한국음향학회지 제32권 제2호 pp. 147~156 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.2 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.2.147

pISSN : 1225-4428
eISSN : 2287-3775

인간로봇 상호작용을 위한 잡음환경에 강인한 음성 끝점 검출 기법

Robust Speech Endpoint Detection in Noisy Environments for HRI (Human-Robot Interface)

박진수, 고한석*†

(Jin-Soo Park* and Han-Seok Ko**)

고려대학교 바이오마이크로시스템기술 협동과정,
*고려대학교 전기전자전파공학부

(접수일자: 2012년 2월 3일; 수정일자: 2012년 12월 24일; 채택일자: 2013년 1월 10일)

초 록: 본 논문에서는 이동하는 로봇에 탑재한 대화체 음성인식기의 주위 잡음 환경에 강인한 새로운 음성 끝점 검출 기법을 제안한다. 기존의 기법은 특징 값의 갑작스러운 변화점을 찾기 위해 에지 검출 필터(edge detection filter)를 적용하여 끝점을 찾았다. 하지만 프레임 에너지의 특징은 잡음 환경에서 불안정하기 때문에 음성의 끝점을 정확하게 찾기 어렵다. 그러므로 두 번의 고속 퓨리에 변환과 통계적 모델 기반의 특징 추출 기법을 제안하여 에지 검출 필터에 적용한다. 제안한 기법이 기존의 기법보다 강인한 특징이 될 수 있음을 본 실험을 통하여 확인하였다.
핵심용어: 음성 끝점 검출, 로그 우도 비, 프레임 에너지, 고속 퓨리에 변환, 에지 검출 필터, 스펙트럼 패턴

ABSTRACT: In this paper, a new speech endpoint detection method in noisy environments for moving robot platforms is proposed. In the conventional method, the endpoint of speech is obtained by applying an edge detection filter that finds abrupt changes in the feature domain. However, since the feature of the frame energy is unstable in such noisy environments, it is difficult to accurately find the endpoint of speech. Therefore, a novel feature extraction method based on the twice-iterated fast fourier transform (TIFFT) and statistical models of speech is proposed. The proposed feature extraction method was applied to an edge detection filter for effective detection of the endpoint of speech. Representative experiments claim that there was a substantial improvement over the conventional method.
Key words: Speech segmentation, Log likelihood ratio, Frame energy, Fast fourier transform, Edge detection filter, Spectral pattern
PACS numbers: 43.72. -p

I. 서 론

HRI(Human-Robot Interface)란 로봇이 사용자 의도를 판단하고, 적합한 반응과 행동을 수행함으로써 인간과의 의사소통 및 상호협력을 가능하게 하는 인식-판단-표현 기술이다. 로봇 또는 홈오토메이션 시스템은 항상 동작하는 도중에 입력되는 신호를 분석하여 사람이 지시하는 명령을 수행한다. 이를 위해서는 마이크로폰에 연속적으로 입력되는 신호 중에 사람의 목소리가 있는지 판단을 해야 음성인식이 가능하다. 사람의 목소리가 있는지 판단하는 음성 끝점 검출은 마이크로폰 입력 신호로부터 음성의 시작점과 끝나는 점을 구분하는 과정이다. 특히, 음성인식에서의 음성 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

†**Corresponding author:** Han-Seok Ko (hsko@korea.ac.kr)
Department of Electrical Engineering, Korea University 5ka-1 Anam-dong, Seongbuk-Gu, Seoul 136-713, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

잡음이 음성인식 성능을 하락시킬 수 있는 가능성을 줄일 수 있다.[1,2] 하지만 잘못된 음성 끝점 검출은 음성인식에 필요한 음성 정보를 잃게 하여 음성인식 성능을 하락시키기도 한다. 따라서, 음성인식에서의 음성 끝점 검출은 매우 중요한 분야라고 볼 수 있다.

음성 끝점 검출 기법으로 가장 많이 알려져 있는 기법은 Rabiner와 Sambur가 제안한 프레임 에너지(frame energy)와 영교차율(zero crossing rate)을 이용한 기법으로 수학적 계산이 간단하며 음성의 기본적인 특징인 에너지와 주파수 성질을 잘 표현하는 장점이 있다.[3,4] ITU-T G.729B에서는 선 스펙트럼 주파수(line spectral frequency), 전체 주파수 대역의 에너지, 낮은 주파수 대역 에너지, 그리고 영교차율을 이용한 음성 검출 표준화 기법을 제안하였다.[5] Wilpon과 Rabiner는 미리 학습된 단어 HMM(Hidden Markov Model) 모델과 잡음 HMM 모델의 유사도를 비교하여 음성 끝점 검출을 하였다.[6] 가우시안 잡음으로부터 영향을 받지 않는 음성의 검출을 위해 HOS(Higher Order Statistics) 방법이 제안 되었고,[7] 여기에 로그 첨도(kurtosis)를 취하고 잡음 환경의 정도를 반영하여 잡음 환경의 변화에 덜 민감하도록 만들어 성능을 향상 시킨 방법이 제안되었다.[8] 음성과 잡음이 주파수 대역에서의 데이터가 다르게 분포 된다는 점을 이용하여 엔트로피로 끝점 검출을 하도록 제안하였다.[9] 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.[10] 이것을 기반으로 프레임 에너지 값으로부터 에지성분을 찾기 위해 에지 검출 필터(edge detection filter)와 상태 천이(state transition)를 적용한 기법이 개발되었다.[11] 또한 음성 신호의 기울기 변화(gradient variations)를 기반으로 하여 에지검출 필터를 적용하는 기법이 연구되었다.[12] 하지만 이러한 기법들은 신호 대 잡음비가 낮은 환경(SNR 0dB 이하)에서는 검출률이 좋지 않기 때문에, 입력 음성 신호의 주기 신호와 비주기 신호를 구분하여 하모닉스 성분을 강조하는 음성 구간 검출 기법들이 연구되었다.[13-15] 그리고 잡음 환경에서의 음성 구간 검출 성능 향상을 위해 음성의 통계 모델 기반의 특징벡터에 SVM(Support Vector Machine)을 적용한 음성 구간 검출 기법이 제안되었다.[16,17]

본 논문에서 제안하고 있는 음성 끝점 검출 기법은 HRI을 위한 음성인식 시스템의 전처리기로서의 채택을 목적으로 하고 있다. 일반적으로 HRI은 음성, 영상, 통신, 제어 등 다양한 기능을 탑재하고 있기 때문에 음성인터페이스에 많은 리소스를 할당하기 어렵다. 그러므로 본 논문에서는 계산량이 적은 에지 검출 필터와 상태 천이 모델을 검출기로 사용하였다. 하지만 HRI에 적용을 하기 위해서는 HRI에서 발생 가능한 배경 잡음에서 좋은 성능을 보이는 것 또한 중요하다. 이를 모두 고려한 기법으로서 잡음 환경에 강인한 특징 추출기법을 제안, 에지 검출 필터와 상태 천이 모델을 적용하여 음성 끝검 검출을 수행한다. 제안한 특징 추출기법은 두 번의 고속 퓨리에 변환을 이용하여, 비음성 구간에서의 잡음 신호의 스펙트럼 패턴과, 음성구간에서 잡음에 음성이 부가된 신호의 스펙트럼 패턴을 구별해준다.[18] 그리고 스펙트럼 절대값을 통계적으로 모델링하여 음성이 부재할 확률과 음성이 존재할 확률간의 로그 우도 비(Log-Likelihood Ratio)를 새로운 특징으로 사용한다.[19,20]

II장에서는 기존의 음성 끝점 검출 기법에 대하여 설명하였고, III장에서는 본 논문에서 제안한 스펙트럼 패턴과 통계적 모델 기반의 음성 끝점 검출 기법에 대하여 설명하였다. IV장에서는 실험 결과를 비교하여 성능을 평가하였다. 마지막으로 V장에서 본 논문의 결론을 맺는다.

II. 기존의 음성 끝점 검출 기법

2.1. 프레임 에너지와 영교차율

조용한 환경에서 가장 효과적으로 사용할 수 있는 방법은 프레임 에너지 기반에 영교차율을 고려한 음성 끝점 검출 기법이다.[3,4] 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 구간

에서 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 프레임에너지와 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.[11] 에지 필터는 Fig. 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.

$$
\begin{aligned}
f(x) &= e^{Ax}[K_1\sin(Ax)+K_2\cos(x) \\
&+ e^{-Ax}[K_3\sin(Ax)+K_4\cos(x)]+K_5+K_6e^s \\
h(i) &= \begin{cases} -f(i), & -W \le i \le 0 \\ f(i), & 1 \le i \le W \end{cases}.
\end{aligned} \tag{1}
$$

여기서 W는 필터길이와 관계되는 변수이며, i는 $-W$부터 W까지 정수이다. $A(A=0.41)$와 $K(K_1=1.538, K_2=1.468, K_3=-0.078, K_4=-0.036, K_5=-0.872, K_6=-0.56)$는 필터 파라미터이다. Fig. 1은 $W=7$일때의 필터 응답 그림이다.

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

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

여기서 n은 프레임 인덱스를 의미한다. $F(n)$에 상태 천이의 동작을 통하여 음성의 시작점과 끝나는 점을 찾을 수 있다. Fig. 2는 시작점과 끝나는 점을 찾기 위한 상태 천이 모델이다.

Fig. 2에서 silence는 비음성 구간을 나타내고 in speech는 음성 구간을 나타낸다. Leaving speech는 음

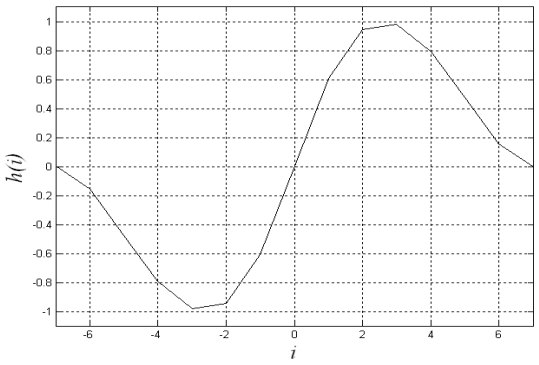

Fig. 1. Edge detection filter h ($W=7$).

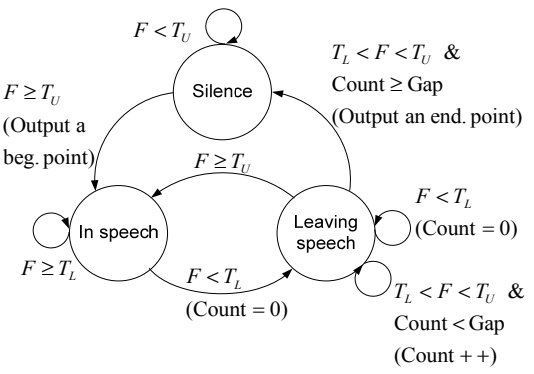

Fig. 2. State transition diagram for endpoint detection.

성 구간이지만 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 판단한다. $F(n)$이 T_U보다 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 $F(n)$이 T_U보다 커지면 다시 in speech 구간으로 돌아간다.

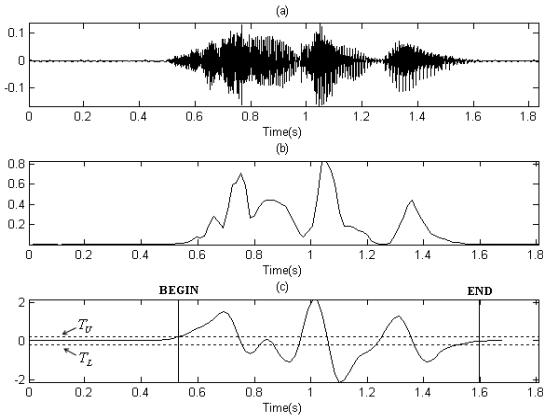

Fig. 3. Results of the frame energy and edge detection filter based endpoint detection in a clean environment, (a) clean speech signal (b) frame energy of signal(a) (c) detection output of edge filter with state transition model.

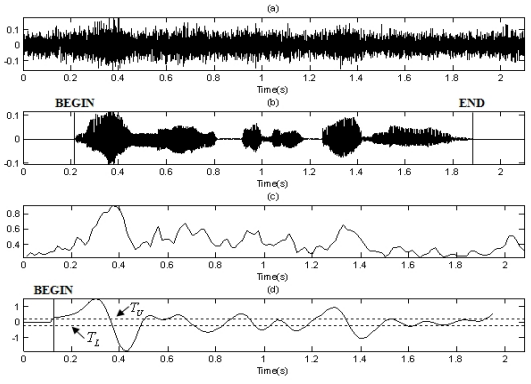

Fig. 4. Results of the frame energy and edge detection filter based endpoint detection in a car-noise environment, (a) noisy speech signal (b) clean speech signal (c) frame energy of signal(a) (d) detection output of edge filter with state transition model.

음성의 시작점과 끝나는 점을 검출하기 위해 프레임 에너지에 위 기법을 적용한 결과를 Fig. 3과 Fig. 4에 나타내었다. Fig. 3은 깨끗한 환경에서 녹음된 음성 신호의 프레임 에너지에 에지 검출 필터와 상태 천이 모델을 적용하여 끝점 검출한 결과이고, Fig. 4는 자동차 잡음 환경 음성 데이터베이스를 입력 신호로 하여 프레임 에너지에 에지 검출 필터와 상태 천이 모델을 적용하여 끝점 검출한 결과이다. 적용한 필터 길이 $W=3$ 이다.

Fig. 3에서와 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. Fig. 4의 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.

III. 스펙트럼 패턴과 통계적 모델 기반의 특징 추출 기법

II장에서는 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 에

지필터를 적용하는 것이 아니라 보다 안정적인 특징 추출 기법을 적용하여 잡음에 강인한 끝점 검출 기법을 제안한다. 제안한 특징 추출 기법은 잡음 환경 음성 신호의 스펙트럼 패턴 구별과 통계적 모델에 기반한 특징 추출기법으로 두 번의 고속 퓨리에 변환을 이용하여 비음성 구간에서의 잡음 신호의 스펙트럼 패턴과, 음성 구간에서 잡음에 음성이 부가된 신호의 스펙트럼 패턴을 구별해준다.[18] 그리고 스펙트럼 절대값을 통계적으로 모델링하여 음성이 부재할 확률과 음성이 존재할 확률간의 로그 우도 비를 새로운 특징으로 사용한다.[19,20] 이 특징 값에 에지 검출 필터를 적용하여 최종적인 끝점 검출을 수행한다.

3.1. 스펙트럼 패턴 기반의 패턴 구별 기법

시간축의 잡음 음성 신호 $x(t)$를 고속 퓨리에 변환(FFT, Fast Fourier Transform)을 이용하여 식(3)과 같이 주파수 축의 신호로 변환한다.

$$X(k^{'}) = \sum_{m=1}^{M} x(m) e^{-j\frac{2\pi}{N}k^{'}m}. \tag{3}$$

여기서 $k^{'}$는 고속 퓨리에 변환 빈 인덱스를 나타내며, M은 고속 퓨리에 변환 사이즈를 나타낸다. $k^{'}$번째 인덱스에 대하여 식(3)을 $X(k^{'})^{1st}$이라 하고, $|X(k^{'})^{1st}|$는 입력 신호의 고속 퓨리에 변환을 통한 $k^{'}$번째 주파수 인덱스에서의 절대값 크기를 나타낸다. 그다음 첫 번째 고속 퓨리에 변환을 통한 절대값 크기(magnitude)를 시간축이라고 가정하여 두 번째 고속 퓨리에 변환 $X(k)^{2nd}$ 을 식(4)과 같이 구한다.

$$X(k)^{2nd} = \sum_{k^{'}=1}^{M} |X(k^{'})^{1st}| e^{-j\frac{2\pi}{N}kk^{'}}, \tag{4}$$

여기서 k는 고속 퓨리에 변환 빈 인덱스를 나타내며, 고속 퓨리에 변환 사이즈는 256 샘플이다. 이러한 고속 퓨리에 변환 과정은 신호의 기본주파수(fundamental frequency)를 계산하여, 스펙트럼 절대값의 연속적인 정점을 가지는 하모닉 성분을 분석하는데 적합한 방법이다. Fig. 5의 (a)는 음성에 잡음이 부가된 신호와 잡음 신호를 첫 번째 고속 퓨리에 변환을 하여 절대값을 보여주고 있다. Fig. 5의 (a)에서 알 수 있듯이 음성에 잡음이 부가된 신호와 잡음 신호의 구별이 어려움을 확인할 수 있다. Fig. 5의 (b)에서 알 수 있듯이 (a)의 결과에 두 번의 고속 퓨리에 변환을 하면, 음성이 존재 하는 구간에서 주기적으로 변동을 거듭하는 하모닉 성분 때문에 음성에 잡음이 부가된 신호와 잡음 신호의 구별이 잘 됨을 확인할 수 있다.

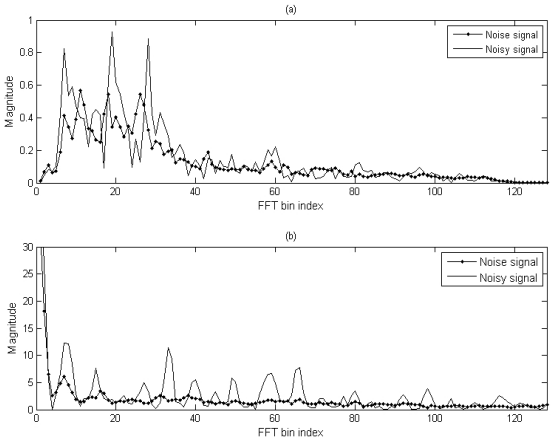

Fig. 5. Difference in pattern of 1st FFT and 2nd FFT, (a) pattern of noisy signal and pattern of noise signal after 1st FFT (b) pattern of noisy signal and pattern of noise signal after 2nd FFT.

3.2. 통계적 모델 기반의 특징추출 기법

음성 신호와 잡음 신호는 독립 가우시안 확률과정(independent Gaussian random process)을 따른다고 가정한 통계적인 모델을 사용한다. 3.1에서 두 번의 고속 퓨리에 변환을 통한 M차원의 벡터는 식(5)로 나타난다.

$$\lambda_Y(k) = \lambda_S(k) + \lambda_N(k). \tag{5}$$

$\lambda_Y(k)$ 는 $|X(k)^{2nd}|$를 의미하며 음성에 잡음이 부가된 신호의 k번째 주파수 인덱스의 절대값 크기를 나타낸다. $\lambda_S(k)$와 $\lambda_N(k)$는 각각 깨끗한 음성 신호, 잡음 신호의 절대값 크기를 의미한다. 현재의 입력 프레임이 음성 구간이라는 가설(hypothesis)

을 H_1, 비음성 구간이라는 가설을 H_0라고 하면 식 (6)과 같다.

$$H_0: speech\,absent:\ Y=N$$
$$H_1: speech\,present:\ Y=S+N, \tag{6}$$

$$P(Y|H_0)=\prod_{k=0}^{M-1}\frac{1}{\pi\lambda_Y(k)}\exp\left(-\frac{|\lambda_Y(k)|}{\lambda_N(k)}\right), \tag{7}$$

$$P(Y|H_1)=\prod_{k=0}^{M-1}\frac{1}{\pi[\lambda_Y(k)+\lambda_S(k)]}\exp\left(-\frac{|\lambda_Y(k)|}{\lambda_N(k)+\lambda_S(k)}\right), \tag{8}$$

$$\Lambda=\frac{1}{M}\log\frac{P(Y|H_1)}{P(Y|H_0)}=\frac{1}{L}\sum_{k=0}^{M-1}\left\{\frac{\lambda_Y(k)}{\lambda_N(k)}-\log\frac{\lambda_Y(k)}{\lambda_N(k)}-1\right\}. \tag{9}$$

H_0와 H_1 일 때의 관측값 Y 생성될 조건부 확률 밀도 함수(conditional probability density function)은 식(7), 식(8)과 같다.

가설 검정(hypothesis test)에 사용되는 강력한 기법중의 하나인 로그 우도 비를 이용하여 식을 간단히 전개하면 식(9)와 같이 된다.

잡음 구간에서는 입력 신호의 특징과 추정된 잡음의 특징의 패턴이 유사하여 로그 우도 비 값이 작게 나올 것이고, 음성 구간에서는 특징의 패턴 변화로 인해 로그 우도 비 값이 크게 나올 것이다. Fig. 6은 자동차 잡음 환경에서 수집한 음성 데이터를 대상으로 제안한 특징에 에지 검출 필터를 적용한 경우 실제 검출 결과를 보여주고 있다. Fig. 6에서 알 수 있듯이 거리가 비음성 구간에서는 아주 작고 음성 구간에서는 크게 나오는 것을 확인 할 수 있으며, Fig. 4와 비교하여 프레임 에너지 보다 변동이 작은 안정적인 특징임을 확인 할 수 있다. 또한 Fig. 6에서의 (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 2.2 에서 언급한 에지 검출 필터의 장점 때문에 에지 검출 필터를 적용하면 보다 안정적인 결과를 얻을 수 있다. 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 잡음에 강인하지 못한 단 점을 보이게 된다. 본 논문에는 적용하지 않았으나 적응적인 잡음 추정방식을 도입한다면 보다 좋은 끝점 검출 결과를 예상 할 수 있을 것이다.[20]

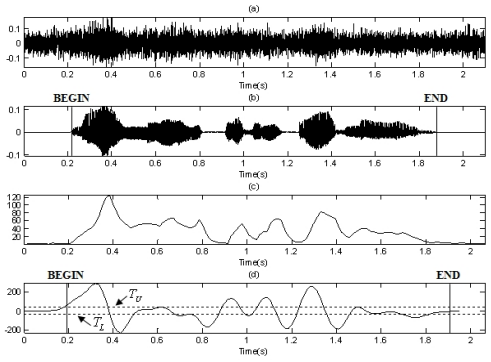

Fig. 6. Results of the proposed method in car-noise environment, (a) noisy speech signal (b) clean speech signal (c) the proposed feature (d) detection output of edge filter with state transition model.

IV. 실험 및 결과 고찰

4.1. 실험 환경

실험을 위하여 Aurora 2.0 에서 제공하는 잡음 환경 음성 데이터베이스를 이용하였다. Aurora 2.0 데이터베이스는 연속 영어 숫자음으로 이루어져 있고 신호대 잡음비가 -5dB부터 20dB까지에 대하여 실험을 하였다. 각 신호 대 잡음비에 따라 1001개의 샘플이 있으며 실험에 사용된 잡음 종류는 자동차 잡음이다. 제안한 기법이 HRI에 적용 여부를 검증하기 위해, 로봇이 구동되는 상황에서 발생한 잡음으로 추가적인 실험을 하였다. 현재 출시된 로봇청소기의 구동 잡

음을 수집하여, Aurora 2.0 의 깨끗한 환경에서 수집된 음성 데이터를 혼합하여 실험을 하였다. 로봇청소기 잡음 환경 음성 데이터베이스의 신호 대 잡음비는 -5 dB이고, 1001개의 샘플이 있다. 3.2에서 언급한바와 같이 본 논문의 잡음 신호를 정상 잡음이라고 가정을 하였기 때문에, 자동차 잡음 외에 로봇청소기 구동 잡음을 실험 데이터베이스로 사용하였다. 특징 추출을 위한 프레임 길이는 256샘플(32 ms)이며 프레임 이동 간격은 128샘플(16 ms)로 하였다. 본 논문에서 제안한 기법의 성능을 평가하기 위해 사전에 Aurora 2.0 의 깨끗한 환경에서 수집된 음성 데이터에 대하여 수작업으로 음성의 시작점과 끝나는 점을 구하였다. 수작업으로 검출된 시작점과 끝나는 점은 제안한 기법으로 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.

4.2. 실험 결과 및 토의

본 논문은 음성 신호의 통계 모델과 에지 검출 필터 및 상태 천이 모델을 기반으로 제안한 음성 끝점 검출 기법이다. 본 논문에서 제안된 기법과 기존의 다양한 기법들과의 성능을 비교하기 위해, 기존의 에지 검출 필터 및 상태 천이 모델을 이용한 기법들[10-12]과 통계 모델 및 SVM을 이용한 기법들[16,17]을 성능 평가를 위한 비교 대상으로 선정하였다.

서론에서 언급한바와 같이 본 논문에서는 HRI에 적용을 하기 위해서는 알고리즘의 계산량이 적어야하고, HRI에서 발생 가능한 배경 잡음에서 좋은 성능을 보여야한다. 기존의 에지 검출 필터와 상태 천이 모델을 이용한 기법들[10-12] 역시 계산량이 적은 기법으로, HRI에 적용이 가능할 것으로 판단되었다. 그리고 통계 모델 및 SVM을 이용한 기법들[16,17] 역시 HRI에서 발생 가능한 배경 잡음 환경에서 이미 우수한 성능을 보였기 때문에, HRI에 적용이 가능할 것으로 예상된다. 기존의 기법들과 제안한 기법을 이용하여 검출한 결과를 수작업으로 검출한 결과와 비교하여 검출 확률과 오검출 확률을 계산하였다. 또한 수작업 검출결과와의 오차 정도도 계산하였다. 성능 지표에 대한 정의는 다음과 같다.

1) 음성 구간을 정확하게 검출한 확률, P_C, 즉 총 음성데이터 개수에 대한 정확하게 검출한 음성데이터 개수의 비율이다.
2) 음성구간을 오검출한 확률, P_F, 즉 총 음성데이터 개수에 대한 오검출된 음성데이터 개수의 비율이다.

P_F는 실제로 음성 구간이지만 비음성 구간으로 판단한 경우를 의미하며 음성을 훼손하는 구간을 의미한다. 검출된 시작점이 수작업으로 검출한 시작점보다 뒤에 위치하거나, 검출된 끝나는 점이 수작업 결과보다 앞쪽에 위치하면 오검출로 인정한다. P_C는 오검출이 아닌 경우, 또는 음성 구간을 훼손하지 않은 경우를 의미하며 시작점 또는 끝나는 점이 정확히 수작업 결과와 같거나 조금 여유있게 판단한 경우를 포함한다.

Table 1은 다양한 신호 대 잡음비에 대하여 기존의

Table 1. Performance between the proposed and conventional method.

		Proposed		Q. Li *et al.* [10]		X. Li *et al.* [11]		Ghaemmaghami *et al.* [12]		Q. Jo *et al.* [16]		Q. Jo *et al.* [17]	
noise	SNR (dB)	P_C	P_F	P_C	P_F	P_C	P_F	P_C	P_F	P_C	P_F	P_C	P_F
Car	-5	88.4	11.6	71.1	28.9	73.2	26.8	76.6	23.4	86.9	13.1	82.0	18.0
	0	91.2	8.8	77.6	22.4	79.2	20.8	86.9	13.1	89.5	10.5	86.5	13.5
	10	93.1	6.9	87.1	12.9	88.1	11.9	94.2	5.8	93.0	7.0	92.7	7.3
	20	95.0	5.0	91.3	8.7	92.1	7.9	96.4	3.6	95.1	4.9	93.3	6.7
Robot cleaner	-5	88.0	12.0	70.1	29.9	72.9	27.1	75.0	25.0	86.0	14.0	81.4	18.6
Average		91.1	8.9	79.4	20.6	81.1	18.9	85.9	14.1	90.1	9.9	87.2	12.8

기법과 제안한 기법의 P_C와 P_F를 보여주고 있다. 신호 대 잡음비가 높은 환경에서는 기존의 기법과 제안한 기법이 모두 좋은 성능을 보였으나, 신호 대 잡음비가 낮은 환경에서 제안한 기법이 기존의 기법보다 높은 성능을 보였다. 제안한 기법의 평균적인 검출 확률은 기존의 기법보다 P_C는 약 1% 에서 11 % 정도 향상 되었으며, P_F 는 약 1% 에서 11 % 정도 감소되었다. 잡음 강도가 높지 않은 신호 대 잡음비가 20 dB와 10 dB인 환경에서는 제안한 기법과 통계 모델 및 SVM을 이용한 기법 [16]이 유사한 성능을 보였으나, 잡음 강도가 높은 신호 대 잡음비가 0 dB와 -5 dB인 환경에서는 제안한 기법이 우수한 성능을 보였다. 이 결과를 통하여 제안한 음성의 구간 검출 기법이 잡음의 강도가 높은 자동차 잡음 환경 뿐만 아니라, 로봇청소기 구동 잡음 환경에서도 좋은 성능을 보인다는 것을 알 수 있다. Table 1의 실험 결과를 기반으로 제안한 기법과 기존의 기법 중 P_C가 가장 높은 기법 [16]을 선택하여, P_C 와 P_F가 평균적으로 몇 프레임 내에서 발생하였는지에 대해 Table 2와 Table 3에 표현하였다.

Table 2. Frame error result of conventional method.

		P_C		P_F	
noise	SNR (dB)	Average error of beginning point (frame)	Average error of end point (frame)	Average error of beginning point (frame)	Average error of end point (frame)
Car	-5	4.0	5.6	4.1	4.8
	0	3.2	5.5	3.5	3.3
	10	2.5	4.6	2.0	2.5
	20	2.0	4.3	2.8	2.4
Robot cleaner	-5	4.2	5.5	4.2	5.2

Table 3. Frame error result of proposed method.

		P_C		P_F	
noise	SNR (dB)	Average error of beginning point (frame)	Average error of end point (frame)	Average error of beginning point (frame)	Average error of end point (frame)
Car	-5	3.0	3.2	4.1	3.4
	0	2.4	1.8	3.4	3.5
	10	2.0	1.7	3.2	3.1
	20	1.8	1.5	3.0	3.3
Robot cleaner	-5	3.1	3.3	3.8	4.0

Table 2는 기존의 기법 [16] 결과와 수작업으로 검출한 결과를 비교한 결과이다. 신호 대 잡음비가 -5 dB인 로봇청소기 구동 잡음 환경에 대하여 수작업 결과와의 평균 오차는 P_C의 경우 시작점은 4.2프레임, 끝나는 점은 5.5프레임이 나왔으며 P_F의 경우 평균 시작점은 4.2프레임, 끝나는 점은 5.2프레임으로 수작업과의 오차가 매우 큰 것을 확인 할 수 있다. 이는 잡음의 강도가 높은 환경에서 끝점 검출 결과가 불안정하다는 것을 보여준다.

Table 3은 제안한 기법의 결과와 수작업으로 검출한 결과를 비교한 것이다. 신호 대 잡음비가 -5 dB인 로봇청소기 구동 잡음 환경에 대하여 수작업 결과와의 평균 오차는 P_C의 경우 시작점은 3.1프레임, 끝나는 점은 3.3프레임이 나왔으며 P_F의 경우 평균 시작점은 3.8프레임, 끝나는 점은 4.0프레임으로 평균 프레임 수치가 작다는 것을 알 수 있다. 이 결과를 통해 제안한 기법이 수작업 결과와 보다 유사한 끝점 검출 결과를 야기시킨다고 판단할 수 있다.

Fig. 7은 로봇청소기 구동 잡음 환경에서 실험한 또 다른 결과이다. Fig. 7과 Table 1에서 알 수 있듯이 제안한 음성의 끝점 검출 기법이 로봇청소기 구동

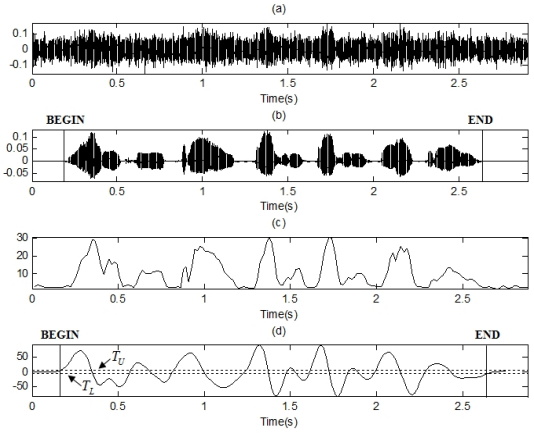

Fig. 7. Results of the proposed method in robot cleaner driving noise environment, (a) noisy speech signal (b) clean speech signal (c) the proposed feature (d) detection output of edge filter with state transition model.

Table 4. Computational cost comparison per frame.

	Proposed	Q. Li *et al*. [10]	X. Li *et al*. [11]	Ghaemmaghami *et al*. [12]	Q. Jo *et al*. [16]	Q. Jo *et al*. [17]
Computational Cost	8995	8450	8997	8850	8240	8150

잡음 환경에서도 좋은 성능을 보인다는 것을 알 수 있으며, HRI에서 음성인식 성능을 높이기 위한 요소 기술로 사용될 수 있다.

마지막으로 제안한 기법이 우수한 성능을 보이더라도 계산량이 많으면 실제로 적용이 어렵기 때문에 계산량을 고려해야 한다. 본 논문의 알고리즘에 사용된 연산을 덧셈과 곱셈은 1, 나눗셈은 5 그리고 지수연산은 10의 가중치를 설정하여, 기존의 기법과 제안한 기법의 계산량을 계산하여 Table 4에 나타내었다. Table 4와 같이 제안한 기법의 계산량이 기존의 기법보다 많지만, 검출율의 성능 향상에 비하여 계산량의 증가는 적다고 볼 수 있다. 이는 실시간 처리에는 문제가 되지 않을 것으로 예상한다.

V. 결 론

본 논문에서는 인간로봇 상호작용을 위하여 음성인식 성능을 높이기 위한 잡음 환경에 강인한 음성의 끝점 검출 기법을 제안하였다. 제안한 기법은 비음성 구간에서의 잡음과 음성 구간에서 잡음에 음성이 부가된 신호의 스펙트럼 포락선의 패턴이 달라질 것으로 예상하여 음성 신호의 두 번의 고속 퓨리에 변환과 통계적 모델 기반의 특징 추출기법을 제안, 에지검출 필터를 적용하는 기법을 제안하였다. 일반적인 프레임 에너지를 특징으로 음성 검출을 하는 것보다 강인한 특징이 될 수 있음을 본 실험을 통하여 확인하였다. 향후에는 비정상적 잡음에도 효과적인 음성 끝점 검출을 위해 적응적인 잡음추정 기법과 연동시키는 실험을 진행할 계획이다.

감사의 글

본 연구는 보건복지부 보건의료연구개발사업의 지원에 의하여 이루어진 것임.(과제고유번호 : A111189)

참 고 문 헌

1. J. Beh, R. H. Baran, and H. Ko, “Dual channel based speech enhancement using novelty filter for robust speech recognition in automobile environment,” IEEE Trans. Consumer Electronics **52**, 583-589 (2006).
2. J. Beh and H. Ko, “Spectral subtraction using spectral harmonics for robust speech recognition in car environments,” LNCS **2660**, 1109-1116 (2003).
3. L. R. Labiner and M. R. Sambur, “An algorithm for determining the endpoints for isolated utterance,” Bell Syst. Tech. J. **54**, 297-315 (1975).
4. L. R. Labiner and B. H. Juang, *Fundamentals of Speech Recognition,* (Prentice Hall, NJ, 1993).
5. ITU-T, *A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70*, (ITU-T Rec. G. 729, Annex B, 1996).
6. J. G. Wilpon and L. R. Labiner, “Application of hidden Markov models to automatic speech endpoint detection,” Comput. Speech Lang. **2**, 321-341 (1987).
7. E. Nemer, R. Goubran, and S. Mahmoud, “Robust voice activity detection using higher-order statistics in the LPC residual domain,” IEEE Trans. Speech Audio Process. **9**, 217-231 (2001).
8. K. Li, M. N. S. Swamy, and M. O. Ahmad, “An improved voice activity detection using higher order statistics,” IEEE Trans. Speech Audio Process. **13**, 965-974 (2005).
9. B. F. Wu and K. C. Wang, “Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments,” IEEE Trans. Speech Audio Process. **13**, 762-775 (2005).
10. Q. Li and A. Tsai, “A matched filter approach to endpoint detection for robust speaker verification,” in Proc. IEEE Work. AIAT (1999).
11. Q. Li, J. Zheng, A. Tsai, and Q. Zhou, “Robust endpoint detection and energy normalization for real-time speech and speaker recognition,” IEEE Trans. Speech Audio Process. **10**, 146-157 (2002).
12. H. Ghaemmaghami, R. Vogt, S. Sridharan, and M. Mason, “Speech endpoint detection using gradient based edge detection techniques,” in Proc. ICSPCS, 1-8 (2008).
13. T. Fukuda, O. Ichikawa, and M. Nishimura, “Long-term spectro-temporal and static harmonic features for voice activity detection,” IEEE J. STSP **4**, 834-844 (2010).
14. K. Ishizuka, T. Nakatani, and M. Fujimoto, “Noise robust

front-end processing with voice activity detection based on periodic to aperiodic component ratio," Speech Communication **52**, 41-60 (2010).
15. T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in Proc. Interspeech, 369-372 (2005).
16. Q. Jo, J. Chang, J. Kim, and N. Kim, "Statistical model-based voice activity detection using support vector machine," IET Signal Process. **3**, 205-210 (2009).
17. Q. Jo, Y. Park, K. Lee, and J. Jang, "A support vector machine-based voice activity detection using effective feature vectors" (in Korean) J. Telecommunications Review **18**, 362-370 (2008).
18. N. C. Maddage, K. Wan, and C. Xu, Wang, "Singing voice detection using twice-iterated composite fourier transform," in Proc. IEEE ICME, 1347-1350 (2004).
19. S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," IEEE Trans. Speech Audio Process. **11**, 498-505 (2003).
20. J. Sohn and W. Sung, "A Voice activity detector employing soft decision based noise spectrum adaptation," in Proc. IEEE ICASSP, 365-368 (1998).

저자 약력

▸ **박 진 수(Jinsoo Park)**

2008년: 경희대학교 전자정보학부 전자공학과 (공학사)
2008년~현재: 고려대학교 바이오마이크로세스템기술 협동과정 석 · 박사 통합과정 재학중
〈관심 분야〉 음성 · 음향 신호처리, 잡음 제거, 음성 끝점 검출

▸ **고 한 석 (Hanseok Ko)**

1982년 5월: 미국 카네기 멜론 대학교 전기공학 (공학사)
1986년 5월: 미국 메릴랜드 대학교 시스템 공학(공학석사)
1988년 5월: 미국 존스 홉킨스 대학교 전기공학 (공학석사)
1992년 5월: 미국 카톨릭 대학교 전기공학 (공학박사)
1995년 3월~현재:고려대학교 전기전자전파공학부 교수
〈관심 분야〉 영상 및 음성 신호처리, 패턴 인식, 데이터 융합