Fig. 12. System block diagram.

실험에 사용된 음성 데이터들은 실내 환경에서 3명의 남자와 1명의 여자에 의하여 발음된 한국어 문장을 사용하였다. 사용된 문장은 국민교육헌장의 문장과 사

용 빈도가 적은 음소들의 경우수의 확보를 위하여 일부 문장을 추가하였다. 발음 속도는 초당 평균 약 4.3자(음절)의 속도를 갖도록 하였고 전체 음소 수는 1,475개이다.

2. 실험 결과 및 결과 분석

수집된 1,475개의 음소에 대하여 divergence test보다 우수한 성능을 갖는 것으로 알려진 forward backward divergence 방법과 본 논문에서 제안된 방법을 이용하여 실험한 결과를 표 1에 보였다.

표 1. 실험 결과
Table 1. Experimental result.

	Forward-backward divergence 방법	제안된 방법
총음소수	1475	1475
추출된 음소수	1980	1649
음소위치 오류	197	46
음소 누락수	220	172

가. 기존 알고리듬에서의 문제점

기존 알고리듬들의 문제점을 성능상의 문제와 실시간 처리상에서의 문제로 나누어 살펴보기로 한다.

1) 성능상의 문제

앞의 II장에서 언급된 알고리듬들을 한국어에 적용시 가장 커다란 문제점들 중에 첫째는, 이들 알고리듬들이 모두 각 음소는 주파수 특성이 stationary한 구간을 갖고 있다는 가정 하에서 출발하고 있다는 것이다. 그러나 자음 음소 중 폐쇄음 계열의 자음, 즉 ㅂ, ㄷ, ㄱ, ㅍ, ㅌ, ㅋ, ㅃ, ㄸ, ㄲ들이 무성음으로 어두 혹은 어미에서 발음될 때는 위에서 언급한 stationary한 구간을 대부분의 경우 갖지 못한다. 따라서 기존의 방법은 이러한 주파수 특성의 변화가 심한 구간들로 이루어진 음소들의 경계를 추출치 못하고 음소를 누락시키거나, 하나의 음소를 여러 개의 음소로 추출하는 오류를 범할 가능성이 높게 된다. 또한 조성호들에 의해서 제안된 알고리듬은 우선 quasi-stationary한 음소의 구간을 찾은 다음 두 개의 quasi stationary한 음소구간들 사이에 있는 경계점을 찾는 방법을 사용하고 있으므로 quasi-stationary한 구간을 갖지 못하는 음소의 경우에는 이를 누락시킬 위험이 있다.

둘째는, 모음과 모음 혹은 모음과 유성자음, 특히 공명음(ㄴ, ㅁ, ㅇ, ㄹ)들과의 사이에서는 주파수 특성의 변화가 매우 점진적으로 일어남으로 인하여 이러한 점진적인 변화구간이 경계점 추출에서 누락되거나, 또 하나의 음소로 과추출되는 오류가 자주 발생된다는 것이다.

셋째는, 파찰음(ㅅ, ㅈ, ㅊ, ㅆ, ㅉ)이나 폐쇄음 계열에서 유기음에 속하는 ㅍ, ㅋ, ㅌ들의 경우는 비교적 폐쇄음들에 비하여 음소의 구간이 기나, 이 구간 내에서 주파수 특성의 변화가 심하다. 따라서 이들 음소 구간에서는 하나의 음소를 여러 개의 음소로 과추출되는 오류가 발생할 우려가 있다는 것이다.

2) 실시간 처리상의 문제

음성인식에서 실시간 처리는 대단히 중요한 문제이다. 왜냐하면 대부분의 음성인식 시스템의 목적은 인간과 컴퓨터와의 대화형 정보교환을 목적으로 하고 있기 때문이다. 기존의 알고리듬들은 이러한 관점에서 몇 가지 문제점들을 갖고 있다.

첫째, divergence test 방법들에서는 정확한 경계점의 추출을 위하여 프레임단위의 시험보다는 음성신호의 각 샘플단위로 시험을 행하고 있다는 것이다. 그러나 앞에서 언급한 대로 음소 경계 부근에서의 음성신호 샘플들은, 음소간의 조음현상으로 인하여, 음소단위의 음성인식에서 유용성이 떨어지는 정보밖에는 갖고 있지 못하게 된다. 따라서 많은 계산량이 필요한 샘플단위의 시험보다는 프레임 단위의 시험이 실시간 처리 측면에서는 더 바람직하다고 할 수 있다.

둘째는 forward-backward divergence test의 경우나 조성호들에 의해서 제안된 알고리듬의 경우 시간축상의 역방향으로 시험을 하여야 하므로 연속음성처리를 위하여는 처리시간의 지연을 피할 수 없다는 것이다. 따라서 이것도 실시간 처리를 지연시키는 하나의 요인으로 작용한다.

나. 제안된 알고리듬의 시험결과 분석

표 1의 결과에서와 같이 제안한 방법은 forward backward divergence test의 경우보다 과추출된 음소의 수를 현저히 줄일 수 있었고, 음소 누락의 오류도 78.2%로 감소하는 결과를 얻었다.

음소 위치 오류의 경우 forward backward divergence test는 음소의 경계를 찾는 방법이고, 제안된 방법은 음소를 대표하는 구간을 찾는 방법이므로, 단순 비교는 곤란하나 상당한 오류 감소가 있는 것을 실험결과에서 알 수 있었다.

과추출의 경우는 표 2에서와 같이 모음이 종성으로 사용되는 경우 끝점 부근에서 가장 많이 발생하였고, 그 외에 파찰음에 해당하는 음소 등에서도 과추출이 관찰되었다. 음소대표구간 누락 오류의 경우는 표 3에서와 같이 모음과 공명음이 연속하여 나타나는 구간, 반모음 w, y가 포함되는 이중모음들 구간, 그리고 중성 자음구간들에서 주로 나타났다. 특히 반모음의 경우는 비교적 짧은 구간에서 주파수 특성이 점진적으로 변하는 경향을 가지므로, 제안된 방법을 통하여 대표구간 추출을 위한 전처리로 추출하는, 안정구간과 천이구간을 이용한 방법으로는 반모음구간에 대한 대표구간 추출 성능이 낮았다.

표 2. 과추출에 대한 분석
Table 2. Analysis of oversegmentation.

	과추출수	비율
종성 모음	106	60.9%
무성 자음	42	24.1%
기 타	32	15.0%
합 계	174	100%

표 3. 누락 오류에 대한 분석
Table 3. Analysis of omission.

	누락오류수	비율
모음 + 공명음	39	22.6%
반모음	42	24.4%
중성자음	29	16.9%
기 타	62	36.1%
합 계	172	100%

제안된 방법의 경우 알고리듬에 세 개의 문턱값 $\lambda_0, \lambda_1, \lambda_2$을 갖고 있다. 이 문턱값들을 낮게 지정하면 과추출 오류가 증가하고, 높게 지정하면 누락오류가 증가하게 된다. 따라서 이들 문턱값들은 신중하게 결정되어야 하며, 본 논문에서 제안된 값들은 많은 실험을 거쳐 실험적으로 결정되었다. 또한 이들 문턱값들을 이용하여 여자가 발음한 연속음성들에 대하여 실험한 결과, 남자음성의 경우와 마찬가지로 높은 성능의 음소 대표구간을 추출할 수 있었고 이로 미루어 제안된 방법이 남녀음성에 대하여 강건함을 확인할 수 있었다.

V. 결론 및 추후 연구 과제

기존의 방법들이 인접 음소간의 조음현상으로 인하여 경계점이 불명확하고, 경계점 부근에서의 음성의 특성은 음소단위의 인식에서의 기여도가 낮음에도 정확한 경계점 추출을 위하여 과도한 노력을 하고 있다.

본 논문에서는 연속음성인식에서, 제한된 최소의 인식단위로 음소를 사용하기 위하여, 연속 음성신호 중에서 각 음소들의 경계점을 찾는 대신 각 음소를 대표할 수 있는 대표구간을 찾는 방법을 제안하였다. 제안된 방법에서는 가중 켑스트랄 거리를 이용한 소구간 시험과 대구간 시험을 통하여 주파수 특성 변화가 적은 안정구간과 변화가 심한 천이구간으로 구분하여 이들로부터 음소들의 특성을 대표할 수 있는 대표구간을 추출하였다. 또한 처리의 실시간성을 높이기 위하여 비교적 간단한 계산으로 구할 수 있는 가중 켑스트랄 거리 측정법에 의하여 각 모델사이의 거리를 구하였다.

실험과정을 통하여 사용된 전체 음소 1,475개에 대하여 과추출율 11.8%, 추출누락율 11.7%, 추출위치불량비율 3.1%의 결과를 얻었다. 이 결과로부터 제안된 방법이 기존 방법에서보다 적은 계산량으로 높은 성능 향상을 보이는 것을 확인할 수 있었고, 특히 폐쇄음이 초성, 종성으로 사용될 때 기존 방법보다 탁월한 성능을 보였다. 이와 같은 결과들은 찾기 어려운 경계점을 찾기보다는 추출이 용이하고 경향이 뚜렷한 음소의 대표구간을 검출함으로써 얻어진 것이라 판단된다.

또한 이렇게 추출한 각 음소 대표구간의 특징벡터들을 연속음성을 고립음소들의 배열의 형태로 바꾸어 줌으로서, 각각의 고립음소들을 최소단위의 고립단어와 같이 취급할 수 있게 함으로써, 다음의 인식단계에서의 인식시스템의 구조를 단순화할 수 있으며, 높은 인식율을 기록하고 있는 고립단어 인식에 사용되는 HMM, 신경회로망 등의 방법을 유용하게 적용할 수 있게 되어 연속음성의 실시간 처리에 기여할 수 있을 것이다.

참 고 문 헌

[1] A. Waibel et al., "Phoneme recognition using Time-Delay Neural Networks," *IEEE Trans. on ASSP*, vol. 37, no. 3, pp. 328-339, March 1989.

[2] A. Ljolje and S. Levinson, "Development of

an Acoustic-Phonetic Hidden Markov Model for Continuous Speech Recognition," *IEEE Trans. on Signal Processing*, vol. 39, no. 1, pp 29-39, Jan. 1991.

[3] X. Huang, "Phoneme Classification Using Semicontinuous Hidden Markov Models," *IEEE Trans. on Signal Processing*, vol. 40, no. 5, pp 1062-1067, May 1992.

[4] 정연찬, *한국어 음운론*, 개문사, p 127-151, 1980

[5] 박창배, *한국어 구조론 연구*, (주)탑출판사, p 11-37, 1990

[6] N. Morgan and H. Bourlard, "Continuous Speech Recognition," *IEEE Signal Processing Magazine*, pp. 25-42, May 1995.

[7] M. Basseville and A. Benveniste, "Sequential detection of abrupt changes in spectral characteristics of digital signals," *IEEE Trans. on Inform. Theory*, vol. IT-29, pp. 709-723, Sept. 1983.

[8] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. on ASSP*, vol. 36, No. 1, Jan. 1988.

[9] 조정호, 홍재근, 김수중, "통계적인 방법에 의한 연결음의 음소분할 알고리듬," *전자공학회논문지*, 제 26권, 제 4호, pp.151-163, 1989년 4월

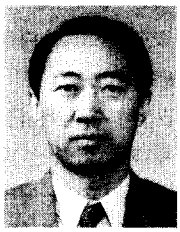
[10] D. V. Hinkley, "Inference about the change point from cumulative sum tests," *Biometrika*, vol. 58, no. 3, pp. 509-523, 1971.

[11] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, pp. 166- 171, 1993.

[12] S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*, pp. 97, 1985.

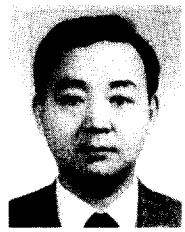
[13] L. Rabiner and M. Sambir, " An algorithm for determining the endpoint of isolated utterances," *BSTJ*, vol. 54, no. 2, pp. 296-315, Feb. 1975.

저 자 소 개



朴 贊 應(正會員)

1950년 8월 4일생. 1977년 8월 서강대학교 전자공학과 공학사, 1989년 8월 서강대학교 전자공학과 공학석사, 1992년 3월 ~ 현재 동 대학원 박사과정, 1992년 대우통신(주) 종합연구소 수석연구원, 1995년 3월 ~ 현재 인덕전문대학 방송통신과 전임강사. 주관심 분야는 음성신호처리, 영상신호처리, 통신시스템 등임



李 夫 熙(正會員)

1948년 8월 22일생. 1971년 2월 서울대학교 공과대학 전기공학과 공학사, 1973년 2월 서울대학교 대학원 전기공학과 공학석사, 1983년 2월 미국 남가주대학교 전기공학과 공학박사, 1983년 3월 ~ 현재 서강대학교 전자공학과 교수. 주관심 분야는 자동제어, 신호처리 등임