말소리와 음성과학 제1권 제4호 (2009)
pp. 111~117

스펙트럼 패턴 기반의 잡음 환경에 강인한 음성의 끝점 검출 기법

Spectral Pattern Based Robust Speech Endpoint Detection in Noisy Environments

박 진 수[1] · 이 윤 재[2] · 이 인 호[3] · 고 한 석[4]
Park, Jinsoo · Lee, Yoonjae · Lee, Inho · Ko, Hanseok

ABSTRACT

In this paper, a new speech endpoint detector in noisy environment is proposed. According to the previous research, the energy feature in the speech region is easily distinguished from that in the speech absent region. In conventional method, the endpoint can be found by applying the edge detection filter that finds the abrupt changing point in feature domain. However, since the frame energy feature is unstable in noisy environment, the accurate edge detection is not possible. Therefore, in this paper, the novel feature extraction method based on spectrum envelop pattern is proposed. Then, the edge detection filter is applied to the proposed feature for detection of the endpoint. The experiments are performed in the car noise environment and a substantial improvement was obtained over the conventional method.

Keywords: endpoint detection, frame energy, spectral pattern, edge detection filter

1. 서 론

음성 끝점 검출은 마이크 입력 신호로부터 음성의 시작점과 끝나는 점을 구분하는 과정이다. 특히, 음성인식에서의 끝점 검출은 음성인식 성능에 큰 영향을 미치는 요소이다. 음성 끝점 검출을 통해 음성 구간만의 신호를 취함으로써, 음성인식에 소요되는 시간을 단축시킬 수 있으며 비음성 구간에 존재하는 잡음이 음성인식 성능을 하락시킬 수 있는 가능성을 줄일 수 있다. 하지만 잘못된 음성 검출은 음성인식에 필요한 음성 정보를 잃게 하여 음성인식 성능을 하락시키기도 한다. 따라서, 음성인식에서의 끝점 검출은 매우 중요한 분야라고 볼 수 있다.

음성 끝점 검출 알고리즘으로 가장 많이 알려져 있는 알고리즘은 Rabiner와 Sambur가 제안한 프레임 에너지 (frame energy)와 영교차율 (zero crossing rate)을 이용한 알고리즘으로 수학적 계산이 간단하며 음성의 기본적인 특징인 에너지와 주파수 성질을 잘 표현하는 장점이 있다[1][2]. Wilpon과 Rabiner는 미리 학습된 단어 HMM 모델과 잡음 HMM 모델의 유사도를 비교하여 음성 끝점 검출을 하였다[3]. 음성과 잡음이 주파수 대역에서의 데이터가 다르게 분포 된다는 점을 이용하여 엔트로피로 끝점 검출을 하도록 제안하였다[4]. 또한 음성의 검출을 음성의 활동 영역과 비활동 영역의 특징이 급격히 변하는 에지 (edge)를 찾아 내는 것으로 이해하고, 영상의 에지를 찾는데 사용되었던 Canny의 에지 검출기가 음성 검출에 응용되었다[5]. 이것을 기반으로 프레임 에너지 값으로 부터 에지성분을 찾기 위해 에지 검출 필터 (edge detection filter)와 상태 천이 (state transition)를 적용한 알고리즘이 개발되었다[6]. 하지만 프레임 에너지 특징은 잡음 환경에서 그 값의 변화가 크기 때문에 단순 입력신호로부터 프레임 에너지만을 이용한 음성 검출은 성능이 많이 하락하게 된다. 또한 이 프레임 에너지에 에지 검출필터를 적용한 결과 역시 불안정하게 되어 안정적인 끝점 검출이 어렵게 된다.

따라서 본 논문에서는 프레임 에너지 보다 잡음 환경에 더욱 강인한 특징을 개발, 에지검출 필터를 적용하여 음성 검출을 수행한다. 프레임 에너지와 같이 본 논문에서 사용된 특징은 그 특징 자체로 문턱치 (threshold)를 적용하여 음성 검출이 가능하나, 에지 검출 필터가 가지는 장점으로 제안하는 알고리즘을 보

1) 고려대학교 jspark@ispl.korea.ac.kr
2) 고려대학교 yjlee@ispl.korea.ac.kr
3) 고려대학교 ihlee@ispl.korea.ac.kr
4) 고려대학교 hsko@korea.ac.kr, 교신저자
본 연구는 서울시 산학연 협력사업(WR080951)의 연구결과로 수행되었습니다.

접수일자: 2009년 9월 14일
수정일자: 2009년 10월 23일
게재결정: 2009년 10월 25일

완하기 위해 에지 검출 필터를 적용하였다. 개발된 특징은 비음성 구간에서의 잡음 신호의 스펙트럼 포락선 패턴 (spectrum envelope patten)과, 음성구간에서 잡음에 음성이 부가된 신호의 스펙트럼 포락선의 패턴이 달라질 것으로 예상하여 기존의 잡음 인식 (noise classification)에서 사용된 스펙트럼 포락선 기반의 특징 추출기법을 적용하여 개발하였다.

2장에서는 기존의 음성 끝점 검출 알고리즘에 대하여 설명하였고, 3장에서는 본 논문에서 제안한 스펙트럼 패턴 기반의 음성 끝점 검출 알고리즘에 대하여 설명하였다. 4장에서는 실험결과를 비교하여 성능을 평가하였다. 마지막으로 5장에서 본 논문의 결론을 맺는다.

2. 기존의 음성 끝점 검출 알고리즘

2.1 프레임 에너지와 영교차율

조용한 환경에서 가장 효과적으로 사용할 수 있는 방법은 프레임 에너지 기반에 영교차율을 고려한 음성 끝점 검출 기법이다[1][2]. 일반적으로 에너지 값은 음성 구간에서 크고, 비음성 구간에서 작게 나타나므로 이러한 성질을 이용하여 문턱치와 비교하여 음성, 비음성을 구별한다.

영교차율은 프레임 구간 안에서 신호 파형이 0값을 통과하는 회수를 말하며, 모음이나 유성음 구간에서 상대적으로 비음성 구간에 비해 작은 값을 나타낸다. 실제 에너지로만 음성과 비음성 구간을 구분하기 힘든 마찰음이나 파열음의 경우, 영교차율이 유성음 보다 크다는 사실을 바탕으로 프레임 에너지에 의해 검출된 결과에 영교차율을 이용하여 결과를 보정해준다. 위의 방법은 비교적 수학적 계산이 간단하며 음성의 기본적인 특징인 에너지를 잘 표현하는 장점이 있지만, 잡음 환경에서 프레임 에너지와 영교차율만 이용한 음성 끝점 검출은 상대적으로 좋은 성능을 야기하지 못하게 된다. 잡음 환경에서는 비음성 구간에서도 높은 에너지 값을 가지는 경우가 있어 정확한 문턱치를 찾기가 힘들며 에너지 값의 편차가 커서 음성과 비음성 구간의 구분이 어려운 단점이 있다.

2.2 에지 검출 필터와 상태 천이를 이용한 끝점 검출

기존의 에지 검출 필터를 이용한 끝점 검출에서는 먼저, 음성 시작 구간에서는 에너지가 커지고 음성이 끝나는 구간에서는 에너지가 감소하는 성질을 이용하여 프레임 에너지 값의 변화가 큰 에지성분을 찾기 위해 에지 검출 필터가 사용되었다. 그리고 에지 검출 필터의 결과를 상태 천이 모델에 적용하여 최종적인 음성의 끝점을 구하였다[6]. 에지 필터는 그림 1과 같이 원점에 대칭인 필터이기 때문에 비음성 구간에서의 에너지 특징의 값이 그 크기에 상관없이 일정하면 필터 결과값이 0에 가까운 결과가 나오다가 음성이 존재하여 에너지 특징값이 커지면 필터 출력 역시 커지게 되며 에너지 특징값이 작아지면 필터 출력 역시 작아지게 된다. 따라서 잡음의 크기에 따라 문턱치를 조절할 필요가 없으며 서서히 변하는 특징값에도 강인한 장점을 가지게 된다.

에지 검출 필터 h는 식(1)과 같다.

$$f(x)=e^{Ax}[K_1\sin(Ax)+K_2\cos(x)+e^{-Ax}[K_3\sin(Ax)+K_4\cos(x)]+K_5+K_6e^{sx}$$

$$h(i)=\begin{cases}-f(i), & -W\le i\le 0\\ f(i), & 1\le i\le W\end{cases} \quad (1)$$

여기서 W는 필터길이와 관계되는 변수이며, i는 $-W$ 부터 W까지 정수, A와 K는 필터 파라미터이다. <그림1>은 $W=7$ 일때의 필터 응답 그림이다.

$(A=0.41, K_1=1.538, K_2=1.468, K_3=-0.078, K_4=-0.036, K_5=-0.872, K_6=-0.56)$

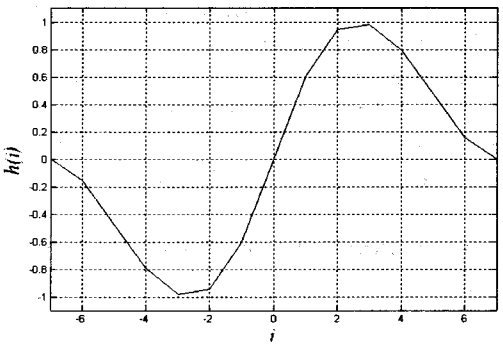

그림 1. 에지 검출 필터 h ($W=7$의 경우)
Figure 1. Edge detection filter h ($W=7$)

프레임 에너지 $g(n)$에 에지 검출 필터 h를 적용하여 출력 $F(n)$를 식(2)와 같이 구할 수 있다.

$$F(n)=\sum_{i=-W}^{W}h(i)g(n+i) \quad (2)$$

여기서 n은 프레임 인덱스를 의미한다. $F(n)$에 상태 천이의 동작을 통하여 음성의 시작점과 끝나는 점을 찾을 수 있다. <그림2>는 시작점과 끝나는 점을 찾기 위한 상태 천이 모델이다[6].

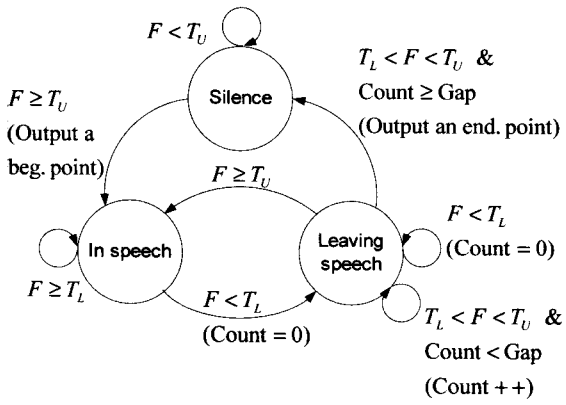

그림 2. 끝점 검출을 위한 상태 천이 모델
Figure 2. State transition diagram for endpoint detection

<그림2>에서 Silence는 비음성 구간을 나타내고 In speech는 음성구간을 나타낸다. Leaving speech는 음성구간이지만 비음성 구간으로 변할 수 있는 단계이다. T_L은 낮은 문턱치 (lower threshold), T_U는 높은 문턱치 (upper threshold), Gap은 끝나는 점을 결정하기 위한 허용치로써 실험적으로 정하는 상수이다. 단, T_U는 항상 T_L 보다 커야 한다.

위 상태 천이 모델을 이용하면, $F(n)$이 T_U 보다 작으면 음성이 없는 비음성 구간 (Silence)으로 판단 한다. $F(n)$이 T_U 보다 커지면 음성이 시작된 것으로 보고 그 부분을 시작점 (In speech)으로 잡는다. $F(n)$이 T_L 보다 작아지면 아직 음성 구간이긴 하지만 비음성 구간으로 바뀔 가능성 (Leaving speech)이 있다고 간주하고, Count ($F(n)$이 T_L와 T_U 사이에 있는 경우 연속적으로 그 사이에 있는 횟수)를 0으로 잡는다. Count가 Gap 보다 작으면 Leaving speech로 판단하고, Count가 Gap 보다 크면 Silence로 판단한다. Silence로 판단되는 그 프레임이 끝나는 점이 된다. $F(n)$이 T_L 보다 작아지면 Count를 0으로 잡고 Leaving speech 단계를 유지한다. 그리고 $F(n)$가 T_U 보다 커지면 다시 In speech 구간으로 돌아간다.

음성의 시작점과 끝나는 점을 검출하기 위해 프레임 에너지에 위 기법을 적용한 결과를 <그림3>과 <그림4>에 나타내었다. <그림3>은 깨끗한 환경에서 녹음된 음성 신호의 프레임 에너지에 에지 검출 필터와 상태 천이 모델을 적용하여 끝점 검출한 결과이고, <그림4>는 음성정보기술산업지원센터 (Speech Information Technology & Industry Promotion Center)에서 현재 배포중인 자동차 잡음 환경 음성 데이터베이스 (Car03) 중의 고속주행 환경에서의 입력 신호의 프레임 에너지에 에지 검출 필터와 상태 천이 모델을 적용하여 끝점 검출한 결과이다. 적용한 필터 길이 $W=3$ 이다.

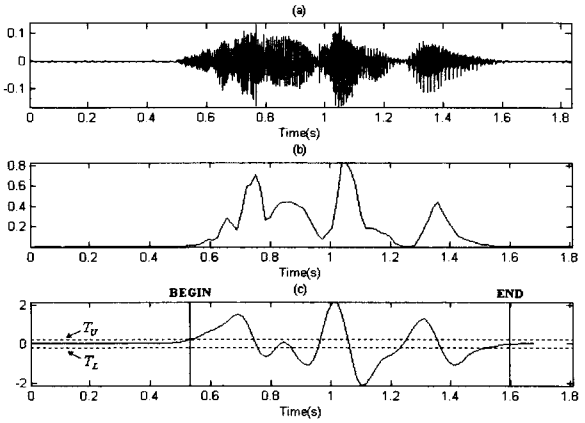

그림 3. 깨끗한 환경의 음성 신호의 프레임 에너지와 에지 검출 필터 및 상태 천이 모델을 적용한 결과
(a)음성 신호 /수신정보/에 대한 정규화된 시간축 파형
(b)음성 신호의 프레임 에너지 (c)에지 필터와 상태 천이 모델을 이용한 음성 끝점 검출결과
Figure 3. Results of the frame energy and edge detection filter based endpoint detection in clean environment

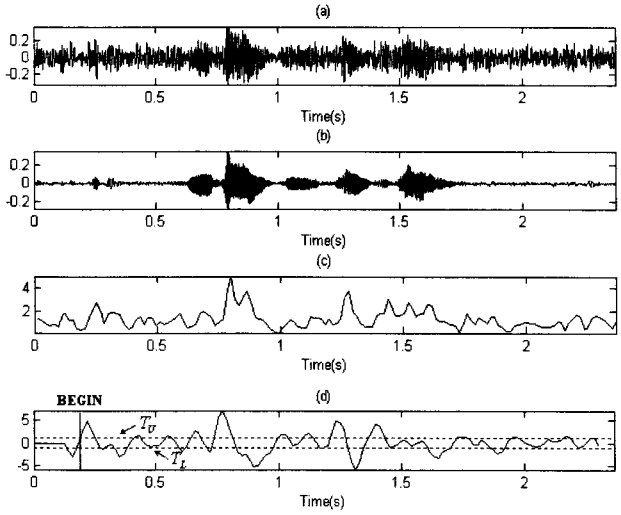

그림 4. 자동차 잡음 환경의 음성 신호의 프레임 에너지와 에지 검출 필터 및 상태 천이 모델을 적용한 결과.
(a)잡음 환경 음성 신호 /출발지설정/에 대한 정규화된 시간축 파형
(b)신호(a)에 주파수 차감법을 적용한 음성 신호에 대한 수동적 끝점 검출 결과
(c)음성 신호의 프레임 에너지
(d)에지 필터와 상태 천이 모델을 이용한 음성 끝점 검출결과
Figure 4. Results of the frame energy and edge detection filter based endpoint detection in car noisy environment.

<그림3>에서와 같이 깨끗한 환경에서의 신호는 프레임 에너지 값도 문턱치를 적용하여 음성 검출이 가능한 정도의 성능이 나오며, 에지 필터 및 상태 천이 모델을 적용하여 음성 검출을 한 결과 역시 잘 나온다는 것을 알 수 있다.

<그림4>의 (a)는 자동차 잡음 환경의 음성 신호, (b)는 (a)신호에 주파수 차감법[7]을 적용하여 수작업으로 음성의 시작점과 끝나는 점을 구한 결과이다. 실험에 사용한 음성 데이터베이스는 깨끗한 음성이 존재하지 않기 때문에 수작업의 검출에 어려움이 있다. 따라서 보다 정확한 음성 검출을 위해 주파수 차감법을 이용하여 잡음 음성을 깨끗한 음성으로 추정한 후 수작업으로 시작점과 끝나는 점을 구하였다. <그림4>의 경우와 같이 잡음 환경에서 프레임 에너지는 그 값이 안정적으로 나오지 못해 좋은 결과를 야기하지 못하며, 에지필터 결과 역시 변동이 큰 결과가 나오기 때문에 음성의 시작점은 크게 벗어난 구간에서 검출되며 끝나는 점도 검출하지 못하는 것을 확인 할 수 있다.

3. 스펙트럼 패턴 기반의 음성 끝점 검출 알고리즘

2장에서는 잡음 환경에서 음성의 끝점 검출을 위하여 프레임 에너지에 영교차율을 고려한 기법, 프레임 에너지에 에지 검출 필터를 고려한 기법을 알고리즘과 실험을 통하여 소개하였다. 하지만 이 기법들은 잡음 환경에서 불안정한 성능의 결과를 보이는 단점이 있다.

따라서 본 논문에서는 단순히 프레임 에너지에 에지필터를 적용하는 것이 아니라 보다 안정적인 특징 추출 기법을 적용하여 잡음에 강인한 끝점 검출 알고리즘을 제안한다. 제안한 특징 추출 기법은 패턴인식에 기반한 특징 추출기법으로 잡음 인식에 사용된 스펙트럼 포락선 기반의 기법을 도입하였다. 잡음의 스펙트럼 포락선의 패턴과 잡음에 음성이 부가된 신호의 스펙트럼 포락선의 패턴이 다르다는 점에 착안하여, 추정된 잡음의 특징과 입력 신호의 특징간의 거리 (distance)를 새로운 특징으로 사용한다. 잡음구간에서는 입력신호의 특징과 추정된 잡음의 특징의 패턴이 유사하여 제안한 특징 값이 작게 나올 것이고, 음성구간에서는 특징의 패턴 변화로 인해 제안한 특징 값이 크게 나올 것이다. 이 특징 값에 에지 검출 필터를 적용하여 최종적인 끝점 검출을 수행한다.

3.1 패턴인식을 위한 특징 추출 기법

제안한 특징은 잡음 인식에 쓰이는 특징으로서 주파수 스펙트럼 포락선의 패턴 기반으로 평균주파수 (mean frequency), 평균주파수를 사이에 두고 저주파 대역과 고주파 대역의 포락선을 라인 피팅 (line fitting)하는 계수를 이용하여 특징을 추출한다.[8]

평균주파수는 입력 신호가 가지는 주된 성분의 주파수를 표현해주는 것으로 식(3)과 같이 수식적으로 나타낼 수 있다.

$$F_{mean} = \frac{\sum_{k=1}^{N/2} |X(k)|}{\sum_{k=1}^{N/2} \frac{|X(k)|}{k}} \tag{3}$$

여기서 $|X(k)|$는 입력 신호의 퓨리에 변환을 통한 k 번째 주파수 인덱스에서의 절대값 크기를 나타내며, N은 퓨리에 변환 사이즈를 나타낸다.

평균주파수를 기준으로 저주파 대역과 고주파 대역의 스펙트럼 포락선을 구별하여 각각의 라인 피팅 방정식을 식 (4)와 같이 구한다.

$$y(k) = a_0 + a_1 \log_2(k) \tag{4}$$

저주파수 대역의 포락선을 라인 피팅하는 계수 a_0와 a_1을 식 (5)로 계산한다.

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{L} \frac{1}{k} & \sum_{k=1}^{L} \frac{\log_2(k)}{k} \\ \sum_{k=1}^{L} \frac{\log_2(k)}{k} & \sum_{k=1}^{L} \frac{(\log_2(k))^2}{k} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^{L} \frac{|X(k)|}{k} \\ \sum_{k=1}^{L} \frac{|X(k)|\log_2(k)}{k} \end{bmatrix} \tag{5}$$

여기서, L은 평균주파수에 해당하는 주파수 인덱스를 의미한다. 저주파수 대역의 라인 피팅은 k를 1부터 L까지 계산한 것의 결과를 사용하며, 고주파수 대역의 라인 피팅은 k를 $L+1$부터 $N/2$까지 계산한 결과를 이용한다. <그림5>는 입력 신호의 스펙트럼 절대값과 그 값을 라인 피팅 한 그림이다.

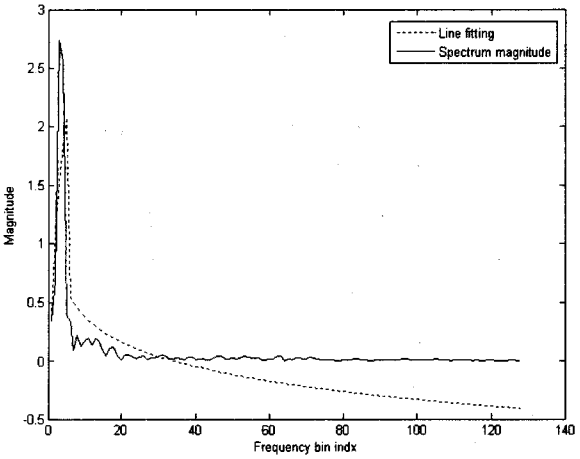

그림 5. 음성 신호의 스펙트럼 절대값과 라인 피팅 결과
Figure 5. Result of the line fitting to the spectrum magnitude

최종적으로 평균주파수, 저주파 대역에서의 라인 피팅을 위한 a_0와 a_1, 고주파 대역에서의 라인 피팅을 위한 a_0와 a_1을 5차 특징으로 구성한다.

3.2 끝점 검출을 위한 특징 추출 기법

3.1에서 구한 특징은 잡음만 있는 비음성 구간에서는 추정된 잡음으로부터 구한 특징과 유사한 패턴의 특징이 추출될 것이고, 음성이 부가되면 추정된 잡음으로부터 구한 특징과 패턴이 많이 달라지게 될 것이다. 이러한 성질을 이용하여 추정된 잡음

으로부터 구한 5차 특징과 입력 신호간의 5차 특징간의 거리를 새로운 특징으로 구한다.

n번째 프레임에서의 특징은 식(6)과 같다.

$$dist(n) = \sqrt{\sum_{i=1}^{P}(input(i) - noise(i))^2} \quad (6)$$

P는 특징 벡터의 차원 수를 의미하며 $input$과 $noise$는 각각 입력신호의 특징과 추정된 잡음의 특징을 의미한다.

<그림6>은 저속주행 환경에서 수집한 음성 데이터를 대상으로 제안한 특징에 에지 검출 필터를 적용한 경우 실제 검출 결과를 보여주고 있다. <그림6>에서 알 수 있듯이 거리가 비음성 구간에서는 아주 작고 음성 구간에서는 크게 나오는 것을 확인할 수 있으며, <그림4>와 비교하여 프레임 에너지 보다 변동이 작은 안정적인 특징임을 확인 할 수 있다. 또한 <그림6>에서의 (c)와 같이 에지 검출 필터를 적용하지 않아도 제안한 특징 자체만으로 문턱치를 적용하여 끝점 검출이 가능함을 예상 할 수 있다. 제안한 특징 추출법은 추정된 잡음과 입력신호간의 유사성을 고려하는 것이기 때문에 입력 잡음의 에너지 크기에 영향을 거의 받지 않는 장점을 가진다. 그러나 2.2에서 언급한 에지 검출 필터의 장점 때문에 에지 검출 필터를 적용하면 보다 안정적인 결과를 얻을 수 있다.

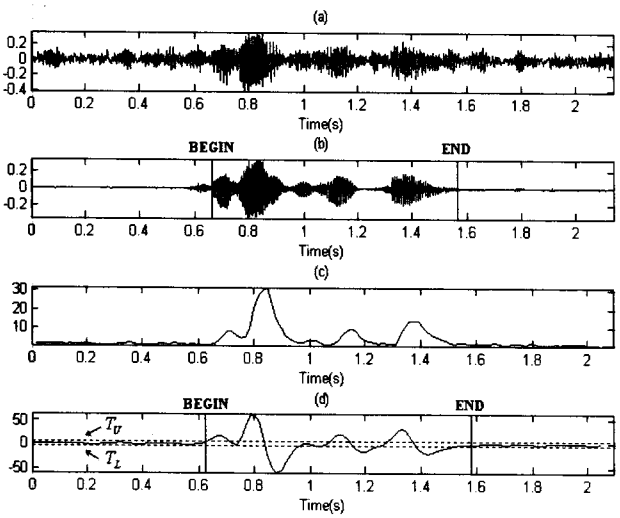

그림 6. 자동차 잡음 환경의 음성 신호에 제안한 알고리즘을 적용한 결과
(a)잡음 음성 신호 /출발지설정/에 대한 정규화된 시간축 파형
(b)신호(a)에 주파수 차감법을 적용한 음성 신호에 대한 수동적 끝점 검출 결과
(c)제안한 특징 결과
(d)에지 검출 필터 적용 및 음성 끝점 검출결과
Figure 6. Proposed algorithm result in car noise environment

본 논문에서는 입력 신호와의 비교를 위한 잡음 특징을 추정하는 방법으로 음성의 처음 몇 프레임은 항상 비음성 구간이라고 가정하는 방식을 사용하였다. 일반적으로 처음 10 프레임 정도로 정하며 이 구간에서 신호의 평균을 취하여 잡음을 추정할 수 있다. 하지만 이러한 방식은 시간에 따라 변하는 비정상(non-stationary) 잡음에 강인하지 못한 단점을 보이게 된다. 본 논문에는 적용하지 않았으나 적응적인 잡음 추정방식을 도입한다면 보다 좋은 끝점 검출 결과를 예상 할 수 있을 것이다.

4. 실험 및 결과

4.1 실험 환경

본 논문의 실험을 위하여 음성정보기술산업지원센터에서 현재 배포중인 자동차 잡음 환경 음성 데이터베이스(Car03)를 이용하였다. 음성 데이터베이스는 자동차의 저속주행(40-60km/h)과 고속주행(70-90km/h) 환경에서 sun-visor의 center 마이크에서 수집한 데이터를 이용하였다. 사용한 음성 데이터베이스는 8kHz로 다운샘플링 된 데이터를 이용하였으며 저속주행과 고속주행 환경에 따라 각각 한국어 음성 200개, 총 400개의 단어에 대해 실험하였다. 저속주행과 고속주행 환경에 대한 잡음의 강도를 나타내기 위해 각 환경에 대한 신호 대 잡음비 (Signal to Noise Ratio)를 구하였다. 음성 데이터의 에너지에 추정한 잡음의 에너지를 빼주어 깨끗한 신호의 에너지를 구하고, 깨끗한 신호와 추정한 잡음의 에너지비로 신호 대 잡음비를 구하였다. 음성구간에서 구한 신호대 잡음 비는 저속주행의 경우 -3dB, 고속주행의 경우 -8dB이다. 특징 추출을 위한 프레임 길이는 256 샘플(32ms)이며 프레임 이동 간격은 128샘플(16ms)로 하였다.

4.2 실험 결과 및 토의

본 논문에서 제안한 알고리즘의 성능을 평가하기 위해 사전에 각각의 음성 데이터에 대하여 수작업으로 음성의 시작점과 끝나는 점을 구하였다. 수작업으로 검출된 시작점과 끝나는 점은 제안한 알고리즘으로 검출한 시작점과 끝나는 점의 정확도에 대한 비교 기준으로 사용하였다. 단 본 논문에서는 프레임 단위로 시작점과 끝나는 점을 구별하기 때문에 결과 비교를 위하여 수작업 역시 프레임 이동 간격을 맞추어 끝점 검출을 하였다. 성능평가를 위해서 프레임 에너지에 에지 검출 필터를 적용한 기법과 제안한 기법을 이용하여 검출한 결과를 수작업으로 검출한 결과와 비교하여 검출여부를 판단하였으며 또한 수작업 검출결과와의 오차 정도도 계산하였다.

error는 실제로 음성구간이지만 비음성 구간으로 판단한 경우를 의미하며 음성을 훼손하는 구간을 의미한다. 검출된 시작점이 수작업으로 검출한 시작점보다 뒤에 위치하거나, 검출된 끝나는 점이 수작업 결과보다 앞쪽에 위치하면 error로 인정한다. detection은 error 가 아닌 경우, 또는 음성 구간을 훼손하지 않은 경우를 의미하며 시작점 또는 끝나는 점이 정확히 수작업

결과와 같거나 조금 여유있게 판단한 경우를 포함한다. 단순히 error와 detection이 일어난 횟수를 확인하는 것이 아니라 error와 detection의 경우, 평균적으로 몇프레임 내에서 발생하였는지에 대해 <표1>, <표2>에 표현하였다.

<표1>은 400개의 자동차 잡음 환경 음성 데이터에 대해 프레임 에너지에 에지 검출 필터를 적용한 결과와 수작업으로 검출한 결과를 비교한 결과이다. 음성의 시작점과 끝나는 점의 detection 개수가 error 개수보다 높은 수치를 나타내고 있지만 수작업 결과와의 평균 오차는 detection의 경우 시작점은 9.6프레임, 끝나는 점은 9.3프레임이 나왔으며 error의 경우 평균 시작점은 5.8프레임, 끝나는 점은 8.5프레임으로 수작업과의 오차가 매우 큰 것을 확인 할 수 있다. 이는 <그림4>와 같이 끝점 검출 결과가 매우 불안정하며 신뢰성이 떨어진다는 것을 보여준다.

<표2>는 제안한 기법의 결과와 수작업으로 검출한 결과를 비교한 것이다. 음성의 시작점과 끝나는 점의 error 개수가 프레임 에너지 결과보다 낮은 수치를 나타내고 있다. 또한 detection의 경우, 프레임의 평균 오차값이 시작점은 0.7프레임, 끝나는 점은 2.2프레임이며 error의 경우 시작점은 1.8프레임, 끝나는 점은 2.4프레임으로 평균 프레임 수치가 작다는 것을 알 수 있다. 이 결과를 통해 제안한 기법이 프레임 에너지를 이용한 것보다 수작업 결과와 보다 유사한 끝점 검출 결과를 야기시킨다고 판단할 수 있다.

표 1. 기존의 프레임 에너지를 이용한 결과
Table 1. Result of the conventional method

	시작점 (개)	시작점 평균오차 (프레임)	끝나는점 (개)	끝나는점 평균오차 (프레임)
error	116	5.8	180	8.5
detection	284	9.6	220	9.3

표 2. 제안한 알고리즘을 이용한 결과
Table 2. Result of the proposed method

	시작점 (개)	시작점 평균오차 (프레임)	끝나는점 (개)	끝나는점 평균오차 (프레임)
error	48	1.8	32	2.4
detection	352	0.7	368	2.2

<그림7>은 고속주행 환경에서 수집한 음성 데이터에 대해 실험한 또 다른 결과 그림이며 제안한 알고리즘 결과가 detection에 해당되는 경우이다. 아래 그림은 수작업 결과와 비교하여 시작점은 0 프레임, 끝나는 점은 3프레임 차이가 난 결과이다. 이 결과를 통하여 제안한 음성의 끝점 검출 알고리즘이 잡음의 강도가 높은 자동차 잡음 환경에서도 좋은 성능을 보인다는 것을 알 수 있다.

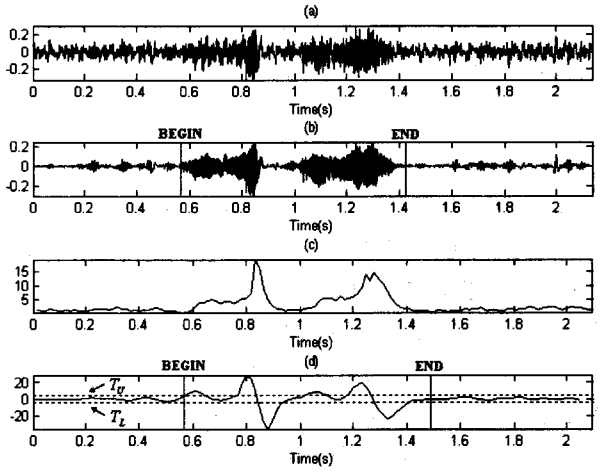

그림 7. 제안한 알고리즘의 끝점 검출결과
(a)잡음 음성 신호 /입력전환/에 대한 정규화된 시간축 파형
(b)신호(a)에 주파수 차감법을 적용한 음성 신호 /입력전환/에 대한 수동적 끝점 검출 결과
(c)제안한 특징 결과 (d)에지 검출 필터 적용 및 음성 끝점 검출결과

Figure 7. Result of endpoint detection for proposed algorithm

본 논문의 실험은 2.4GHz 컴퓨터와 Matlab을 이용하여 실험하였으며 기존의 에너지 기반의 기법은 프레임당 0.0013초가 걸렸으며 제안한 기법은 0.0021초가 걸렸다. 제안한 기법의 연산이 기존의 기법보다 많지만 matlab을 이용한 실시간 (실시간 프레임 이동 속도: 0.016초)에는 문제가 되지 않는 것을 확인하였다.

또한 본 논문에서는 제안한 기법을 자동차 환경 잡음에 대해서만 평가를 하였다. 제안한 특징 추출은 잡음 인식에 사용된 특징 추출 기법에 기반을 한 것이며 기존의 잡음 인식에서는 11가지 잡음 환경에 대해 본 논문의 특징과 함께 한가지 특징을 더 추가하여 90% 에 가까운 인식률을 얻었다[8]. 이는 제안한 기법의 특징이 다양한 잡음의 스펙트럼 패턴을 분별하는데 효과적임을 의미하며, 음성 검출에서도 효과적으로 사용될 수 있을 것이라 예상된다.

5. 결 론

본 논문에서는 자동차 잡음 환경에서 음성의 끝점 검출률 향상을 위한 새로운 음성 끝점 검출 기법을 제안하였다. 제안한 기법은 비음성 구간에서의 잡음과 음성구간에서 잡음에 음성이 부가된 신호의 스펙트럼 포락선의 패턴이 달라질 것으로 예상하여 기존의 잡음 인식에서 사용된 스펙트럼 포락선 기반의 특징 추출기법을 적용하였다. 일반적인 프레임 에너지를 특징으

로 음성 검출을 하는 것보다 강인한 특징이 될 수 있음을 본 실험을 통하여 확인하였다. 향후에는 비정상적 잡음에도 효과적인 끝점 검출을 위해 적응적인 잡음추정 기법과 연동시키는 실험을 진행할 계획이다.

참고문헌

[1] Labiner, L. R., Sambur, M. R., (1975). "An algorithm for determining the endpoints for isolated utterance", *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297-315, Feb.

[2] Labiner, L. R., Juang, B. H., (1993). *Fundamentals of speech recognition*, pp. 460-461, Prentice Hall.

[3] Wilpon, J. G., Labiner, L. R.., (1987). "Application of hidden Markov models to automatic speech endpoint detection", *Computer Speech and Language*, no. 2, pp. 321-341, Nov.

[4] Wu, B. F., Wang, K. C., (2005). "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 762-775, Sept.

[5] Li, Q., Tsai, A., (1999). "A matched filter approach to endpoint detection for robust speaker verification," in *Proc. IEEE Workshop on Automatic Identification*, Summit, NJ, Oct.

[6] Li, Q., Zheng, J., Tsai, A., Zhou, Q., (2002). "Robust endpoint detection and energy normalization for real-time speech and speaker recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146-157, March.

[7] Boll, S. F., (1979). "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, April.

[8] Kates, J. M., (1995). "Classification of background noises for hearing-aid applications," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461-470, Jan.

• **박진수 (Park, Jinsoo)**
고려대학교 대학원 지능신호처리 연구실
서울시 성북구 안암동 5가
Tel: 02-926-2909 Fax: 02-3291-2450
Email: jspark@ispl.korea.ac.kr
관심분야: 음성검출, 음성 신호처리
2008~현재 고려대학교 대학원 바이오마이크로시스템기술 협동과정 석박사통합과정 재학중

• **이윤재 (Lee, Yoonjae)**
고려대학교 대학원 지능신호처리 연구실
서울시 성북구 안암동 5가
Tel: 02-926-2909 Fax: 02-3291-2450
Email: yjlee@ispl.korea.ac.kr
관심분야: 음성 신호처리, 음성인식, 반향제거
2003~2009 고려대학교 대학원 전자컴퓨터공학과 석박사통합과정 졸업
2009~현재 고려대학교 대학원 박사후 연구원

• **이인호 (Lee, Inho)**
고려대학교 대학원 지능신호처리 연구실
서울시 성북구 안암동 5가
Tel: 02-926-2909 Fax: 02-3291-2450
Email: ihlee@ispl.korea.ac.kr
관심분야: 멀티채널 잡음제거, 음성 신호처리
2008~현재 고려대학교 대학원 영상정보처리 협동과정 석박사통합과정 재학중

• **고한석 (Ko, Hanseok)** 교신저자
고려대학교 전기전자전파공학과
서울시 성북구 안암동 5가
Tel: 02-3290-3239 Fax: 02-3291-2450
Email: hsko@korea.ac.kr
관심분야: 영상 및 음성 신호처리, 패턴인식, 데이터 융합
1995~현재 고려대학교 전기전자전파공학과 교수