FIR filtering에 의한 끝점추출에 관한 연구*

A Study on the Endpoint Detection by FIR Filtering

이 창 영**
(Chang-Young Lee)

ABSTRACT

This paper provides a method for speech detection. After first order FIR filtering on the speech signals, we applied the conventional method of endpoint detection which utilizes the energy as the criterion in separating signals from background noise. By FIR filtering, only the Fourier components with large values of [amplitude × frequency] become significant in energy profile. By applying this procedure to the 445-words database constructed from ETRI, we confirmed that the low-amplitude noise and/or the low-frequency noise are separated clearly from the speech signals, thereby enhancing the feasibility of ideal endpoint detections.

Keywords: speech detection, endpoint detection, speech recognition, analysis

1. 서 론

음성 구간을 배경잡음으로부터 구분해 내는 작업은 음성신호 처리 전반에 걸쳐서 매우 중요한 요소이다. 예를 들어 TASI(Time Assignment Speech Interpolation)와 같은 통신 기술에서는, 주어진 채널 수의 수 배에 이르는 통화자들을 수용하기 위하여 통화자가 발성을 하고 있는 동안에만 채널을 할당해 주어야 하고, 따라서 통화자들의 idle time 즉 silence를 검출해 내는 것이 필수적이다 [1]. 음성인식에서도 끝점추출(endpoint detection)은 결코 소홀히 다루어질 수 없는 문제이다. 10 개의 숫자를 인식하는 한 실험에 의하면, 이상적으로 끝점을 추출했을 경우의 인식률이 93 %인데 반하여, 그 끝점의 양쪽을 60 ms 씩 증감시켰을 때는 3 %의 인식률 저하가 측정되었다 [2].

끝점추출을 어렵게 만드는 요인은 화자(speaker)의 발성법과 발성 환경의 두 가지로 구분할 수 있다. 전자의 경우로서, 화자가 입맛을 다신다던가 심호흡을 한다던가 또는 발성에 앞서서 건조한 입술을 열면서 여러 가지 잡음을 내는 일이 있다. 끝점추출은 대부분 입력된 음성 신호의 에너지를 조사함으로써 이루어지는데, 화자의 발성법에 따른 이러한

* 이 연구는 동서대학교 교내 학술연구비에 의해 지원되었음.
** 동서대학교 컴퓨터공학과

잡음들의 에너지는 음성신호에 비해 무시될 수 없는 크기이고, 따라서 끝점추출을 어렵게 만드는 것이다. 발성 환경에서도 여러 소음들이 불가피하게 개입된다. 자동차 소음과 같은 특수한 배경잡음이 존재하는 경우의 끝점추출에 대한 연구가 이루어지기도 했지만 [3], 일반적으로는 올바른 끝점추출을 위하여 소음이 없는 환경에서의 발성이 요구된다. 하지만 그러한 조용한 환경이 실현되기 어렵기도 하거니와, 그런 조용한 환경하에서 음성 신호가 입력된다 하여도, 마이크·녹음기·전화선 등의 기구나 전송장치에서 발생되는 소음 및 왜곡(distortion)도 적지 않다.

본 연구에서는, 하나의 소음 유형으로서 저주파 배경잡음이 존재하는 경우의 끝점 처리 방법을 다룬다. ETRI(전자통신연구소)에서 구축한 445 단어의 음성 DB에 대한 끝점추출을 수행하면서, 우리는 많은 음성 신호에 저주파의 배경잡음이 존재하고, 이로 인하여 종전의 에너지 계산에 의한 방법으로는 자동적인 끝점추출이 실패하는 경우가 적지 않다는 것을 발견하였다. 이에, 이러한 저주파 배경잡음을 어떻게 처리하여 적절한 끝점추출을 자동적으로 수행할 수 있을까 하는 문제에 접근하게 되었다.

2. 기존의 끝점추출 방법과 그 문제점

그림 1은 한 남성에 의해 발성된 단어 <갯수>의 파형을 나타낸다. 샘플링 주파수는 16 kHz이며 16 비트로 양자화 되었다. 이 파형에 $N=100$개의 데이터를 한 frame으로 하는 에너지를 계산한다. 이 계산 전에 음성 신호 $x(i)$에서 dc bias는 적절한 방법에 의해 제거되어야 한다. 일부 화자들이 음성신호 발성 이전에 내는 mouth click과 같은 순간적 잡음을 제거하기 위하여 여러 가지 부수적인 기술들이 동원되고 있지만 [4], 기본적으로는 식 (1)의 에너지가 어느 문턱값(threshold) 이상이 되는 조건을 찾는 것이 끝점추출의 내용이다. 그림 1에 연직으로 그어진 두 개의 선은 이상적인 양 끝점을 나타내고 있다.

$$E(j) = \sum_{i=j-N/2}^{j+N/2-1} x^2(i) \tag{1}$$

그림 2는 식 (1)에 의해 계산된 에너지를 보여주고 있다. 저주파 배경잡음으로 인하여 silence 구간의 에너지가 음성 신호에 비해 작지 않음을 볼 수 있다. 그림 2에 그어진 연직선으로부터 알 수 있듯이, 왼쪽의 이상적인 끝점을 추출하려면 $10^6 \sim 10^7$의 문턱값을 택하여야 하는데, 이 값은 화자에 따라 그리고 발성하는 단어에 따라 모두 다르다는 것이 문제이다. 그 값을 작게 선택하면 silence 구간이 너무 많이 포함되게 되며, 크게 선택하면 음성 신호가 누락되는 중대한 결함이 발생할 위험이 높아진다. 또한, 오른쪽의 이상적인 끝점추출은 사실상 불가능하다는 것을 그림으로부터 짐작할 수 있다.

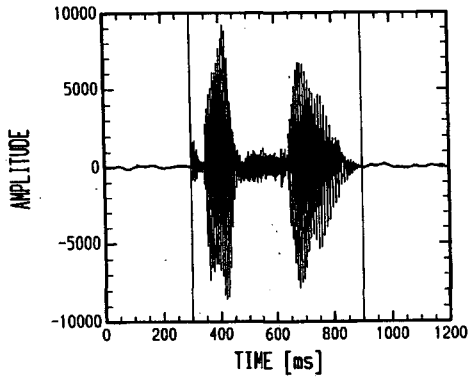

그림 1. 남성에 의해 발성된 단어 <갯수>의 파형

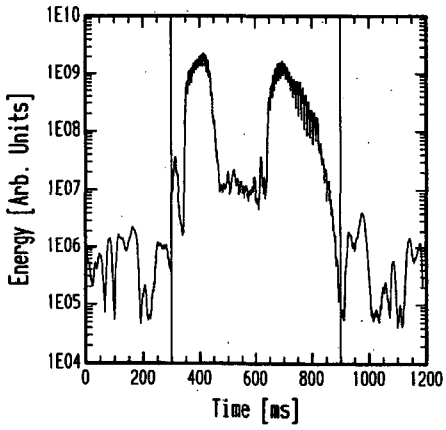

그림 2. 그림 1의 에너지

그림 3과 4는 다른 남성에 의해 발성된 <스위치>라는 단어의 파형과 그의 에너지를 보여주고 있다. 역시 두 연직선은 이상적인 양 끝점을 나타낸다. 그림 4의 왼쪽 부분의 배경잡음을 제거하려면 문턱값을 2×10^6 정도로 해야 하는데, 그것은 곧 /스/라는 음절의 탈락을 의미한다. 일반적으로 모음 /으/의 에너지는 noise level로부터 겨우 분간되는 정도로 작은데다가, 또한 <스위치>라는 단어의 경우에는 음절 /스/의 길이도 대단히 짧기 때문에 ($\approx 50\,\mathrm{ms}$), 그 경계를 명확하게 구분하는 끝점추출이 대단히 어려운 것이다.

요약하면, 기존의 방법을 사용하여 모든 음성 신호에 대해 만족스러운 끝점추출을 수행하는 문턱값을 찾기는 쉬운 일이 아닐 뿐더러, 심지어는 만족할 만한 문턱값이 존재하지 않을 수도 있다. 즉, 작은 문턱값을 택할 경우에는 저주파 배경잡음에 의해 너무 많은 silence 구간이 포함되어 끝점추출이 별 의미가 없게 되며, 이의 해결을 위해 문턱값을 증가시키면 /스/와 같은 낮은 에너지의 음성이 탈락되어 버리는 일이 생기는 것이다. 효과적인 신호 처리를 위하여 silence 구간을 제거하는 일은 대단히 긴요하긴 하지만, 한편으로는 음성신호 일부가 탈락되는 일이 발생하지 않도록 해야 하는데, 이러한 모순은 끝점추출에서 늘 존재하는 문제이다.

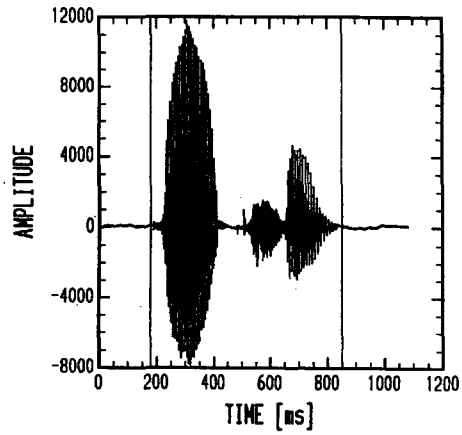

그림 3. 다른 남성에 의해 발성된 단어 <스위치>의 파형

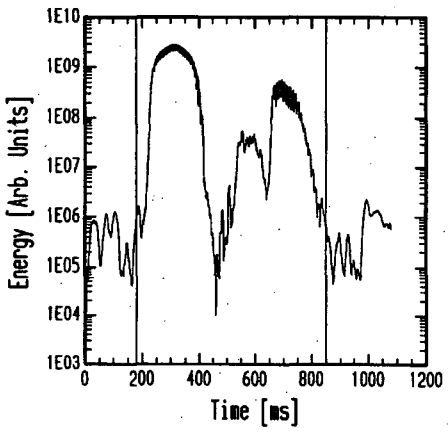

그림 4. 그림 3의 에너지

3. FIR filtering에 의한 끝점추출

저주파의 배경잡음을 해결하기 위한 방법으로서, 우리는

$$y(i) = x(i) - \mu x(i-\delta) \tag{2}$$

와 같이 변환된 데이터를 끝점추출에 시도하였다. μ와 δ는 조정될 수 있는 값이다. 위와 같은 변환의 배경은 이러하다. 저주파 잡음을 제거하기 위한 가장 직관적인 방법은 주어진 신호에 그 신호의 inverse를 더해주는 것인데, 그러면 그것은 곧 모든 신호의 소멸을 의미한다. 그리하여 신호를 δ만큼 shift시키고 그 inverse를 취한 후 factor μ를 곱하여 원래 신호에 더해주는 것이다. 이 처리는 식 (2)의 보다 일반적인 형태인 M차 FIR system, 즉

$$\tilde{x}(i) = \sum_{j=0}^{M} a_j x(i-j)$$

의 한 특수한 경우로서 해석될 수 있다. 이러한 filtering은 신호 처리에 있어서 더러 유용하게 쓰이는데, 한 예로 LPC와 같은 특징 벡터를 추출하는 데 앞서서 frame에 대한 windowing과 더불어 spectral flattening을 목적으로 다음과 같은 1차 FIR filtering이 적용된다 [5].

$$\tilde{x}(i) = x(i) - a x(i-1)$$

따라서 우리의 연구는 음성 신호에 FIR filtering을 적용한 후 끝점추출을 행하는 것으로 볼 수 있다.

또한, 식 (2)는 다른 견지에서도 해석 가능하다. 주어진 함수

$$x(t) = A\sin(\omega t) \tag{3}$$

가 있을 때, 이 함수를 δt만큼 shift시켜 그 inverse와 더하면

$$y(t) = A\,[\sin(\omega t) - \sin(\omega(t-\delta t)] \tag{4}$$

가 되는데, $\omega\,\delta t \ll 1$인 경우에 위 식은

$$\begin{aligned} y(t) &= A\,[\sin(\omega t) - \sin(\omega t)\cos(\omega\delta t) + \cos(\omega t)\sin(\omega\delta t)] \\ &\approx (A\omega\delta t)\cos(\omega t) \end{aligned}$$

로 근사시킬 수 있다.

식 (3)의 크기가 단지 진폭 A에 의해서만 결정되는 것과는 달리, 식 (4)에 의해 변환된 식의 크기는 $A\omega$에 비례해 주어진다. 따라서 음성 신호에 (4)의 처방을 하면 [Amplitude×Frequency]의 곱이 큰 Fourier 성분만이 의미가 있게 되고, 이는 곧 진폭이 무시될 수 없는 크기의 잡음일지라도 그 주파수가 작은 것은 변환에 의해서 제거됨을 의미한다.

식 (2)에서의 μ와 δ는 가장 효과적인 끝점추출을 위해 조정되어져야 하는 값이지만, 간단하게 $\mu=1$, $\delta=1$로 두어도 그 효과는 쉽게 나타난다. 그림 5는 그림 1을 $\mu=1$, $\delta=1$의 값을 사용한 식 (2)에 의해 변환한 그림이며, 그림 6은 그림 5의 파형에 대한 에너지를 보여주고 있다. 그림 1의 음성신호 양쪽에 구불구불하게 보였던 저주파 배경잡음이 그림 5에서는 평평하게 바뀐 것이 확실하게 보여지며, 그림 6으로부터 알 수 있듯이 silence 구간과 음성 신호 구간의 에너지가 선명하게 구분됨으로써, 음성 추출이 훨씬 더 효과적으로 수행될 것을 기대할 수 있다.

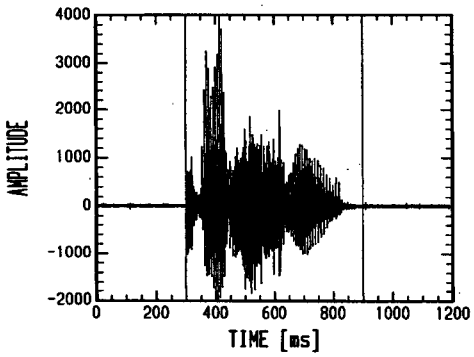

그림 5. 그림 1에 1차 FIR filtering을 적용하여 변환시킨 파형

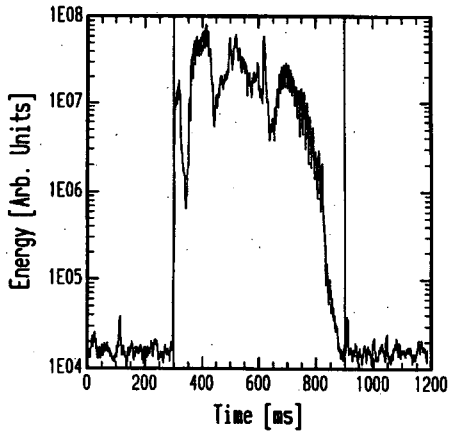

그림 6. 그림 5의 에너지

그림 7은 그림 3의 파형을 식 (2)에 의해 변환한 것이다. 역시 $\mu=1$, $\delta=1$의 값이 사용되었다. 저주파 배경잡음이 평평하게 됨을 다시 확인할 수 있다. 그림 8은 그림 7에 대한 에너지를 나타낸다. 그림 4의 왼쪽 연직선 근처에 있는 음절 /스/의 에너지가 저주파 배경잡음과 비슷한 정도인데 반해, 그림 8에서는 뚜렷한 구분이 지어지고 있음을 볼 수 있다. 이는 식 (2)의 FIR filtering에 의해, 저주파의 배경잡음이 상대적으로 고주파인 /스/음에 비해 어느 정도 약화된 결과에 따른 것이다. 그림 6과 그림 8로부터, FIR filtering을 통해 변환된 파형에 대해 $N=100$으로 계산한 식 (1)의 에너지가 대략 3×10^4의 값을 넘는 구간을 음성 신호 구간으로 택함으로써, 기존의 방법으로 처리가 어려웠던 끝점추출이 능률적으로 수행됨을 기대할 수 있다.

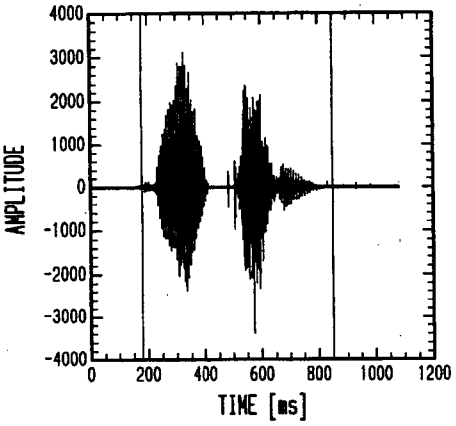

그림 7. 그림 3에 1차 FIR filtering을 적용하여 변환시킨 파형

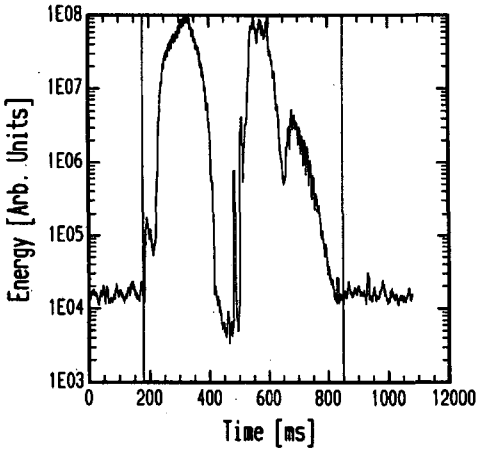

그림 8. 그림 7의 에너지

4. 요약 및 결론

종전의 끝점추출은 주어진 입력 음성신호 자체의 에너지가 어느 문턱값을 넘는가 하는 판별에 의해 이루어지는데, 저주파의 배경잡음이 있는 경우 그 방법은 실패할 가능성이 높다. 이를 극복하기 위해 문턱값을 증가시키면 모음 /으/와 같이 에너지가 작은 음성들이 끝점추출에서 탈락하는 일이 발생한다. 이처럼, 종전의 방법에서 단지 문턱값의 조절만으로는 끝점추출에 한계가 있는 것이다.

본 연구에서는 1차 FIR filtering으로 음성신호를 변환한 후, 이 변환된 신호에 대해 종전과 같은 끝점추출을 수행하는 방법을 제시하였고, 그 결과 기존의 방법에 비해 10 % 미만의 에러로 끝점추출이 수행됨을 확인하였다. 이러한 변환에 의해 음성 신호에서 (진폭×주파수)가 동시에 큰 성분만이 걸러진다. 즉, 에너지가 작은 잡음, 또는 에너지는 어느 정도 커도 주파수가 작은 배경잡음이 효과적으로 제거되는 것이다. ETRI의 445 단어 음성 DB에 본 연구에서 고안된 방법을 적용한 결과, 저주파 배경잡음이 변환에 의해 평평한 noise level로 바뀌어, 종전의 방법에 의해 구분이 어려운 경우에서도 끝점추출이 올바로 수행됨을 확인하였다.

참 고 문 헌

[1] *"Transmission Systems for Communications."* 1970. Bell Telephone Laboratories, Rev. 4th ed., New Jersey.

[2] Wilpon, J. G., Rabiner, L. R., and Martin, T. B. 1984. "An Improved Word-Detection Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints." *AT&T Tech. J.*, 63(3), 479-498.

[3] 서동권 외 4인. 1998. "음성인식을 위한 자동차 소음환경에서의 끝점추출." *한국음향학회지* 제 17권 1호, 76-79.

[4] Deller, J. R., Proakis, J. G., Hansen, J. H. 1993. *"Discrete-Time Processing of Speech Signals."* Macmillan, New York, 246-251.

[5] Rabiner, L. R. and Juang, B. 1993. "Fundamentals of Speech Recognition." Prentice-Hall, New Jersey, 112-117.

접수일자 : '99. 2. 12.
게재결정 : '99. 3. 25.

▲ 부산시 사상구 주례동 산 69-1
동서대학교 컴퓨터공학과
Tel: (051) 320-1572 (O), (051) 324-1325 (H)
Fax: (051) 312-2389
e-mail: seewhy@kowon.dongseo.ac.kr