Ramp Edge Detection을 이용한 끝점 검출과 음절 분할에 관한 연구

유 일 수, 홍 광 석
성균관대학교 정보통신공학부
전화: (031) 290-7196

A Study on Endpoint Detection and Syllable Segmentation System Using Ramp Edge Detection

Il-Soo Yu, Kwang-Seok Hong
School of Information and Communication Eng., Sungkyunkwan University
E-mail: gildda@nate.com

Abstract

Accurate speech region detection and automatic syllable segmentation is important part of speech recognition system. In automatic speech recognition system, they are needed for the purpose of accurate recognition and less computational complexity, In this paper, we propose improved syllable segmentation method using ramp edge detection method and residual signal peak energy. These methods were used to ensure accuracy and robustness for endpoint detection and syllable segmentation system. They have almost invariant response to various background noise levels. As experimental results, we obtained the rate of 90.7% accuracy in syllable segmentation in a condition of accurate endpoint detection environments.

I. 서론

음성인식시스템에서 인식 성능과 계산량 감소에 큰 영향을 주는 요소는 바로 정확한 음성 영역의 검출(끝점 검출)과 음절 영역 분할에 있다. 잘못된 끝점 검출은 입력 음성의 오인식의 근본적인 원인이며, 정확한 음성영역 검출은 묵음 영역을 적당히 제거 함으로서 묵음영역에서 오는 에러를 최소화 할 수 있다. 또한 음절 분할은 단어-단위 인식과 연속-음성 인식에서 인식할 후보 모델을 제한 함으로서, 좀 더 정확한 인식과 계산량을 크게 감소 시킬 수 있는 장점이 가지므로 그 필요성이 증가되고 있다.[4] 끝점 검출 방법으로는 Energy Threshold, Pitch Detection, Spectrum Analysis, Cepstral Analysis, ZCR등, 많은 방법들이 소개되어 왔다.[1] 하지만 대부분의 끝점 검출 방법은 낮은 SNR(Signal-to-Noise Rate)에 약하거나 많은 계산량을 요구를 함으로서 실시간 음성 구간 검출에 에로 사항을 갖는 경우가 많다. 본 논문에서는 [Qi Li외]가 소개한, 에너지 정보를 이용한 Ramp Edge Detection 방법과 3상태 끝점 결정 알고리즘을 적용하여 음성영역을 검출 하였다. 그리고 동시에 미리 Ramp Edge로 설계된 최적 필터를 변경 함으로써, 입력된 전체 음성 영역(문장 또는 단어)의 검출 뿐만 아니라 소-음성 구간에 해당하는 음절 영역을 검출 할 수 있다는 것을 실험을 통해 확인하였다.

따라서, 본 논문은 Ramp Edge Detection 방법으로 끝점 검출과 음절 분할을 동시에 실시간으로 수행할 수 있는 음절 분할 끝점 검출 알고리즘을 제안하고자 한다. 그러나 음절 분할은 음성 신호의 에너지 정보만으로는 정확히 수행하기 어려운 문제점을 갖는다. 특히 연음과 빠른 발성 음성 신호의 경우 변화에 둔감한 에너지 정보만으로 음절을 명확히 구분 짓기 어렵다. 이 문제를 보안하기 위해 LCR(Level Crossing Rate), PVR(Peak Valley Rate), Spectral Analysis, 정규화 자기상관 계수, 또는 Pitch 정보 등이 이용되고 있다[4,5] 하지만, 본 논문은 좀 더 효율적인 분할을 고려해 주기 위해 음성 신호에 대한 역 LPC 필터를 거친 잔여 신호의 피치 에너지 정보를 이용하여, 다시 Ramp Edge Detection 방법을 적용하였다.

2장에서는 끝점 검출 및 분할 검출 파라미터들을 소개 하며, 3장에서는 2장에 소개된 방법들을 사용하여, 본 논문의 음절 분할 끝점 검출 알고리즘을 제안한다. 4장에서는 본 제안 내용을 검증하기 위한 실험에 대한 결과를 제시하고 5장에서는 결론을 맺는다.

II. 끝점 검출 및 분할 파라미터

본 논문에서는 음성 영역 검출 및 분할을 위해 사용되는 특징 파라미터로 에너지와 Ramp Edge의 최적 필터, 그리고 역 LPC 필터를 거친 잔여 신호의 피크 에너지를 사용하였다.

2.1 에너지

에너지 정보는 묵음에서 음성 구간을 구별 지어주는 좋은 특징 정보로 이용될 수 있어, 끝점 검출에 많이 사용되고 있다. 프레임당 에너지는 아래의 식(1)과 같다.

$$x_s(t) = 10\log_{10} \sum_{i=n_t}^{n_t+N-1} s(i)^2 \qquad (1)$$

$x_s(t)$ 는 음성신호의 프레임당 에너지이며, $s(i)$ 는 입력 신호의 데이터, n_t 는 현재 데이터의 샘플의 위치, N 는 한 프레임 길이, t 는 프레임 수를 나타낸다.

2.2 Ramp Edge의 최적 필터

묵음에서 음성 신호가 시작 되는 부분과 음성 신호가 끝나고 묵음이 시작하는 부분의 에너지의 윤곽 곡선을 Ramp Edge로 가정한다.[1] 일반적으로 Ramp Edge는 아래의 식(2)와 같이 함수로 표현된다.[2]

$$c(x) = \begin{cases} 1 - \dfrac{e^{-sx}}{2}, & \text{for } x \geq 0 \\ \dfrac{e^{sx}}{2}, & \text{for } x \leq 0 \end{cases} \qquad (2)$$

x 는 프레임 수이고 다른 종류의 Edge들에 맞춰줄 수 있는 s는 임의의 양수의 상수 값으로 표현한다. 이 식(2)를 바탕으로 Petrou와 Kittler[3]에 의해 영상처리의 Edge Detection에 적합한 최적 필터가 설계되었다. 이 최적 필터는 끝점 검출에 이용하기에 적합하여 [Qu Li 외]의해 사용되었으며, 식(3)과 같다.

$$\begin{aligned} f(x) = {} & e^{Ax}[K_1 \sin(Ax) + K_2 \cos(Ax)] \\ & + e^{Ax}[K_3 \sin(Ax) + K_4 \cos(Ax)] \\ & + K_5 + K_6 e^{sx} \end{aligned} \qquad (3)$$

A와 K_i는 필터 파라미터들이며, $f(x)$는 전체 필터의 반에 해당한다. 전체 필터의 계수는 식(4)와 같다.

$$\begin{aligned} h[t] = {} & [-f(-W \leq i \leq 0), f(1 \leq i \leq W)] \\ & 0 \leq t \leq 2W, \quad -W \leq i \leq W \end{aligned} \qquad (4)$$

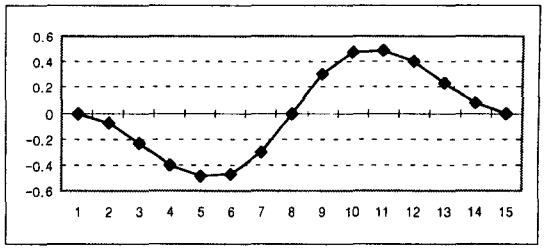

그림 1. 설계된 Ramp Edge의 최적 필터의 모양

i와 t는 정수이며, t는 배열의 인덱스와 대응된다. 본 논문에서는 W =7, s =1, A =0.41, 그리고 [K_1 ··· K_6]= [1.583, 1.468,-0.078,-0.872,-0.56]을 사용하며 그림1과 같은 모양을 갖는다.[3]

2.3 잔여 신호의 피크 에너지

연음과 빠른 발성 음성 신호와 같은 경우, 변화에 둔감한 에너지 특징 파라미터로만 음절을 구분 짓기 어렵기 때문에 이런 변화에 민감한 특성을 갖는 파라미터로 역 LPC 필터를 거친 잔여 신호의 피크 에너지가 사용될 수 있다. 잔여 신호는 음성 발성 기관의 소스 신호와 대응되며, 유성음은 임펄스 열로 무성음의 파열음은 임펄스, 그 외 무성음은 White Noise로 분류한다. 연음이나 빠른 발성에서의 소스 신호는 시작부분에서 강하게 나타나며 뒤로 갈수록 약한 특성을 가지며, 조음 기관의 조음에 의해서도 그 신호의 강약이 구분되는 특성을 갖는다. 역 LPC 필터는 식(5)와 같이 z-domain으로 표현했다.[3]

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad (5)$$

$$E(z) = GU(z) = A(z)S(z)$$

$H(z)$ 은 발성 모델의 전달함수이며, $S(z)$ 은 음성 신호, G 는 음량(Gain), $U(z)$ 는 소스 신호, a_i 는 LPC 계수, p 는 LPC 계수의 차수, $A(z)$ 은 역 LPC 필터, $E(z)$ 은 잔여 신호이다. 본 논문에서는 p=8을 사용하였다. 식(5)로 얻어진 잔여신호를 프레임당 피크 에너지를 구한다. 피크 에너지는 식(6)으로 표현한다.

$$x_e(t) = 10\log \underset{n_t \leq i < n_t+N}{Max} [e(i)] \qquad (6)$$

$x_e(t)$ 는 잔여신호의 프레임당 피크 에너지이며, $e(i)$ 는 입력 신호의 데이터, n_t 는 현재 데이터의 샘플의 위치, N는 한 프레임 길이, t 는 프레임 수를 나타낸다.

III. 끝점 검출 및 음절 분할 방법

본 논문은 Ramp Edge의 최적 필터를 끝점 검출과 음절 분할에 적용하기 위해 식(7)의 Moving Average 필터를 적용한다.

$$F(t) = \sum_{i=-W}^{W} h[W+i]x(t+i) \qquad (7)$$

$F(t)$ 는 Moving Average 필터의 출력 값이고, t 는 현재의 프레임 번호이다. 그리고 $h[\cdot]$ 는 Ramp Edge의 최적 필터 계수이며, $x(\cdot)$ 는 음성신호의 프레임당 에너지이다. 이렇게 구해진 $F(t)$ 의 출력 값은 [Qi Li외]이 소개한 3상태 끝점 결정 알고리즘을 사용하여, 입력 신호의 음성 구간의 시작점과 끝점을 검출하게 된다. 그림 2에 3상태 끝점 결정 알고리즘과 본 논문에서 제안하는 음절 분할을 포함한 상태 천이도를 나타내었다. 여기서 T_L와 T_H의 임계 값의 결정의 최대 출력 F값을 이용하여 T_L=-0.33*|F|, T_H= 0.33 *|F|으로 하였다. 그리고 GAP는 잔여음성 상태의 지속시간을 의미하며, 단위는 프레임 단위를 사용한다. 본 논문에서는 GAP를 20으로 하였다.

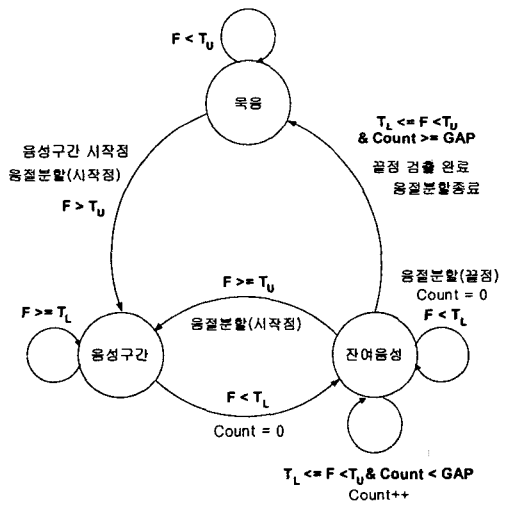

그림 2. 3상태 끝점 결정 및 음절 분할 상태 천이도

그림3에 끝점 결정 및 음절 분할 알고리즘으로 실제 음성 구간 검출 및 음절 분할을 나타내고 있다. 그림3에서 보는 것과 같이 '굉장히'의 발성 음성의 전체 음성 구간의 검출은 잘 되었다. 하지만, 음절 분할에 있어 3음절을 갖는 단어임에도 불구하고, '굉'과 '장히'로 2음절로 나뉘었다. '장히' 부분은 연음으로 발음되어 기존의 음성 신호의 에너지 정보만으로는 음절 분할이 어려움을 보여 주고 있다.

본 논문에서는 이 문제점을 해결하기 위해 식(5, 6)의 역 LPC 필터를 거친 음성 신호의 잔여 신호의 피크 에너지 정보를 이용하여 똑같이 Ramp Edge

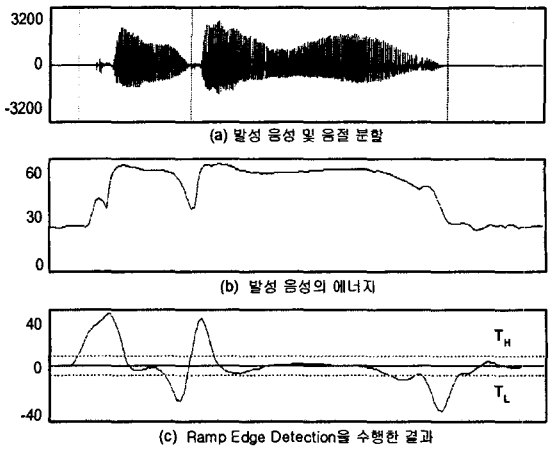

그림 3. 발성 음성 '굉장히'의 끝점 검출 및 음절 분할

Detection 방법을 사용하여 연음과 빠른 발성 음성의 음절의 문제점을 보안하였다. 그림4에 이 방법을 사용하여 '굉장히'의 발성 음성을 정확히 3음절로 분할한 것을 보여주고 있다. 여기서 T_L와 T_H의 임계 값의 결정의 최대 출력 F값을 이용하여, T_L=-0.25*|F|, T_H= 0.25 *|F|으로 하였다.

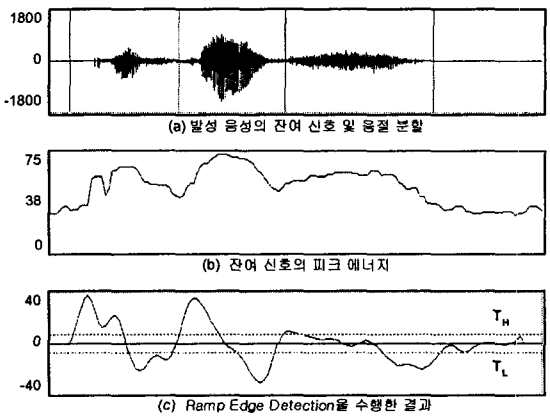

그림 4. 발성 음성 '굉장히'의 추가 음절 분할

IV. 실험 및 결과

본 논문의 음성 데이터는 남자 3명과 여자 2명으로부터 PBW(Phonetical Balanced Word)의 50단어를 선별하여, 실험실 잡음 환경에서 11KHz의 샘플링으로 음성 녹음 하였다. 한 프레임의 길이는 23ms이며, 9ms의 프레임 간격으로 분석하였다.

끝점 검출 및 음절 분할을 동시에 실험하기 위해 음성 구간의 앞 뒤로 15프레임의 묵음 구간을 삽입하여 Batch 모드로 실험하였다. 끝점 검출의 유무는 육안으로 확인 하였고, 음절 분할은 각 단어의 본래 음절 수와 자동으로 계산된 음절 수의 비교하였고, 분할 영역의 오류는 수작업으로 검증하였다.

표1에 끝점 검출과 음절 분할을 수행하여 얻어진 끝점 검출과 분할음절수의 결과를 보여주고 있다. 그리고 표2에는 최종 음절 분할의 검증 및 확인 결과를 보여주고 있다. 실험 결과 끝점 검출율은 100%로 얻었고, 음절 분할율은 90.7%을 얻었다. 각 단어에 대해 정확한 분할이 이뤄진 분할 완성률은 68%였으며, 30%의 분할 삭제와 2.4%의 분할 삽입 오류가 발생했다.

표 1. 끝점 검출과 분할 음절 수 (50단어)

화자	음절수	분할음절수	끝점 검출(%)
화자1(여)	185	161	100
화자2(여)	185	171	100
화자3(남)	185	167	100
화자4(남)	185	171	100
화자5(남)	185	169	100
평 균 (%)		90.7	100

표 2. 최종 음절 분할 결과 (50단어)

화자	단어	분할완성	삭제	삽입
화자1(여)	50	28	20	2
화자2(여)	50	35	13	2
화자3(남)	50	33	16	2
화자4(남)	50	38	12	0
화자5(남)	50	36	14	0
평 균 (%)		68	30	2.4

V. 결론

정확한 끝점 검출과 음절 분할은 음성인식의 성능과 계산량 감소에 많은 영향을 주는 전처리의 중요한 단계이다. 본 논문은 끝점 검출 방법과 음절 분할을 기존의 접근 방법과는 다른 음성 신호의 에너지와 잔여 신호의 피크 에너지만을 이용하여 Ramp Edge Detection 방법으로 수행하였다. 실험결과에서 보듯이 끝점 검출 100%, 음절 분할 90.7%라는 좋은 성능을 보였다. 따라서, 음성인식시스템에 적용하여 인식 성능의 향상을 기대해 본다. 그리고 음성인식 시스템에서 Ramp Edge Detection 방법을 실시간으로 처리하기 위해서는 T_L와 T_H의 임계 값의 결정을 잘 조절하는 것이 중요하다. 이 문제점은 향후 해결 과제로 남아있다.

<감사의 글>

본 연구는 한국과학재단 목적기초연구(R05-2002-001007-0)지원으로 수행되었음.

Reference

[1] Qi Li, Jinsong Zheng, Tsai, A., Qiru Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition", IEEE Transactions on Speech and Audio Processing, Vol. 10, Issue. 3, pp. 146-157, Mar 2002,

[2] Petrou, M., Kittler, J., "Optimal edge detectors for ramp edges", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, Issue. 5, pp. 483-491, May 1991,

[3] Thomas F. Q., "Discrete-Time Speech Signal Processing - Principles and Practice", Prentice Hall, 2002

[4] 문입섭, 박기영, 김종교, "대용량 음성인식을 위한 음소분할 알고리즘에 관한 연구", 공학연구, 제27권, pp. 99-104, 1997.

[5] 한학용, 고시영, 허강인, "우리말 연속음성의 음절 분할법", 한국음향학회지, 제20권, 제3호, pp. 70-75, 2001.