음성학적 지식과 DAC 기반 분할 알고리즘

구 찬 모[†]·왕 지 남[††]

요 약

본 논문에서는 음절이 잘 발달되어 있는 한국어에 대해서 신뢰할 수 있는 완전 자동화된 레이블링 시스템을 제안한다. 음운 및 음향학적인 정보를 최대한 이용하고 분할에러를 줄이기 위해서 조절 메카니즘의 하나로 DAC개념을 사용하여 음성을 *speechlet*으로 나누고 분할 된 음성 구간에 대해서 레이블링을 시도하는 *DAC*기반 분할알고리즘이다. HMM방법이 획일적이고 확정적인 성능을 갖는 반면 본 제안 방법은 음성학적인 특화지식을 컴포넌트로 개발 추가 계속 향상시킬 수 있는 프레임워크를 제시하고 있다는 점에서 주요 의의가 있다고 하겠다. HMM과 같은 통계학적인 방법을 이용하지 않고 음운학적, 음향학적 지식만을 이용하는 새로운 방법은 수행속도와 음성학적인 특화 지식컴포넌트를 확장함에 따라 일관성이 있으며 효과적 방법으로 적용가능 할 것이다. 제안 방법을 검증하기 위하여 실험결과를 제시하였다.

Phonetic Acoustic Knowledge and Divide And Conquer Based Segmentation Algorithm

Chan-Mo Koo[†] · Gi-Nam Wang[††]

ABSTRACT

This paper presents a reliable fully automatic labeling system which fits well with languages having well-developed syllables such as in Korean. The ASL System utilize DAC (Divide and Conquer), a control mechanism, based segmentation algorithm to use phonetic and acoustic information with greater efficiency. The segmentation algorithm is to divide speech signals into *speechlets* which is localized speech signal pieces and to segment each *speechlet* for speech boundaries. While HMM method has uniform and definite efficiencies, the suggested method gives framework to steadily develope and improve specified acoustic knowledges as a component. Without using statistical method such as HMM, this new method use only phonetic-acoustic information. Therefore, this method has high speed performance, is consistent extending the specific acoustic knowledge component, and can be applied in efficient way. we show experiment result to verify suggested method at the end.

키워드 : 자동분할 및 레이블링(Automatic segmentation and labelling), 신호 지역화(Signal Localization), 사례연구(Case Study), 분할 및 정복(Divide-And-Conquer)

1. 서 론

레이블링이란 연속된 음성신호로부터 음소단위 혹은 음절단위로 이루어지는 경계 점을 찾고 분할하여 분할 된 영역에 해당 음소나 음절을 표시 연계하는 작업으로 음성인식 혹은 음성합성에 기초적인 인프라 데이터를 작성하는데 이용된다. 해당 대부분의 레이블링 작업은 음성신호의 파형을 보여주는 툴을 이용하여 수 작업에 의한 레이블링 방법을 이용하고 있다. 따라서, 레이블링 방법을 체계화한 방법을 찾기가 어렵고 사람의 음향학적, 시각적 능력의 차로 인해서 사람에 의한 레이블링은 일관성이 많이 결여되어 있고 많은 시간이 소요되는 것이 일반적이다.

† 준 회 원 : 아주대학교 대학원 산업공학과
†† 정 회 원 : 아주대학교 산업공학과 교수
논문접수 : 2001년 10월 11일, 심사완료 : 2002년 1월 8일

ASL(Automatic Segmentation and Labeling)방법의 개발 도입은 레이블링 시간을 상당히 줄일 수 있으며 일관된 레이블링 결과를 줄 수 있어 지금까지 연구가 진행되어 왔다[3-5, 7, 8, 10]. 하지만, 아직까지 신뢰성에는 많은 한계를 보여주고 있으며 대부분의 ASL시스템은 HMM(Hidden Markov Model)과 같은 통계학적인 패턴인식접근법을 이용하고 있고 사람에 의한 수정작업을 병행하기도 한다[5, 9, 10].

본 논문에서는 음절이 잘 발달되어 있는 한국어에 대해서 신뢰할 수 있는 완전 자동화된 레이블링 방법을 제안한다. 음소의 음운학적, 음향학적인 정보를 이용하며 음소의 발음표기가 주어지면 주어진 음소 결합의 음운 및 음성학적 지식을 이용 음소의 경계를 찾아내는 방법을 제시한다. 음소의 경계를 찾는데 유사성을 측정하는 기법중의 하나인 SVF (Spectral Variation Function)을 이용한 방법을 또한 제시한다[4].

본 논문에서 제시한 레이블링 방법은 효과적인 조절 메카니즘으로서 DAC(Divide And Conquer)방법에 기반하고 있다. 먼저, 음성신호를 *speechlet* 이라고 불리우는 여러 개의 지역적 신호들로 분할하고 분할 된 신호들은 몇 가지 음운학적 사례를 포함하는 정의된 음향학적 패턴 중의 하나와 연계 대응시킨다. 각각의 음운학적 사례는 음향학적 지식을 포함하고 있으며 각 *speechlet*의 초기 음운학적 경계를 제공한다. 최종 음운학적 경계는 초기 음운학적 경계의 이웃들의 유사성 척도인 SVF를 통해서 결정하는 방법을 사용한다.

본 논문은 HMM과 같은 통계학적인 방법을 이용하지 않고 음운학적, 음향학적 지식을 이용하여 보다 수행시간에 빠른 속도로 진행가능하며 또한 잘 정의 된 음향학적 지식은 신뢰할 수 있는 DAC 기반 분할 알고리즘을 제시하는 방법으로 평가된다. 전체성능은 특화 된 음성 지식을 더욱 발견하여 추가한다면 계속 향상 될 수 있을 것이며 기존의 HMM방법이 획일적이고 확정적인 성능을 갖는 반면[5] 본 제안 방법은 음성학적인 특화지식을 컴포넌트로 개발 추가 적용하여 계속 향상시킬 수 있는 프레임워크를 제시하고 있다는 점에서 주요 의의가 있다고 하겠다.

2. 음운학적 가정

ASL 시스템에서 FBANK(Filter Bank), LPC(Linear Predictive Coding), 에너지, 켑스트럼 같은 다양한 스펙트럴 파라메터가 이용되는데 본 논문에서는 FBANK를 적용하였다. 먼저, 선택된 스펙트럴 파라메터에 대해서 신뢰할 수 있는 레이블링 방법을 설계 개발하기 위한 음향학적인 선결요건에 대해서 다음과 같이 정리 분석한다.

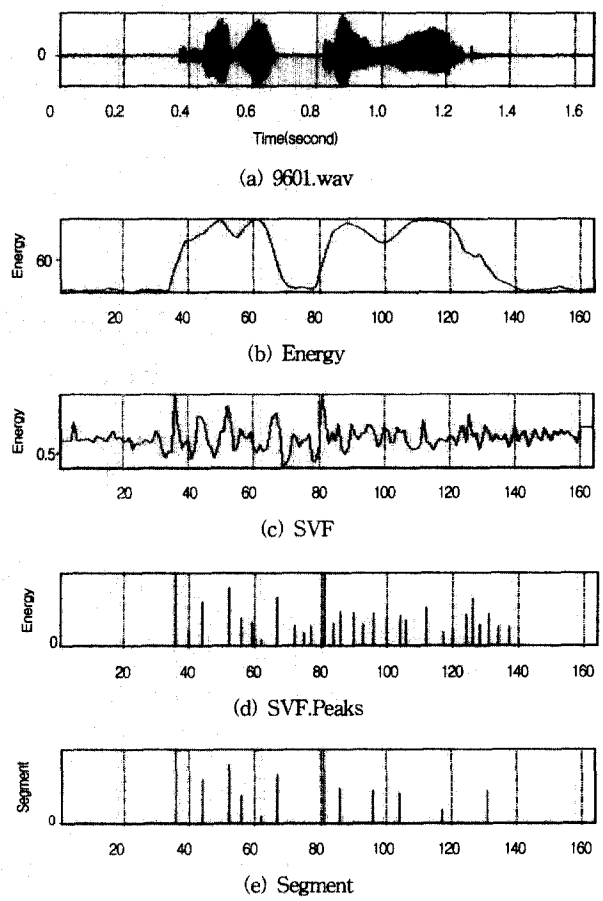
음향학적인 선결요건에 대해서 기존의 MLS (Multi-level segmentation)를 개발한 Glass[6]의 개념과 연계 될 수 있다.

가정 (1) : 음성은 반 안정적 음성 신호들의 시간적 연속이다.

가정 (2) : 음향학적 신호는 지역적 환경에 영향을 받아 변화되는 짧은 사건이다.

가정 (3) : 음성 신호는 음운학적, 음향학적인 정보를 이용해서 검출 가능한 부분들의 결합이다. 예를 들어, 음절이 발달된 언어에서 모음과 자음은 쉽게 구별이 된다.

가정 (1)은 같은 분할 영역 내에 있는 음성 벡터는 다른 분할 영역 내에 있는 음성 벡터들 보다 더 서로간에 유사하다는 의미이다. 따라서, 분할 문제는 근접한 프레임의 유사성에 의존하는 지역적 클러스터링문제로 표현된다. 유사성 측정의 수단으로 모든 프레임들은 이전의 그리고 이후의 프레임들과 비교되어진다. (그림 1)에서 여성에 의해서 발음이 된 "9601"(g-u r-yu-g2 gg-o-ng i-l)의 웨이브, 에너지, SVF, SVF 피크, 그리고 최종 음운학적 경계를 보여주고 있다. 해밍창을 사용했으며, 창크기는 250000, 채널의 수는 50, 목표율은 100000인 FBANK 스펙트럼을 이용했다.



(그림 1) 여성에 의해서 발음된 "9601(구육공일)"의 (a) 웨이브, (b) 에너지 수준, (c) SVF, (d) SVFG Peak, (e) 최종 음소 경계

가정 (2)의 결과로 음성 신호는 과분할 혹은 미분할 될 수 있으며 하나의 음소를 가진 음성신호가 여러 개로 분할될 수 있고, 여러 개의 음소를 가지지만 하나의 음소로 분할될 수도 있다. (그림 1)의 (d)는 SVF가 모든 하위 지역에 대해서 과분할 되어 있는 것을 보여준다. 과분할 혹은 미분할 문제를 해결하기 위해서 두 개의 근접한 지역을 합치거나 하나의 지역을 두 개 혹은 그 이상으로 나누어주는 메카니즘이 필요하게 된다.

가정 (3)은 본 논문에서 새롭게 사용한 것이며 여전히 연구분석이 필요하고 음절이 발달된 언어에 의존적이다. 하지만, 대부분의 경우 모음은 에너지 레벨 혹은 고주파와 저주파 에너지의 비율의 지역 극대치와 관련이 크다. 일부 무성자음의 경우 에너지 레벨에서 지역 극대치가 있지만 무성

자음은 고주파 에너지만 크고 저주파 에너지는 아주 작으며 모음은 포만트를 형성하지만 무성자음은 고주파 에너지만 높을 뿐이기 때문에 모음과 잡음은 음운학적, 음향학적 정보로부터 찾을 수 있다. 우선 모음의 중간점과 잡음구간을 찾아내고 이를 분할 점으로 사용하여 신호 지역화을 수행한다.

MLS 시스템의 예로서 Glass의 MLS 알고리즘을 분석하여 보면 MLS 알고리즘은 앞서 언급한 선결조건 (A1)과 (A2)가 사용되었다. 하지만, 분리/삽입과정을 이용하지는 않는다. 다음은 Glass의 MLS 알고리즘을 이해하기 쉽게 변경해서 유사코드로 옮겨 놓았다.

- Find boundaries utilizing local clustering :

$$\{b_i : 0 \le i \le N;\ b_i < b_{i+1}\}$$

- Create the corresponding regions :

$$\{R_i : 0 \le i \le N;\ R_i = (b_i, b_{i+1})\}$$

- While(over-segmented) do
  measure

$$\{D_i : 0 \le i < N-1;\ D_i = d(S(R_i), S(R_{i+1}))\}$$

  find $i_c = \arg \min_{0 \le i < N-1} \{D_i : D_i \neq 0\}$

  eliminate b_{i_c};

  merge R_{i_c} to R_{i_c+1};

  set $D_{i_c} = 0$;

  End do

b_0, b_N은 에너지 혹은 영점 교차 등을 통해서 구해지는 음성신호의 시작점과 끝점을 나타낸다. MLS 알고리즘의 성능은 유사성측정 $d(\cdot,\cdot)$에 의해 좌우된다. 하지만, 이러한 유사성 측정 기법이 음소 경계에서 반드시 큰 값을 가진다고 할 수 없기 때문에 과 결합이 발생하게 된다. 따라서, 실제 응용에서는 반자동으로 사용되었다.

3. 유사성측정기법 : SVF

$\{S_n\}_{n=0,\cdots,N}$을 스펙트라 혹은 켑스트라와 같은 파라메터화한 음성벡터라고 하면 p와 q는 미리 결정된 영이 아닌 정수들이다. 먼저, 지역평균을 빼서 지역 나머지 $R_n^{(p)}$를 구한다.

$$R_n^{(p)} = S_n - S_n^{(p)},\quad S_n^{(p)} = \frac{1}{2p+1} \sum_{k=-p}^{p} S_{n+k} \qquad (1)$$

그러면, SVF $\{F_n^{(p,q)}\}$는 지역 나머지의 프레임들간의 거리를 측정함으로써 구할 수 있다.

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

지역 나머지 $\{R_n^{(p)}\}$은 음성 벡터의 지역 분산을 강조하는데 관계가 있기 때문에 p값이 너무 큰 값을 갖지 않도록 하고 반면에, $\{S_n\}$의 지역 평균인 $\{S_n^{(p)}\}$을 위해서 p는 어느 정도는 크기를 유지해야 한다. 정수 q는 SVF의 피크를 구하기 위해서 충분히 작아야 한다. 많은 수치적 실험을 통해서 이 정수들이 다음을 만족한다라는 것을 발견했다.

$$q \le p \le 3 \sim 4 \qquad (3)$$

본 논문에서는 p = 3, q = 2를 사용하였다. 이 값을 이용해서 SVF를 구해보면 근접한 지역 나머지가 유사하면 거의 영의 값을 가지고, 서로 다르면 큰 값을 가진다는 것을 알 수 있다.

4. 한국어의 음소정의

본 연구에서는 강인한 자동 음소 레이블러 뿐만 아니라 음운학적, 음향학적 정보를 담고 있는 한국어 음소 정의를 제시한다. 27개의 자음, 17개의 모음, 2개의 특별음소(il, ung), 그리고 비음성구간(S)을 합하여 총 47개의 음소들을 정의하였다. 특별음소 일(il)과 웅(ung)은 사람에 의해서도 그 경계를 구분하기가 어렵기 때문에 하나의 음소로 정의한 것이다. 각각의 음소는 자음과 모음, 유성음과 무성음, 파열음, 마찰음, 비음 등의 정보를 담고 있다. 그리고, 귀착, 동화, 축약, 생략, 첨가, 이화 현상의 한국어의 음운규칙[1,2]을 고려하여 연속 음성의 음소표기를 시도하였다. 예를 들어, 귀착현상으로서 "앉다"를 /a-n dd-a/로 표기하였고, 구개음화로서 "굳이"를 /g-u j1-i/로 표기하여 본 연구에 적용하였다. <표 1>은 구체적인 음소의 음운학적 지식을 표현하기 위해서 사용한 컬럼 정보를 보여 준다. 그리고, <표 2>는 자음에 대한 음소정의를 보여 주고 있으며, <표 3>은 모음과 일부 다르게 정의한 음소를 보여준다.

〈표 1〉 한국어 음소(칼럼정보)

KrPhone : The Korean Phoneme
Column information 0. phonemes 1. vowel? (0 : consonant ; 1 : semi-vowel ; 2 : vowel) 2. voice? (0 : unvoiced ; 1 : semi-voiced ; 2 : voiced) 3. stop? (0 : non-final ; 1 : final-consonant ; 2 : stop) 4.plosive? (0 : non-plosive ; 1 : semi-plosive ; 2 : plosive) 5. nasal? (0 : non-nasal ; 1 : nasal) 6. landmark? (0 : none ; 1 : low ; 2 : mid(none) ; 3 : high) 7. energy position (0 : unknown ; 1 : low ; 2 : middle ; 3 : high ; 31 : high&low, etc)

〈표 2〉 한국어 음소(자음)

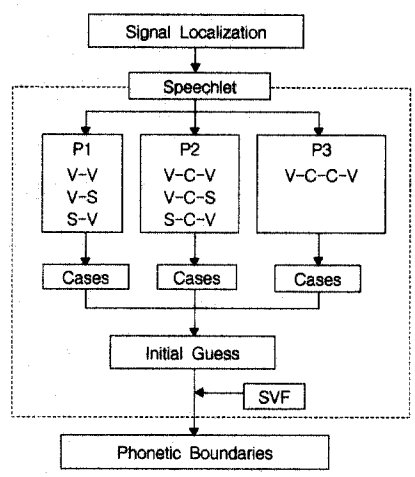
자 음			
음 소	심 볼	칼 럼 정 보	비 고
ㄱ(초성)	g	0 0 0 2 0 0 2	가을
ㄱ(중성)	g1	0 1 0 2 0 0 2	동굴
ㄱ(종성)	g2	0 1 2 0 0 0 2	육
ㄲ	gg	0 0 0 2 0 0 2	까닭
ㄴ	n	0 2 1 0 1 1 1	
ㄷ(초성)	d	0 0 0 2 0 0 2	다리
ㄷ(중성)	d1	0 1 0 2 0 0 2	바다
ㄷ(종성)	d2	0 1 2 0 0 0 2	낟,낫,낯
ㄸ	dd	0 0 0 2 0 0 2	딸
ㄹ(초성)	r	0 1 0 0 0 0 1	다른
ㄹ(종성)	l	1 2 1 0 0 0 12	보통실
ㅁ	m	0 2 1 0 1 1 12	
ㅂ(초성)	b	0 1 0 2 0 0 23	바람
ㅂ(중성)	b1	0 2 0 2 0 0 23	나비
ㅂ(종성)	b2	0 1 2 0 0 0 23	합
ㅃ	bb	0 0 0 2 0 0 23	빨래
ㅅ	s	0 0 0 1 0 0 23	
ㅆ	ss	0 0 0 1 0 0 23	
ㅇ(종성)	ng	0 2 1 0 1 0 12	
ㅈ(초성)	j	0 0 0 2 0 0 0	자리
ㅈ(중성)	j1	0 1 0 2 0 0 0	이제
ㅉ	jj	0 1 0 2 0 0 0	짜임새
ㅊ	ch	0 0 0 2 0 3 23	처음
ㅋ	k	0 0 0 2 0 0 0	칼
ㅌ	t	0 0 0 2 0 0 0	탈
ㅍ	p	0 0 0 2 0 0 3	파리
ㅎ	h	0 0 0 0 0 0 0	

〈표 3〉 한국어 음소(모음과 예외)

모음과 예외			
음 소	심 볼	칼 럼 정 보	비 고
ㅏ	a	2 2 0 0 0 0 12	아
ㅑ	ya	2 2 0 0 0 0 12	야
ㅓ	eo	2 2 0 0 0 0 12	어
ㅕ	yeo	2 2 0 0 0 0 12	여
ㅗ	o	2 2 0 0 0 0 1	오
ㅛ	yo	2 2 0 0 0 0 1	요
ㅜ	u	2 2 0 0 0 0 1	우
ㅠ	yu	2 2 0 0 0 0 12	유
ㅡ	eu	2 2 0 0 0 0 0	으
ㅣ	i	2 2 0 0 0 0 13	이
ㅔ,ㅐ	e	2 2 0 0 0 0 0	에,애
ㅖ,ㅒ	ye	2 2 0 0 0 0 0	얘,예
ㅘ	wa	2 2 0 0 0 0 12	와
ㅚ,ㅙ,ㅞ	we	2 2 0 0 0 0 0	외,왜,웨
ㅝ	weo	2 2 0 0 0 0 12	워
ㅟ	wi	2 2 0 0 0 0 0	위
ㅢ	eui	2 2 0 0 0 0 0	의
일	il	2 2 0 0 0 0 13	일
웅	ung	2 2 0 0 0 0 1	웅
묵음	sil	0 0 0 0 0 0 0	묵음

5. DAC 기반 자동 분할 알고리즘

본 논문에서 제시하는 전체적인 DAC기반 자동분할 알고리즘이 (그림 2)에 제시되고 있다. 분할개념 으로 연속음성 신호들이 정의된 *speechlet*으로 분할하고 신호를 지역화 하는 개념으로 연계되고 정복 알고리즘으로 세분화된 특성화 접근방법을 이용하고 있다.



(그림 2) 한국어의 ASL 과정

지역화에 이용된 분할개념으로 한국어의 특성을 분석하여 음절을 V, C-V, V-C, C-V-C(C = 자음, V = 모음)로 구성하였으며 음절의 앞뒤에 묵음(S)이 올 수 있다.

음성신호가 모음과 자음에 의해서 지역화 되면, 각각의 지역화 된 신호의 음소들은 다음 패턴중의 하나로 분류가 된다.

$$\begin{aligned} p_1 &: \text{V-V, V-S, S-V} \\ p_2 &: \text{V-C-V, V-C-S, S-C-V} \\ p_3 &: \text{V-C-C-V} \end{aligned} \tag{4}$$

예를 들어, 다음창(/S d a eu m ch a ng S/)을 발음하였을 때 /S d a/는 S-C-V가 되고 /a eu/은 V-V, /eu m ch a/는 V-C-C-V, /a ng S/은 V-C-S가 된다. 본 논문에서는 이 지역화 된 신호 조각을 *speechlet*이라고 정의하였다. 패턴이 p_k인 한 *speechlet*에서 k개의 음소 경계를 찾으면 된다.

일단 잡음구간과 모음을 정확하게 구분하고 나면 각 *speechlet*에 대한 구체적인 분할 전략을 적용시킨다. 예를 들어, 패턴 V-C-S에 대해서 C와 S의 경계는 에너지 레벨을 사용해서 쉽게 구별할 수 있다. C는 비음, 폐음, 반모음 등의 특성을 가지고 있고, 모음에 대해서도 포만트 구조가 다를 수 있기 때문에 많은 선택 가능한 사례가 존재한다. 하지만, 세분화되고 특화 된 사례연구를 통해서 V와 C사이의 음소 경계를 포함하는 구간에 대한 초기 추측을 할 수 있고 실제 음소 경계는 구간사이의 SVF 중에서 가장 큰

피크를 선택 할 수 있다. 신호 지역화에 의해 연속된 신호를 분할하고 각 분할 된 *speechlets*에서 특성화 된 음운학적, 음향학적 지식을 이용 음소 경계를 찾는 것이 바로 기존 연구와 다른 차별성이라 할 수 있다.

6. 신호 지역화

신호 지역화는 전체 음성구간을 잡음구간과 모음의 중간점을 이용하여 *speechlet*으로 나누고 나누어진 지역적 신호들에 대해서 연속적으로 분할을 수행하는 프로세스이다. 잡음구간을 구별해 내는 것은 쉽기 때문에 모음을 찾는 전략에 대해서 자세하게 설명하고자 한다.

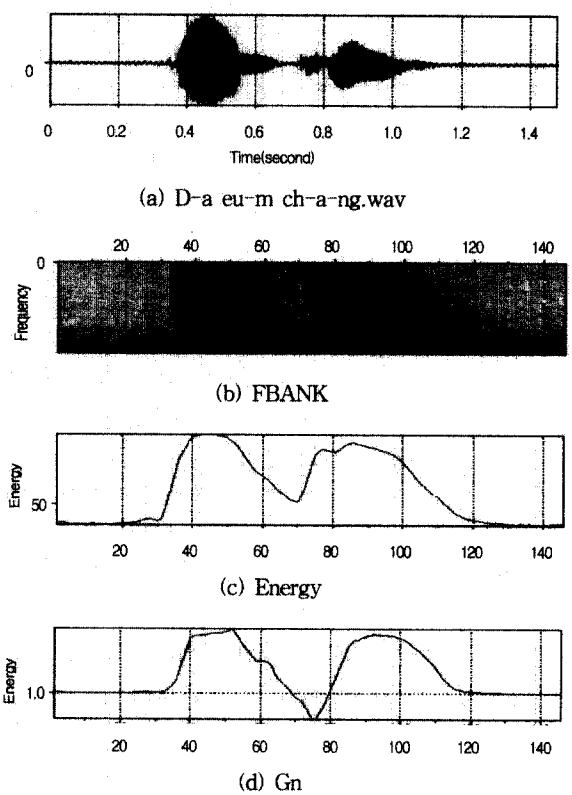
대부분의 경우 모음은 에너지 레벨에서 지역적 극대치를 가진다. 하지만 모든 지역 극대치가 모음이라고 할 수는 없다. /s/, /ss/, /ch/, /k/, /t/, 그리고, /p/와 같은 무성 자음들은 높은 에너지를 가질 수 있기 때문이다. 원칙적으로 자음에 대해서는 포만트가 나타나지 않는 다는 사실을 이용해서 구별하거나 자음에 대해서 영점교차율이 높다는 것을 이용해서 구별할 수 있다. 모음은 그 에너지가 포만트의 형태로 전 범위의 주파수대에서 에너지가 나타나지만 무성자음은 고주파 밴드에서만 에너지가 나타난다. 따라서, 고주파 에너지가 상당히 높은 지역 극대치를 무시할 수 있다. 하지만, 모음과 자음을 구별하는 문제는 단순하지 않다. 반모음(semi-vowel : /l/, /r/)과 비음(nasals : /m/, /n/, /ng/)의 에너지 수준 과 모음의 에너지 수준을 구별하는 전략을 수립해야 한다.

모음의 에너지 수준에서 지역 극대치가 나타나지 않을 수가 있다. 예를 들어, 다음창(/d a eu m ch a ng/)을 발음할 경우 /eu/ 음에서 지역 극대치가 나타나지 않을 수 있다. 특히, 모음이 연속으로 오고 아주 빠르게 발음이 되는 경우에 발생빈도가 높다. 그리고 이 모음들(특히, a o, u eo)은 제 1 포만트와 제 2 포만트에서의 주파수가 유사하다는 것을 알 수 있다. 이러한 문제는 저주파와 고주파의 에너지 비율을 통해서 해결할 수 있다. 이 에너지 비율은 다음과 같이 정의된다.

$$G_n = \sum_{i=0}^{m/2-1} S_n(i)^2 / \sum_{i=m/2}^{m-1} S_n(i)^2 \qquad (5)$$

이때 m은 음성 벡터, S_n은 영역이다.

(그림 3)에서 남자에 의해서 발음된 다음창(/d a eu m ch a ng/에 대한 웨이브 형태를 그렸다. FBANK, 에너지 수준, 그리고 에너지 비율 $G=\{G_n\}$를 보여주고 있다. 지역 극대치가 43, 76, 그리고 85 프레임에서 나타난다. 반면에 저주파와 고주파 에너지 비율에 의해서 51과 92 프레임이 선택이 된다. 76 프레임은 무성 자음과 관련이 있다. 즉, 76 프레임에서 고주파 밴드를 가진다. 따라서, 지역 극대치는 43, 51, 85, 그리고 92가 되고 실제 모음의 개수는 3이기 때문에 하나를 제거해야한다. 음성이 29 프레임에서 자음으로 시작을 하기 때문에 첫 번째 모음은 35~50 프레임 사이에 있을 것으로 기대할 수 있다. 게다가, 에너지 수준 $E=\{E_n\}$과 에너지 비율 G가 43 프레임에서 크기 때문에 35~50 프레임 사이에 첫 번째 모음이 있다는 것을 더 확신할 수가 있다. 만약 85 프레임이 두 번째 모음이라고 하면 3번째 모음은 92 프레임이 되어야 하고 두 개의 자음이 85와 92 프레임사이에 있어야 하는데 이는 불가능하다. 따라서, 51 프레임이 두 번째 모음으로 확정된다. 따라서 43, 51, 그리고 85 프레임이 모음으로 결정된다.



(그림 3) 남자에 의해서 발음된 "다음창"의 (a) 웨이브 (b) FBANK, (c) 에너지 수준, (d) 에너지비율(G)

본 연구에서 사용되는 모음인식과정을 일반화하고 요약하면 다음과 같다.

- 음성 웨이브 신호로부터 STM(short time magnitude)을 구하고, 스펙트럼으로부터 에너지 수준 E와 저주파와 고주파의 에너지 비율(G)에서의 모든 지역 극대치로부터 피크를 구한다.
- 무성 자음와 관련이 있는 모든 피크들을 제거한다 : 모음과 함께 올 때 다음 피크에서 모음이 올 가능성을 검토하라.
- 의사결정 기준을 만족하는 선택된 피크들에 대해서 확정 태그를 붙인다 ; 처음과 마지막 모음에 대한 피크를

결정하고 ($2\gamma+1$) 프레임(양쪽에 γ 프레임 씩)를 평균한 후에도 여전히 지역 극대치인 피크들을 선택한다. $\gamma=3\sim5$가 적당하다. 예를 들어, $\gamma=4$에서 43과 85 프레임이 (그림 2)의 평균화를 통해서도 여전히 피크들로 남는다면 모음으로 확정된다. 이는 Max search와 Max adjust 모듈로 구성되어 있다.

- "음소길이" 체크를 한다.
- 결과가 합당한지를 조사한다. 만약 결과 합당하지 않다면 /fail/ 태그를 붙인다.

7. 사례 연구

본 논문에서 사례연구란 각 *speechlet*의 시작점과 끝점이 결정된 후 음운학적, 음향학적 지식을 이용해서 *speechlet* 내의 각각의 세분화 된 초기 경계들을 결정하는 것을 말한다. 총 패턴의 수는 3가지이며 패턴에 따라 결정하여야할 각 *speechlet*의 초기 경계들의 수도 자동 결정된다. 각 사례별로 초기경계들의 결정방법에 대해서 자세히 설명하고자 한다.

7.1 P1(경계 수가 한 개인 경우)

7.1.1 VV

모음검출에서 에너지와 고주파와 저주파의 에너지 비율을 이용해서 지역적 최대치를 찾았고 이를 *speechlet*의 경계로 사용하였다. 따라서, 상대적으로 모음과 모음이 연속으로 오는 경우 그 사이에서 지역적 최소가 존재한다. 그러므로, 지역적 최소에서 바로 모음과 모음의 경계가 존재한다라고 할 수 있다. 만약, 지역적 최소가 여러 개 존재하는 경우는 가장 최소 에너지를 가지는 지역적 최소를 초기 경계로 한다.

7.1.2 SV

음성신호의 시작점과 끝점 추출에 의해서 구해진 음성의 시작점을 그 초기 경계로 한다. 경계를 구하기 쉽고 정확하다.

7.1.3 VS

음성신호의 시작점과 끝점 추출에 의해서 구해진 음성의 끝점을 그 초기 경계로 한다. 경계를 구하기 쉽고 정확하다.

7.2 P2(경계 수가 두 개인 경우)

7.2.1 VCV

/ng/을 제외한 종성자음은 다음에 오는 모음에 의해서 영향을 받기 때문에[2] VCV 경우에서 /ng/을 제외한 모든 자음은 초성으로 발음이 된다. 따라서, 대부분의 경우 V-C 사이에서 최소 에너지를 가지게 된다.

자음이 비음인 /n/, /m/이면 V-C 사이에서 최소 에너지를 가지는 점을 첫 번째 초기 경계로 한다. 두 번째 초기 경계는 C-V사이의 지역적 최소 에너지를 가지는 지점으로 한다. 왜냐하면, /n/, /m/은 비음인 동시에 유성자음이기 때문에 에너지가 존재하며 C-V(그리고 V-C)의 경계에서 지역적 최소 에너지를 가지기 때문이다.

비음 /ng/인 자음이 오는 경우는 두 번째 초기경계지점에서 최소 에너지를 가지게 된다. 비음 /ng/가 종성이기 때문이다. 첫 번째 경계는 V-C 사이의 지역적 최소 에너지를 가지는 점을 이용한다. 지역적 최소가 존재하지 않는다면 큰 peak을 찾거나 두 번째 경계에서 threshold값을 적용하여 결정한다.

자음이 파열음인 경우 고주파 에너지 레벨이 높다는 음운학적 지식을 이용한다. 또한 파열음을 발음하기 위해서 *short pause* 현상이 일어난다. 따라서, 최소 에너지를 가지는 점을 첫 번째 경계로 하고 두 번째 경계는 고주파 에너지를 가지는 점에서부터 최소 에너지를 가지는 점을 그 경계로 한다.

자음이 /r/, /h/인 경우는 지역적 최소 에너지를 갖는 프레임을 첫 번째 초기 경계로 하고 이점을 기준으로 2~3 프레임의 threshold값을 적용해서 두 번째 초기 경계를 결정한다.

7.2.2 VCS

자음은 폐쇄음이거나 비음인 종성으로 쓰이는 자음밖에 올 수가 없다.

무성음이고 폐쇄음인 자음의 경우는 대부분 8프레임을 초과하지 않는다. 따라서, 두 번째 경계 점에서 8프레임정도 뒤쪽의 피크를 첫 번째 경계로 한다. 두 번째 경계 점은 음성의 끝점을 그 초기 경계로 한다.

비음인 자음(/n/, /l/, /m/)이 오는 경우는 이미 결정된 두 번째 경계로부터 지역적 최소 에너지를 가지는 점을 찾는다. 만약, 최소 에너지가 존재하지 않는 경우 두 번째 경계로부터 6~8 프레임의 threshold값을 적용해서 첫 번째 초기 경계를 구한다.

7.2.3 SCV

종성으로 쓰이는 자음은 절대로 올 수 없다.

묵음 다음에 비음인 자음이 오는 경우 에너지비율(G)가 가장 작아지는 지점을 두 번째 경계로 한다. 첫 번째 경계는 음성의 시작점을 그 경계로 한다.

묵음 다음의 자음이 무성음이고 파열음인 경우 고주파수 레벨의 에너지가 높다. 모음 또한 파열 자음만큼은 아니지만 상당히 높은 에너지를 보이므로 자음에서 모음으로 변화하는 지점에서 상대적으로 에너지가 낮아진다. 따라서, 고주파 레벨 에너지를 구해서 고주파 에너지가 가장 작아지는 점을 찾아 두 번째 경계로 한다.

7.3 P3(경계 수가 세 개인 경우)

7.3.1 VCCV

첫 번째 자음은 종성 자음이고 두 번째 자음에는 종성으로 쓰이는 자음은 올 수가 없다. 이때 종성자음은 비음인 유성음 혹은 폐쇄음인 무성자음일 수 있다. 그리고, 종성으로 쓰이지 않는 자음은 비음인 유성음 혹은 파열음이거나 /r/, /h/인 무성자음일 수 있다. 따라서, 다음 4가지의 경우가 존재한다.

1) V-VC-UC-V

두 번째 자음인 UC(Unvoiced Consonant)를 발음하기 위해서 short pause 현상이 일어남으로 자음과 자음 사이에서 최소 에너지를 갖는다. 따라서, 최소 에너지를 가지는 지점을 두 번째 경계로 정하고 첫 번째 경계는 V-VC(Voiced Consonant) 사이에서 지역적 최소 에너지를 가지는 지점을 찾아 초기 경계로 이용한다.

두 번째 자음이 무성음이고 파열음인 경우 고주파수 레벨의 에너지가 높다. 그리고, 모음 또한 파열 자음만큼은 아니지만 상당히 높은 고주파 에너지를 보이므로 자음에서 모음으로 변화하는 지점에서 상대적으로 고주파 에너지가 낮아진다. 따라서, 고주파 레벨 에너지를 구해서 에너지가 가장 작아지는 점을 찾아 세 번째 경계로 한다.

두 번째 자음이 /r/, /h/인 경우는 두 번째 경계에서 2~3 프레임정도의 threshold값을 적용해서 세 번째 경계를 결정한다.

2) V-UC-VC-V

UC에는 폐쇄음인 종성자음이 온다. 그리고 VC로 올 수 있는 자음은 /n/, /m/ 뿐이다. 따라서, 폐쇄음을 발음한 이후 극히 짧은 묵음 구간이 발생하며 최소 에너지를 가지는 지점이 바로 UC-VC의 경계가 된다. 첫 번째 경계는 무성음이고 폐쇄음인 자음의 경우는 대부분 8 프레임을 초과하지 않는다는 사실을 이용하여 두 번째 경계 점에서 8프레임정도 뒤쪽의 피크를 첫 번째 경계로 한다. 세 번째 경계는 VC-V 사이에서 지역적 최소 에너지를 가지는 지점을 찾아 초기 경계로 이용한다.

3) V-UC-UC-V

첫 번째 UC에는 폐쇄음인 종성자음이 온다. 그리고, 두 번째 UC는 파열음, 혹은 /r/, /h/음이다. 폐쇄음을 발음한 이후 극히 짧은 묵음 구간이 발생하므로 이 지점에서 최소 에너지가 존재하게 되고 두 번째 경계로 사용된다. 첫 번째 경계는 무성음이고 폐쇄음인 자음의 경우는 대부분 8프레임을 초과하지 않는다는 것을 이용한다. 따라서, 두 번째 경계 점에서 8프레임정도 뒤쪽의 피크를 첫 번째 경계로 한다.

두 번째 자음이 무성음이고 파열음인 경우 고주파 수 레벨의 에너지가 높다. 그리고, 모음 또한 파열 자음만큼은 아니지만 높은 고주파 에너지를 보이므로 자음에서 모음으로 변화하는 지점에서 상대적으로 낮은 고주파 에너지가 존재한다. 따라서, 고주파 레벨 에너지를 구해서 에너지가 가장 작아지는 점을 찾아 세 번째 경계로 한다.

두 번째 자음이 /r/, /h/인 경우는 두 번째 경계에서 2~3 프레임정도의 threshold값을 적용해서 세 번째 경계를 결정한다.

4) V-VC-VC-V

두 자음 모두 유성 자음으로 이루어진 경우로 두 번째 VC에는 /n/, /m/만 올 수 있다. VC-VC 사이에서 지역적 최소 에너지가 존재한다면 그 점을 두 번째 경계로 한다.V-VC, VC-V의 경계는 지역적 최소 에너지를 가지는 지점을 찾아 그 첫 번째와 세 번째 경계로 이용한다.

8. 음소분할 실험 결과

〈표 4〉 분할 알고리즘의 성능

구 분	In 20ms		In 30ms		In 40ms		Total boundaries
S-V	160	97.56%	162	98.78%	162	98.78%	164
S-C-V	1412	91.81%	1480	96.23%	1507	97.98%	1538
V-C-C-V	427	75.31%	462	81.48%	492	86.77%	567
V-C-V	982	83.93%	1067	91.20%	1109	94.79%	1170
V-S	483	95.08%	487	95.87%	494	97.24%	508
V-C-S	707	83.97%	758	90.02%	786	93.35%	842
V-V	172	68.25%	196	77.78%	216	85.71%	252
Total	4343	86.15%	4612	91.49%	4766	94.54%	5041

<표 4>는 W대에서 개발한 PRW 3813종 40set(남녀 총 500명분) 중에서 남녀 각각 5명씩, 총 900여 단어를 대상으로 실험을 수행한 결과를 나타내고 있다. 새로운 알고리즘의 성능을 평가하기 위해서 음운학적 지식이 있는 사람에 의한 레이블링을 수행하였으며, 20ms 내에서 86.15%, 30ms에서 91.49%, 40ms에서 94.54%의 정확성을 보여주고 있다.

본 제안 알고리즘의 성능을 비교하기 위하여 HTK(Hidden Markov Model Tool Kit)[11]을 사용하여 PRW 3813종 40 set 모두를 음소단위로 훈련시키고 HMM모델을 만들었으며 이 HMM을 이용하여 임의로 선택한 3명에 대해 레이블링을 시도하고 사람에 의한 레이블링과 비교하였다. HMM을 이용한 레이블링 방식의 성능은 <표 5>과 같이 20ms에서 62.08%의 정확성을 보였다. 본 자동 레이블링 알고리즘은 HMM을 이용한 레이블링 방식보다 우수하다고 할 수 있으며 특히, V-S와 S-V의 경우는 20ms의 내에서 95.08%, 97.5%의 높은 성능을 보인다. 반면 V-V과 V-C-C-V의 경우는 각각 68.25%, 75.31%를 보여 각 경우별로 차이를 보여주고 있다.

〈표 5〉 HMM을 이용한 분할 성능

구 분	In 20ms		In 30ms		In 40ms		Total boundaries
S-V	63	71.59%	72	81.82%	77	87.5%	88
S-C-V	256	62.44%	307	74.78%	343	83.66%	410
V-C-C-V	109	59.89%	144	79.12%	163	89.56%	182
V-C-V	248	54.15%	305	66.59%	364	79.48%	458
V-S	49	66.22%	60	81.08%	65	87.84%	74
V-C-S	73	55.73%	104	79.39%	111	84.73%	131
V-V	60	64.52%	75	80.65%	83	89.25%	93
Total	858	62.08%	1067	77.65%	1206	86%	1436

이는 본 연구의 핵심인 각 케이스별로 특화 된 음운학적, 음향학적 지식을 이용 할 경우 성능을 향상시킬 수 있는 높은 가능성을 보여주고 있다. 제안 방법은 현재까지 개발된 음운학적, 음향학적인 지식만을 사용하였다는 점에서 차별화 될 수 있는 방법이다. 즉 특화 되고 세분화 된 음성학적인 지식을 계속 연구하여 추가시 전체적인 성능 향상이 가능하리라 판단된다. 이러한 연구는 각 언어가 가지고 있는 다양한 자음 모음의 조합 특성을 더욱 파악 분석하고 특화 된 지식으로 추출하면 더욱 향상시킬 수 있다.

9. 결 론

음성신호에 대해서 음운학적, 음향학적 정보를 이용하는 DAC기반의 자동 레이블링 알고리즘을 소개하였다. 이 자동 레이블링 알고리즘은 현재 20ms이내에서 86%정도의 정확성을 보이고 있다. 전체성능은 특화 된 음성 지식을 더욱 발견하여 추가한다면 계속 향상 될 수 있을 것이다.

음성학적인 특화지식 컴포넌트의 차후 연구로서 모음과 모음이 연속해서 오는 경우 특히, 이중모음이 오는 경우에 대한 연구가 더 진행되어야 하며 자음에 대해서도 더욱더 다양한 전략이 필요하다. 특히, 모음 다음에 유성자음(/l/, /m/, /n/, /ng/)이 오는 경우 그 경계를 구분하기가 어렵고, VCCV 경우에서 폐쇄음인 자음 다음에 /s/, /ss/ 자음이 오는 경우에는 폐쇄음 다음에 short pause가 아니을 수 있다. 이러한 문제점들에 대한 성능 개선을 위해서 많은 음성DB에 대해 자동레이블링 알고리즘의 성능을 평가하고 새로운 전략을 수립하여야 한다.

본 DAC기반 자동분할 알고리즘은 HMM 방식과는 달리 훈련과정이 필요 없으며 대용량 음성DB에 대한 레이블링 작업을 빠른 시간으로 수행할 수 있고 잘 정의 된 음성학적인 지식을 적용하여 보다 일관성이 있는 방법으로 발전 가능성이 있다고 사료되어진다. 보다 다양하고 특화 된 음운학적, 음향학적인 지식을 계속 연구하여 얻어진 음성학적인 지식을 계속 추가하여 나간다면 현재까지 방법보다도 더욱더 성능 면에서 향상시킬 수 있는 방법으로 발전 확대가 가능하다는 점에서 본 연구의 의의를 찾을 수 있다.

참 고 문 헌

[1] 엄태수, "한국어의 음운규칙 연구", pp.172-264, 1999.

[2] 허웅, "국어음운학", 정음사, pp.129-280, 1984.

[3] E. Vidal and A. Marzal, "A review and new approaches for automatic segmentation of speech signals, Signal Processing V : Theories and Applications (L. Torres, E. Masgrau, and M. A. Lagunas, eds.)," Elsevier Science Publisher, Amsterdam, pp.43-53, 1990.

[4] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labelling of speech based on hidden Markov models," Speech Communication pp.357-370, 1993.

[5] J. P. Hosom, "Automatic time alignment of phonemes using acoustic-phonetic information," Ph.d. thesis, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, May, 2000.

[6] J. R. Glass, "Finding acoustic regularities in speech : Application to phonetic recognition," Ph.D. Thesis, MIT Press, May, 1988.

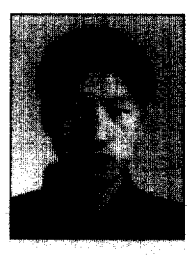
[7] K. Kvale, "On the connection between mannual segmentation conventions and error made by automatic segmentation," Proceedings of ICSLP'94 (Yokohama, Japan), pp. 1667-1670, September, 1994.

[8] P. Cosi, "SLAM : A PC-based multi-level segmentation tool, Speech Recognition and Coding : New Advances and Trends (A. J. R. Ayuso and J. M. L. Soler, eds.)," Compuer and System Sciences, Vol.147, Springer-Verlag, New York, pp.124-127, 1995.

[9] R. Andr e-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," IEEE Trans. ASSP 36, pp.29-40, 1988.

[10] S. Cox, R. Brady, and P. Jackson, "Techniques for accurate automatic annotation of speech wave forms," Proceedings of ICSLP'98 (Sydney, Australia), pp.1947-1950, December, 1998.

[11] "http : //htk.eng.cam.ac.uk," Hidden Markov Model Tool Kit Homepage.



구 찬 모

e-mail : cmkoo@madang.ajou.ac.kr

2000년 계명대학교 산업공학과(공학사)

2002년 아주대학교 산업공학과(공학석사)

현재 아주대학교 산업공학과 박사과정

관심분야 : ERP, SCM, CRM, E-Business, Signal Processing

왕 지 남

e-mail : gnwang@madang.ajou.ac.kr

1983년 아주대학교 공과대학 산업공학과(공학사)

1985년 한국과학기술원 산업공학과(공학석사)

1992년 미 Texas A&M 대학 산업공학과(공학박사)

현재 아주대학교 산업공학과 부교수

관심분야 : Neural Network, 시스템 진단, 감시 및 제어, 제조시스템의 데이터 통신, 지능형 분산 정보 시스템 설계, CIM, CALS, 초고속망 응용기술, Computer Vision