한국음향학회지 제20권 제8호 pp. 24~30 (2001)

음소 음향학적 변화 정보를 이용한 한국어 음성신호의 자동 음소 분할

Automatic Phonetic Segmentation of Korean Speech Signal Using Phonetic-acoustic Transition Information

박 창 목*, 왕 지 남**
(Chang-Mok Park*, Gi-Nam Wang**)

*아주대학교 대학원 산업공학과, **아주대학교 기계 및 산업공학부
(접수일자: 2001년 9월 14일; 채택일자: 2001년 10월 9일)

본 논문에서는 발음표기가 주어진 상황에서 음성 신호의 자동 음소 분할에 관한 것이며 음소의 경계를 음소 음향학적인 변화특성에 따라 3가지 형태로 분류하여 각각에 적합한 분할 알고리즘을 개발하였다. 형태 1은 묵음 · 유성음 · 무성음간의 분할이며 히스토그램분석으로 구한 문턱값으로 초기 분할 후, 웨이블릿 계수의 SVF (Spectral Variation Function)를 이용하여 분할하였다. 형태 2는 연속적인 모음의 분할이며 각 모음변화특성을 템플릿으로 구성하여 분할에 활용하였다. 형태 3은 모음과 유성자음 혹은 유성화 자음의 분할이며 특성 주파수대역의 진폭변화를 이용하여 후보구간을 정한 후, 켑스트럼 계수의 SVF를 이용하여 최종적인 분할을 수행하였다. 본 실험에서는 분할 성능을 테스트하기 위하여 한국어 PBW Speech DB에서 342개의 단어를 자동으로 분할한 후, 수작업으로 분할한 결과와 비교하였다. 전체적인 자동 분할 성능은 20 msec내에서 81.5%의 분할 성능을 보였다.

핵심용어: 음성 신호, 음소 분할, 음소 음향학적 특성, 웨이블릿
투고분야: 음성처리 분야 (2.4)

This article is concerned with automatic segmentation for Korean speech signals. All kinds of transition cases of phonetic units are classified into 3 types and different strategies for each type are applied. The type 1 is the discrimination of silence, voiced-speech and unvoiced-speech. The histogram analysis of each indicators which consists of wavelet coefficients and SVF (Spectral Variation Function) in wavelet coefficients are used for type 1 segmentation. The type 2 is the discrimination of adjacent vowels. The vowel transition cases can be characterized by spectrogram. Given phonetic transcription and transition pattern spectrogram, the speech signal, having consecutive vowels, are automatically segmented by the template matching. The type 3 is the discrimination of vowel and voiced-consonants. The smoothed short-time RMS energy of Wavelet low pass component and SVF in cepstral coefficients are adopted for type 3 segmentation. The experiment is performed for 342 words utterance set. The speech data are gathered from 6 speakers. The result shows the validity of the method.

Keywords: *Speech signal, Phonetic segmentation, Phonetic-acoustic characteristic, Wavelet*
ASK subject classification: *Speech signal processing (2.4)*

책임저자: 박창목 (cmpark@madang.ajou.ac.kr)
442-749 수원시 팔달구 원천동 산 5번지
아주대학교 대학원 산업공학과
(전화: 031-219-2429; 팩스: 031-219-1610)

I. 서 론

음성 신호의 음절 및 음소분할은 음성 인식, 음성 합성 등 음성 응용 시스템을 위한 음성 데이터베이스 구축에 기본적인 도구가 되고 있다. 그러나 이러한 음소 분할을 완벽하게 자동으로 수행하는 것은 매우 어려운 일이며, 심지어 전문가들의 수작업 결과들도 일치하지 않는 현상을 보인다. 이러한 이유는 음소들간의 상호 조음 현상, 단어 내에서의 음소특성변화 그리고 화자들간의 음소특성변화 때문에 음소 경계가 모호하기 때문으로 사료된다. 이러한 어려움 때문에 음소 경계에 대한 특성보다는 통계적으로 모델링된 음소 열과 분할 대상 음성과의 시간축 정렬에 의한 방법을 대부분 사용하고 있다. 이러한 방법 중 가장 대표적인 것이 HMM (Hidden Markov Models)을 이용한 통계적 패턴인식 방식이다[1]. HMM을 이용한 음소분할은 각 음소 모델을 학습하기 위한 데이터에 의존적이며, 그 학습 절차에도 많은 영향을 받기 때문에 강건한 분할 시스템을 만들기 위해서는 방대한 선행작업이 필요하며, 학습 데이터를 만들기 위한 음소 분할과정 또한 피할 수 없는 선행작업이다.

이에 본 연구에서는 주어진 발음기호정보, 음소의 음향학적인 특성과 그 변화시점에 대한 특성을 이용하여 음소 분할을 자동으로 수행할 수 있는 시스템을 제안하였다. 제안된 시스템은 먼저 형태 1을 분할한 후, 형태 1에서 추출된 유성음 구간에서 다시 형태 2에 대한 분할을 수행하고, 최종적으로 형태 3을 분할하는 절차를 통해 계층적인 음소분할을 수행하게 된다. 형태 1은 묵음, 유성음, 무성음간의 분할이며, 이산 웨이블릿 변환을 통하여 유성음에 해당되는 대역 신호를 추출한 후, 히스토그램 분석으로 문턱 값을 정하여 초기 분할을 수행하였다. 좀더 세밀한 분할을 위해서 웨이블릿 계수의 SVF (Spectral Variation Function)를 사용하였다. 형태 2는 연속되는 모음의 분류이며, 한국어의 모음을 17개로 분류한 후 모음변이 부분에서의 스펙트로그램 특성을 수집하여 템플릿을 구성하여 연속 모음의 분할을 수행하였다. 형태 3은 유성자음과 유성음의 분할이며 특정 주파수 대역의 에너지에서 유성음이 지역 최고점을 가지고, 유성자음은 지역 최소점을 나타내는 특성을 이용하여 각 경계가 존재하는 구간을 정한 후, 각 구간에서 켑스트럼 계수의 SVF를 이용하여 최종적인 분할을 수행하였다. 실험을 위하여는 한국어의 음성 데이터 베이스 (PRW Speech DB, PBW Speech DB)를 이용하여 성능 시험을 하였다. 실험 결과에서는 그 적용 가능성을 보여 주고 있다.

표 1. 한국어 음소의 분류
Table 1. Classification of phonemes.

분류 기호	음소 종류
V	모음
C1	유성 자음, 유성음화 자음
C2	파열음, 파찰음, 마찰음
S	묵음

표 2. 음소 변화 형태 (A: 모든 음소)
Table 2. Phonetic transition patterns.

변화형태 기호	형태 1	형태 2	형태 3
Cases	S/A A/S C2/V or C1 V or C1/C2	V/V	V/C1 C1/V C1/C1

II. 한국어의 음소 변화 패턴 분류

본 연구에서는 한국어의 음소를 4가지 (표 1)로 분류하였다. V에 해당되는 음소는 포만트 (200 Hz~4000 Hz)에 의한 특징을 가지고 있으며, C1에 해당되는 음소들은 낮은 주파수에 많은 에너지가 모여있는 것이 V음과 비슷하지만 V와는 다르게 포만트 특징이 명확하지 않은 특징을 보이거나 포만트의 대역폭과 진폭이 V와 다른 특징을 보인다. 이에 반하여 C2는 높은 주파수 대역 (4000 Hz 이상)에 에너지가 모여있는 특징을 가지고 있다.

결국 한국어 단어는 V, C1, C2, S의 연속적인 결합이라고 정리할 수 있으며, 음소 변화 형태는 3가지 (표 2)로 정리 가능하다. 형태 1인 경우는 저주파 대역과 고주파 대역간의 급격한 변화, 혹은 전체 주파수 대역에서의 급격한 변화로 특징 지워질 수 있으며, 형태 2인 경우는 포만트의 괘도변화, 그리고 형태 3인 경우는 특정 저주파 대역에서의 주파수 특성 변화로 특징 지워질 수 있다. 실제 한국어 발음에서의 대부분 음소 변화는 표 2와 같은 범주에 속하며, 예외적으로 C2/C2 (종성 자음/초성 자음)인 경우는 항상 그 사이에 짧은 묵음 구간이 존재하기 때문에 형태 1 범주에 속한다고 볼 수 있다.

III. 형태 1 분할 절차

3.1. 히스토그램 분석을 통한 초기 분할

형태 1 분할을 위하여 본 연구에서는 시간과 주파수 영역에서 음성 신호의 분해 성능이 좋은 것으로 알려진 DWT (Discrete Wavelet Transform)를 사용하였다[2]. 웨

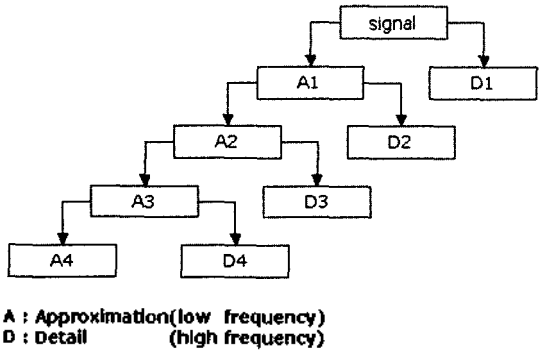


그림 1. DWT에 의한 신호 분해
Fig. 1. Signal decomposition using DWT.

표 3. 분해된 신호의 주파수 대역[2] (sampling rate=16 kHz)
Table 3. Frequency range of decomposition.

	주파수 대역 (Hz)
D1	4780 ~ 8000
D2	2914 ~ 4737
D3	1457 ~ 2369
D4	729 ~ 1184
A4	1 ~ 716

이블릿 계수는 10차 Daubechies를 사용하였으며, 그림 1은 DWT에 의한 신호 분해를 나타내며, 표 3은 분해된 각 신호의 주파수 대역을 나타낸다.

분해된 신호는 5 msec단위 (shift size=3 msec)의 프레임에 헤밍 윈도우를 씌운 후 각 프레임의 표준 편차를 구하고 메디안 필터를 통해 평활화하여 5차의 특징 벡터 $[d1\ d2\ d3\ d4\ a4]$를 구한다. 형태 1의 초기 분할은 음성 신호에서 유성음 부분 (V 와 C1), 무성음 부분 (C2), 묵음 부분 (S)을 추출하는 것이며, 본 연구에서는 유성음과 음성 구간을 추출하기 위해 $E1(=0.8a4+0.2d4)$과 $E2(=a4+d4+d1)$를 각각 정의하였는데, 그림 2에서처럼 유성음 구간에서는 $a4$와 $d4$신호에서 높은 진폭을 나타내고, 무성음 구간에서는 $d1$에서 높은 진폭을 나타내기 때문이다. 각 구간을 결정하기 위한 적절한 문턱치는 히스토그램 분석을 통해서 구할 수가 있다. 즉 그림 2에 보는 바와 같이 $E1$신호는 비 유성음 구간에서 일정하고 작은 진폭을 나타내며, $E2$신호는 묵음 구간에서 마찬가지로 일정하고 매우 작은 진폭을 나타내므로, 진폭에 대한 히스토그램에서 빈도수가 많은 진폭을 이용하여 각각의 문턱치를 구하여 대략적인 유성음 구간과 음성 구간을 구할 수 있다. 또한 음성 구간이면서 유성음 구간이 아닌 부분은 무성음 구간으로 결정하게 된다. 실제 본 연구에서는 0~1사이로 정규화된 $E1$과 $E2$의 진폭을 50등분하여 각 구간에서의 빈도 수를 구한 후, 최고 빈도를 가진 구간을 I 라고 할 때 α만큼 증가시킨 구간의 중심 값을 문턱치로 정하였다. α는 유성음의 추출을 위해서는 8, 음성의 추출을 위해서는 2를 선택하였다.

각 문턱치를 이용하여 유성음 구간과 음성 구간을 추출한 경우, 양단에 나타나는 에러를 없애주기 위해 다음과 같은 규칙을 순차적으로 적용하였다.

① 30 msec 이하인 묵음구간은 음성구간으로 병합한다.
② 30 msec 이하인 음성구간은 묵음구간으로 간주한다.
③ 30 msec 이하인 유성음 구간은 비유성음 구간으로 간주한다.
④ 30 msec 이하인 비유성음 구간은 유성음 구간으로 간주한다.

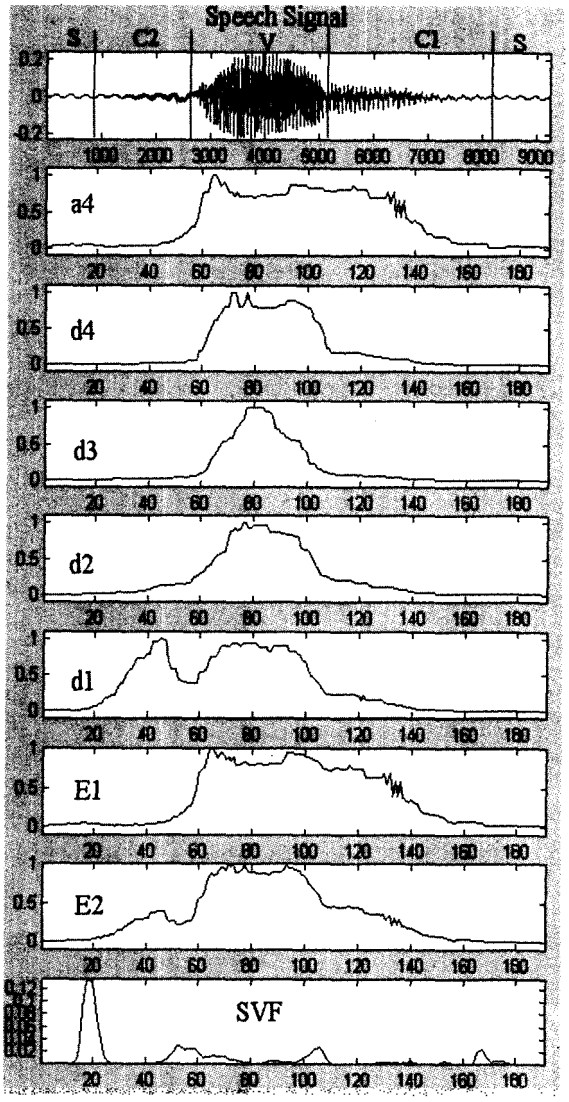


그림 2. 음성신호 /삼/, 5단계 분해, E1, E2 와 SVF
Fig. 2. Speech Signal/삼/, 5 decomposition, E1, E2 and SVF.

3.2. SVF를 이용한 정밀 분할과 발음기호정보를 통한 에러 수정

S_n ($n=1, \cdots, N$)을 N개의 음성특징벡터라고 할 때, SVF (Spectral Variation Function), F_n^q은 식 (1)로 구할 수 있다.

$$R_n = S_n - \overline{S}$$

$$F_n^q = \frac{1}{2}\left(1 - \frac{1}{q^2}\sum_{i,j=1}^{q} \frac{R_{n-i} \cdot R_{n+j}}{\|R_{n-i}\| \cdot \|R_{n+j}\|}\right) \quad (1)$$

SVF는 주위 프레임들간의 정규화된 거리를 표현하는데 유용한 방법이며, 그림 2처럼 각 음소의 경계에서 지역 최고점을 보여준다. SVF ($q=3$)에 사용되는 음성특징벡터는 DWT로 앞에서 구한 벡터를 사용하였으며, 이는 Wavelet변환의 시간/주파수 해상도의 장점을 활용하여 시간상 좀더 정밀한 SVF를 구할 수 있게 한다. 결국 앞에서 구한 초기 분할에서 좌우 15 msec 내의 이웃 프레임에서 가장 큰 SVF를 최종적인 음소 분할 지점으로 정하게 된다.

앞에서 음성 구간이면서 유성음 구간이 아닌 부분은 무성음 구간으로 간주하기 때문에 묵음 구간의 양쪽 끝에 잘못된 무성음 구간이 생길 수 있다. 이러한 구간을 수정하기 위하여 주어진 발음정보기호를 사용한다. 각각의 무성음이 초성인지 종성인지 파악하여 초성이면 다음 음절에 유성음 구간이 있어야 하므로 이에 위배되면 유성음 구간으로 병합하고, 종성이면 바로 전 음절에 유성음 구간이 있어야 하므로 이에 위배되면 유성음 구간으로 병합한다.

IV. 형태 2 분할 절차

4.1. 연속적인 모음경계의 특성

형태 1에서 분할된 유성음 구간에는 연속적인 모음 혹은 유성자음 (C1)과 모음들이 존재하게 된다. 이러한 유성음 구간에서 연속적인 모음을 분할하는 것이 형태 2 분할 절차이다.

앞에서 언급한 바와 같이 V-V인 경우는 포만트 (공명주파수)의 변화로 특징 지워질 수 있다. 이러한 변화는 다른 음소경계보다 비교적 느리며 그 변화속도를 예측하기 힘들어 그 경계를 구분하기가 수작업으로도 힘들고 상호조음현상 때문에 경계가 모호하다. 더구나 이중모음인 경우 한 모음 안에서 포만트의 변화가 있기 때문에 연속적인 모음 분할을 더욱 어렵게 한다.

본 연구에서는 비록 모음의 포만트가 각 사람마다 조금씩 다르고, 조음 환경에 따라 영향을 받지만, 두 개의 다른 모음이 변하는 구간의 형태는 항상 독특한 특징을 나타낸다는 사실에 착안하여 임의의 사람으로부터 이러한 특징 정보를 수집하여 연속적인 모음 분할에 활용하였다.

실제 한국어의 모음은 21개이나 서로 비슷한 모음을 묶어 17개로 축약할 수 있으며, 동일한 모음이 연속적으로 오는 경우는 하나의 모음으로 간주하게 되면 전체적으로 모음변화구간은 272개가 존재한다. 모음변화정보는 경계 부분 (100 msec 구간)의 스펙트로그램에서 4000 Hz 이하에 해당하는 2차원 벡터가 되며, 이는 모음의 포만트가 200~4000 Hz내 존재하기 때문이다[3]. 본 연구에서는 모음을 잘 모델링하며 계산시간이 비교적 빠른 선형예측분석을 통한 스펙트로그램을 사용하였다[4].

4.2. 상관함수를 이용한 템프릿 매칭 (templet matching)

본 연구에서는 분할대상 음성에서 모음변화템플릿과 유사한 특성을 가진 부분을 찾기 위하여 상관함수를 사용하였다. $f(x,y)$와 $g(x,y)$를 이산변수 x, y의 함수라 할 때, 두 함수의 상관함수 (cross correlation)는 식 (2)로 정의되어진다.

$$f(x,y) \circ g(x,y) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)g(x+m, y+n) \quad (2)$$

$F^*(u,v)$를 $f(x,y)$의 푸리에 변환의 공액복소수, $G(u,v)$를 $g(x,y)$의 푸리에 변환, $\mathfrak{I}^{-1}(\cdot)$를 역푸리에 변환이라 정의할 때 빠른 상관함수는 식 (3)으로 계산 가능하다.

$$f(x,y) \circ g(x,y) = \mathfrak{I}^{-1}(F^*(u,v)\ G(u,v)) \quad (3)$$

위 식을 활용하기 위해 $f(x,y)$는 모음변화템플릿으로 정의하고, $g_i(x,y)$, $1<i<K$,은 대상 음성 스펙트로그램에서 시간상으로 i번째 블록으로 정의하였다. x, $0<x<M-1$은 프레임 인텍스이고 y, $0<y<N-1$는 주파수 인텍스를 나타낸다. 상관함수를 사용하면 모음변화템플릿과 비슷한 특성을 가진 부분에서는 그림 3과 같이 중앙에 피크가 형성되는 상관함수를 나타내게 된다.

$$R_i = \frac{E(f(x,y)g_i(x,y)) - E(f(x,y))E(g_i(x,y))}{\sqrt{Var(f(x,y))}\sqrt{Var(g_i(x,y))}} \quad (4)$$

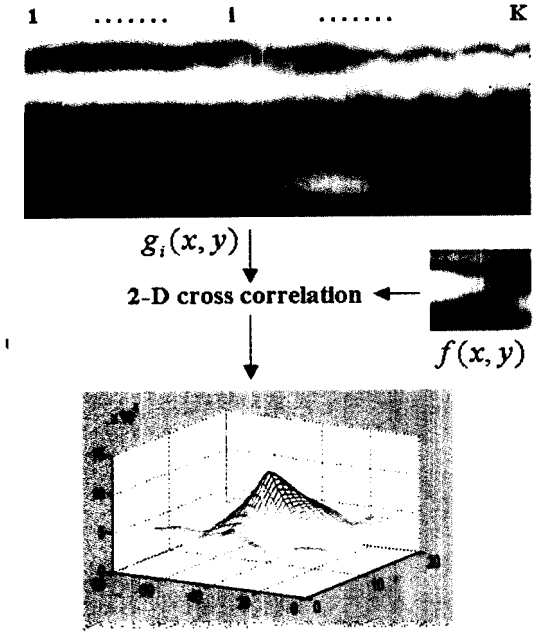


그림 3. 2차원 상관함수
Fig. 3. 2-dimensional cross correlation.

두 함수 사이에 정규화된 상관도를 계산하는 방법은 식 (4)로 정의되며, 실제 두 함수의 쉬프트정보를 모르는 상황에서 상관도 R_i는 식 (5)와 같이 근사화될 수 있다.

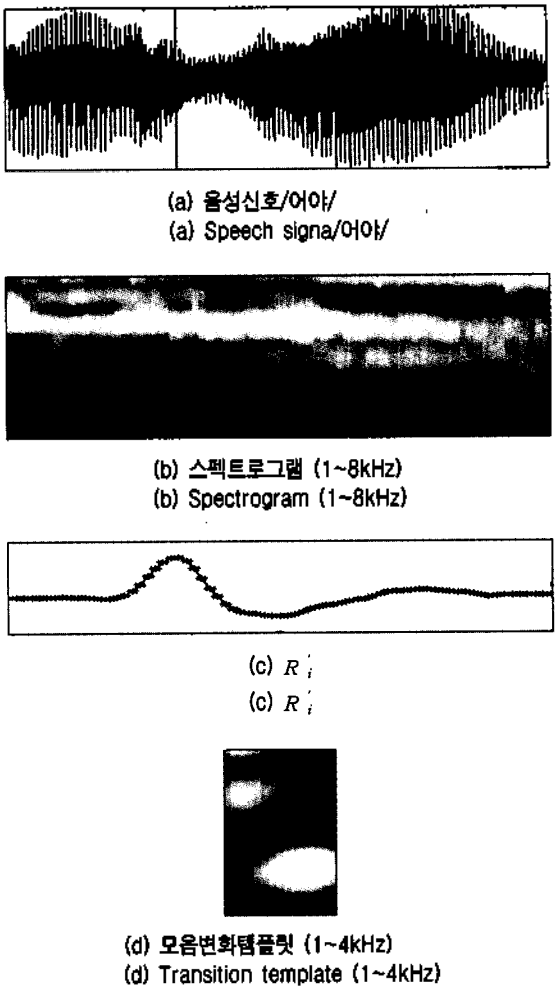
또한 피크가 중앙에 올 때 좀 더 완벽한 일치를 나타내므로 중앙의 첨도를 측정하는 δ를 사용하여 R_i'를 식 (6)에서 정의하였다. 식 (6)에서 $C_i(x,y)$는 i번째 상관함수를 나타낸다.

$$R_i \approx \frac{\frac{\max((f(x,y)\ast g_i(x,y))}{MN} - E(f(x,y))E(g_i(x,y))}{\sqrt{Var(f(x,y))}\sqrt{Var(g_i(x,y))}} \tag{5}$$

$$\delta = \sum_{n=\frac{N}{2}-5}^{\frac{N}{2}+5} C_i(M/2,n) - \sum_{n=\frac{N}{2}-5}^{\frac{N}{2}+5} C_i(M-1,n) \tag{6}$$

$$R_i' = R_i + \delta$$

연속모음분할 과정에서 K개의 순차적인 블록으로부터 $R_1', R_2', \cdots, R_K'$를 측정할 수 있으며 이 중 가장 큰 값을 보이는 블록이 모음경계부분임을 알 수 있다. 그림 4는 "어"와 "야"가 연속으로 오는 경우인데, "야"가 이중모음이기 때문에 포만트의 변화가 뒤쪽에 관측됨을 볼 수 있지만, (a)번 그림에서 수작업에 의한 모음경계와 비교해 볼 때, R_i'가 모음경계를 지시해 주고 있음을 알 수 있다.



(a) 음성신호/어야/
(a) Speech signa/어야/

(b) 스펙트로그램 (1~8kHz)
(b) Spectrogram (1~8kHz)

(c) R_i'
(c) R_i'

(d) 모음변화템플릿 (1~4kHz)
(d) Transition template (1~4kHz)

그림 4. 연속모음 분할 예
Fig. 4. Example of V-V segmentation.

V. 형태 3 분할 절차

형태 1 · 2 의 분할을 마치게 되면 C1과 모음의 경계혹은 연속적인 유성자음만 남게 된다. 연속적인 유성자음은 두 음절에서 "종성/초성"의 형태로 드물게 나타나는 바 본 논문에서는 하나의 유성자음으로 간주하였다. 결국 형태 3에서는 한국어의 ㄴ, ㅁ, ㅇ, ㄹ에 해당하는 유성자음, 유성화된 자음 (ㄱ, ㄷ, ㅂ, ㅈ, ㅎ)들을 모음과 구별하기 위한 절차가 필요하며, 이러한 유성자음들과 모음은 저주파대역에서 그 주파수 특성이 다르지만 그 경계가 모호하여 분할하기 힘들다. 특히 종성으로 사용된 "ㄹ"인 경우와 유성화된 자음은 더욱 어렵다.

그림 5에서 보는 바와 같이 C1에 해당하는 음소와 모음이 연속적으로 오는 경우 C1이 모음보다 포만트 주파수 폭이 좁아지고 포만트가 상실되며, 또한 진폭이 약해지는 특성을 가지고 있다. 특히 비음 (ㄴ, ㅁ, ㅇ)인 경우 1000

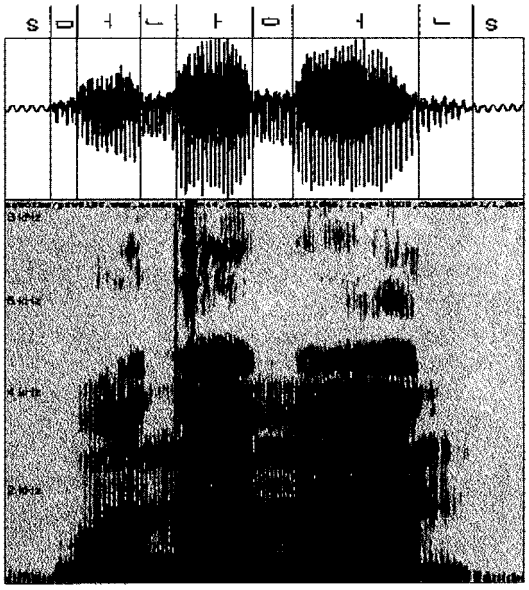


그림 5. 형태 3 분할과 스펙트로그램/머나먼/
Fig. 5. Type 3 segmentation and spectrogram/머나먼/.

Hz와 3500 Hz 근처에 영점을 가지고 있으며, 유음 (ㄹ)인 경우 3500 Hz 근처에 영점을 가지고 있다. 또한 모음사이의 유성음화된 자음은 700 Hz이내에 에너지가 집중되어 있다[5]. 이러한 특성은 음성 신호의 포락선의 궤도를 형성하며, 형태 1의 분할에서 구한 분해된 신호 중 D2와 D4에서 확연히 나타남을 볼 수 있다. 결국 이러한 특성으로 인해 D2 (2914~4734 Hz)와 D4 (729~1184 Hz)의 궤도를 보면 C1 음소에서는 지역 최소점을 보이고, 모음에서는 지역 최고점을 나타내는 것을 볼 수 있다. 결국 이러한 지역 최소점과 지역 최고점을 정확히 추출하면 1차적 분할을 위한 후보 구간들이 선정되고, 각 후보 구간에서 두 개의 음소로 분할하는 서브 문제로 축소된다.

5.1. 후보 구간 선정

후보 구간을 선정하기 위해 S (=D2+D4)를 생성한 후, 5 msec 단위 (shift size=3 msec)의 프레임에 헤밍 윈도우를 씌운 후 각 프레임의 표준 편차를 구하고 메디안 필터를 통해 평활화된 s를 구하였다. 이어서 s 미분을 구한 후 미분들의 영 교차점을 찾아서 지역 최소점과 지역 최고점을 찾으면 된다. s의 i번째 프레임의 미분치는 식 (7)과 같이 구해진다.

$$\Delta s_i = s_{i+t} - s_{i-t} \tag{7}$$

본 연구에서는 평활화 방법은 길이 20인 Median filter를 사용하였으며, 표본화주파수가 16 kHz일 때 t를 10으로 설정하였다. 영 교차점은 식 (8)과 같이 구하여진다.

$$ZR_i = (1 - sign(\Delta s_{i-1})sign(\Delta s_i))\ sign(\Delta s_{i-1}) \tag{8}$$

$$sign(x) = \begin{pmatrix} 1, & \text{if } x>0 \\ 0, & \text{if } x=0 \\ -1, & \text{if } x<0 \end{pmatrix}$$

기울기가 (+)에서 (−)로 이동하는 지점, 즉 지역 최고점은 ZR_i 가 2가 되며, 그 반대인 경우는 −2가 된다.

5.2. SVF를 이용한 분할

후보구간이 선정되면 각 구간 내에서 가장 높은 SVF를 가지는 프레임을 형태 3의 음소 경계로 결정하게 된다. 형태 3의 특성은 앞에서 설명한 바와 같이 포만트 변화에 많은 정보를 가지고 있으며 D2와 D4가 존재하는 주파수 대역인 1000 Hz와 3500 Hz 주위에서 많은 변화가 있다. 형태 3에서는 S로부터 SVF를 구하였다. 왜냐하면 발성의 마지막에 유성자음이 나타나는 경우 그림 5에서 보는 바와 같이 유성 자음 중간에 제 2 · 3 포만트가 소실되는 현상이 발생하여 음성 신호를 이용하여 SVF를 구할 경우 분할에 오류를 줄 수 있기 때문이다. SVF의 계산을 위해 유성음을 잘 모델링하는 것으로 알려진 켑스트럼 계수 (12 차수, window size=20 msec, frame size=4 msec)를 음성의 특징 벡터로 사용하여 SVF를 구하였다.

그림 6은 형태 3의 분할 예를 보여주고 있다. 점선으로 표시된 부분이 후보 경계를 나타내는 지역 최고점과 지역 최소점을 나타내며, 실선부분이 SVF를 이용한 최종 음소 경계를 나타내고 있다.

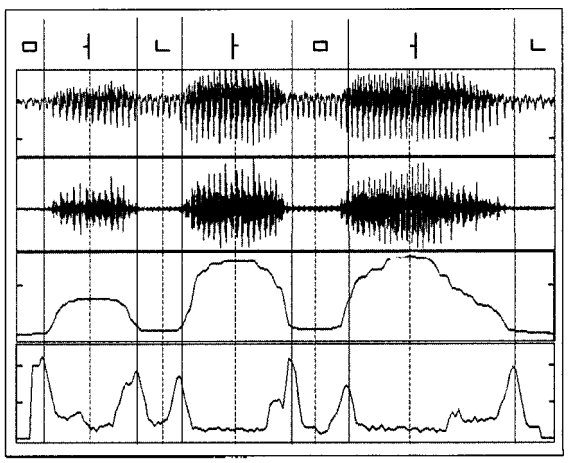


그림 6. 형태 3 분할예/머나먼/ (음성신호, S, s, SVF)
Fig. 6. Example of type 3 segmentation/머나먼/ (Speech Signal, S, s, and SVF).

표 4. 실험 결과
Table 4. Experimental results.

오차 범위/형태	형태 1 (%)	형태 2 (%)	형태 3 (%)
10 msec	92.0	80.2	67.4
20 msec	93.0	82.9	68.8
40 msec	96.3	88.5	72.6
60 msec	97.7	90.1	77.6

VI. 실험결과 및 결론

본 실험에서는 한국어 PRW Speech DB에서 20세 남자 1명으로부터 테스트 단어에 존재하는 154개의 연속 모음 변화 템플릿을 수집하였으며, 분할 성능을 테스트하기 위해서 PBW Speech DB에서 2음절, 3음절, 4음절 형태의 342개 단어를 대상으로 6명의 화자가 발음한 데이터를 사용하였다. 6명 화자의 구성은 10대 · 20대 · 40대 여자, 10 · 20 · 40대 남자로 구성되어 있다. 수작업으로 분할한 결과와 비교하여 형태별로 표 4와 같은 결과를 얻었다.

실험 결과에서 보는 바와 같이 형태 1의 분할 성능이 비교적 좋으며, 형태 3의 분할 성능이 가장 낮은 것으로 나타났다. 형태 2의 결과로부터 특성 화자로부터 수집한 모음 변화 정보를 이용하여, 여러 단어 · 여러 화자에 나타나는 모음 변화의 경계에 활용할 수 있다는 사실과 형태 1과 형태 3의 분할 알고리즘에 대한 활용 가능성을 입증할 수 있었다. HMM을 이용한 분할 성능은 20 msec내에서 전체적으로 79.9%의 성능을 보인다고 보고되고 있다[6]. 본 실험 결과에서는 동일한 에러범위에서 81.5%의 결과를 보이고 있어 HMM과 비등한 성능을 보이고 있음이 증명되었다.

참 고 문 헌

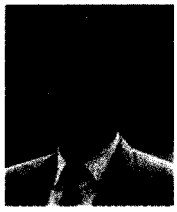
1. F. Brugnara, D. Falavigna and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357-31, 1993.
2. B. T. Tan, R. Lang, H. Schroder, A. Spray and P. Dermody, "Applying wavelet anaysis to speech segmentation and classification," *Wavelet Applications*, vol. Proc. SPIE 2242, pp. 750-761, 2, 1994.
3. L. R. Rabiner, B. H. Juang, "Fundamentals of speech Recognition," Prentice Hall, 1993.
4. T. Robinson, "Speech Analysis," Tony Robinson Lent term, 1998.
5. 지민제, "한국어의 조음 및 음향 음성학," 음성 신호처리 기술 제1권, 한국과학기술원/삼성 첨단기술센터 산학협동강좌, 1996.
6. 홍성태, 김제우, 김형순, "자동 음성분할 및 레이블링 시스템의 성능향상," 말소리, no. 35-36, pp. 175-188, 1998.

저자 약력

● 박 창 목 (Chang-Mok Park)

1996년 2월: 아주대학교 산업공학과 학사
1998년 2월: 아주대학교 산업공학과 석사
2000년 2월: 아주대학교 산업공학과 박사수료
2001년 2월: 마크애니 연구소 주임연구원
※ 주관심분야: 컴퓨터 비전, 패턴인식

● 왕 지 남 (Gi-Nam Wang)



1983년 2월: 아주대학교 산업공학과 학사
1985년 2월: 한국과학기술원 석사
1992년 12월: Texas A&M 대학 박사
현재: 아주대학교 기계 및 산업공학부 부교수
※ 주관심분야: 신경망, 시스템 감시 및 제어, 지능형 분산정보시스템