한국음향학회지 제31권 제3호 pp. 161~169 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.3 (2012)

음성인식 로봇을 위한 동시통화검출 기반의 강인한 음성 끝점 검출

Robust End Point Detection for Robot Speech Recognition Using Double Talk Detection

문성규 · 박진수* · 고한석*

(Sung-Kyu Moon, Jin-Soo Park*, and Hanseok Ko*)

고려대학교 영상정보처리협동과정, *고려대학교 전자전기전파공학부

(접수일자: 2012년 2월 15일; 채택일자: 2012년 2월 29일)

초 록: 본 논문에서는 반향이 큰 로봇 환경에 강인한 음성 끝점 검출 방법을 제안한다. 양방향 대화 로봇과 같이 반향 대 신호 비가 -5 dB 이하인 반향환경에서는, 반향제거기의 성능이 저하되어 사용자 음성 에너지와 비슷한 크기의 에너지를 갖는 잔여반향이 생긴다. 잡음에 강인한 기존의 음성 끝점검출 방법이라도, 사용자 음성과 비슷한 수준의 에너지를 갖는 잔여반향은 음성으로 오검출하기 때문에 정확한 음성 끝점검출이 어렵다. 반향 환경에 강인한 끝점검출을 위해, 본 논문에서는 음성/반향 구간 판별에 좋은 성능을 보이는 동시통화검출의 결과를 기존의 음성끝점검출 방법과 AND 연산하여 음성끝점검출기를 구성하였다. 제안하는 방법의 평가를 위해 반향이 큰 환경에서 고립단어 인식을 실험하였고, 다양한 실험환경에서 기존 음성 끝점검출 방법보다 평균 30 % 이상의 인식 성능 향상을 확인할 수 있었다.
핵심용어: 동시통화검출, 음성 끝점 검출, 음성인식 로봇, 반향 제거기
투고분야: 음성 음질 개선 음성처리 분야 (2.3) 잡음 감쇠 (2.3)

ABSTRACT: This paper presents a robust speech end-point detector using double talk detection in echoic conditioned speech recognition robot. The proposed method consists of combining conventional end-point detector result and double talk detector result. We have tested the proposed method in isolated word recognition system under echoic conditioned environment. As a result, the proposed algorithm shows superior performance of 30 % to the available techniques in the points of speech recognition rates.
Key words: Double talk detector, Speech end-point detector, Speech recognition robot, Acoustic echo cancellation
ASK subject classification: Speech Signal Processing (2.3)

I. 서 론

음성합성 기술과 음성인식 기술의 발달을 통해, 로봇이 대화를 통해 사용자의 의도를 판단하고 적절한 반응과 행동을 수행하는 HRI (Human-Robot Interface) 기술이 구현되고 있다. 로봇의 적절한 반응과 행동 수행에 앞서서, 로봇이 사용자의 의도를 정확히 판단하기 위해서는 잡음환경에 강인한 높은 성능의 음성 인식 기술이 요구된다. 높은 성능의 음성 인식을 위해서는 마이크 입력 신호 중 음성인식의 시작과 끝을 알릴, 음성구간의 시작점과 종료점 검출이 필수적이다. 음성신호 시작점과 종료점을 검출하는 음성 끝점검출 (End Point Detection, EPD)을 통해 비음성 구간의 잡음이 음성 인식의 성능을 하락시키는 것을 방지하고, 필요한 구간만을 입력 받으므로 음성인식에 소요되는 시간을 단축시킬 수 있다. 반대로 정확하지 못한 끝점검출은 음성구간을 비음성구간으로 간주하여 음성정보를 무시하기 때문에, 인식 성능을 저하시킬 수 있으므로 정확한 끝점검출이 요구된다. 음성 끝점검출의 성능을 하락시키는 주된

***Corresponding author:** 고한석 (hsko@korea.ac.kr)
136-701 서울특별시 성북구 안암로 145 고려대학교
전기전자전파공학부
(전화: 02-3290-3239)

요인은 잡음이다. 마이크에서 잡음을 제거하여 시스템 사용자의 음성만을 깨끗하게 입력받기 위해 spectral subtraction, wiener filter, statistical method 등 다양한 방법들이 연구되어 왔다 [1-5]. 스피커를 통한 출력과 마이크를 통한 입력을 동시에 수행하는 시스템인 Loudspeaker-Enclosure Microphone (LEM) 시스템에서는 일반적인 잡음 외에, 스피커 출력 신호가 다시 마이크로 들어가는 회귀신호 즉 반향 (Echo)의 제거가 필요하다. 다양한 선행 연구를 통해 적응필터를 기반으로 반향경로 (echo path)를 추정하여 반향을 제거하는 반향제거기 (Acoustic Echo Canceller, AEC)가 개발되었다 [6-11].

기존 연구들이 효과적으로 반향을 제거하지만, 신호 대 반향 비 (Signal to Echo Ratio, SER)가 -5 dB 이하인 경우 대부분 좋은 성능을 보이지 않는다 [6-11]. 기존 연구들이 SER -5 dB 이하인 환경에서의 반향제거에 초점을 두지 않은 이유는 대부분의 반향이 존재하는 시스템들이, SER이 -5 dB 이하일 만큼 반향이 신호에 비해 크지 않기 때문이다. 일반적으로, 반향제거가 연구되어 온 환경은 휴대전화나 핸즈프리 통화 상황이다. 이 경우, 마이크는 스피커 보다 사용자 입에 가깝게, 혹은 두 거리가 비슷하게 위치하는 경우가 대부분이기 때문에 SER이 -5 dB 이하일 경우는 많지 않다 [11]. 그러나 Text To Speech (TTS) 기능이 탑재된 음성인식 로봇의 경우, SER이 -5 dB 이하로 내려가는 경우가 위에서 언급된 통화상황에 비해 자주 발생한다. 로봇과의 자연스러운 대화 상황에서는, 로봇으로부터 화자가 어느 정도 거리를 두고 발화를 한다. 따라서 휴대전화 마이크에 입을 가져가는 통화상황에 비해, 사용자와 마이크의 거리가 스피커와 마이크 거리보다 멀다. 또, 휴대전화 스피커에 귀를 가까이 가져가는 통화상황에 비해, 로봇과의 대화 상황에서는 사용자의 거리를 고려해서 스피커의 출력도 상대적으로 크기 때문에 SER이 열악한 상황이 비교적 자주 발생한다.

일반적인 양방향 대화 로봇 시스템은 그림 1과 같다. TTS 시스템의 합성음이 로봇의 스피커를 통해 사용자에게 전해지고, 사용자 음성이 마이크를 통해 로봇에게 입력되는 시스템이다. 로봇과의 자연스러운 대화를 위해서는 로봇이 자신의 합성음을 인식해

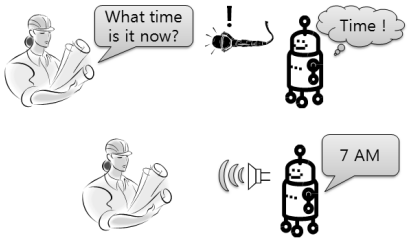

그림 1. 양방향 대화 로봇 시스템
Fig. 1. Full duplex robot speech recognition system.

버리지 않기 위해 반향 제거가 필수적이다. 그러나 음성인식 로봇의 경우 위에서 언급된 것과 같이 SER이 -5 dB 이하로 내려가는 경우가 발생하기 때문에 AEC의 성능이 하락하여 잔여반향 (Residual Echo)이 생기게 된다. 이 잔여 반향을 EPD가 음성으로 오검출 할 경우, 정확한 음성인식이 이루어질 수 없다.

본 논문에서는 음성 인식 성능 개선을 위해 AEC의 성능을 향상시켜서 잔여반향을 줄이는 방법 대신, EPD를 반향 환경에 강인하게 개선하는 간단한 방법을 제안한다. 적응필터 기반의 AEC에서는 필터의 안정적인 수렴과 발산을 방지하기 위해 동시통화검출기 (Double Talk Detector, DTD)가 사용된다. DTD는 음성과 반향 두 가지가 동시에 존재하는 구간을 검출하기 때문에 DTD라는 이름이 붙었다. 구체적으로 DTD는 음성과 반향이 동시에 존재하는 구간과 음성만 존재하는 구간을 검출한다. 동시통화구간에는 마이크 입력에 음성신호가 존재하는데, 이 구간에서 스피커 출력신호인 반향 성분만으로 AEC의 적응필터 계수를 갱신할 경우 올바른 반향 경로의 추정이 이뤄지지 않는다. 이런 잘못된 추정을 방지하기 위해 DTD는 동시통화구간을 검출하여 구간동안 적응필터 갱신을 중지시키는데 사용된다 [12-13]. DTD는 스피커 출력 신호와 마이크 입력 신호를 이용하여 음성 구간을 검출하기 때문에, 마이크 입력신호만을 갖고 음성 구간을 검출하는 EPD 보다 반향/음성구간을 구별하는데 효과적이다. 따라서 본래 DTD의 목적인 필터갱신 중지에 추가적으로, 음성구간 검출을 위해 DTD를 사용하는 것을 제안한다. 하지만 DTD가 순간잡음을 음성으로 오검출하는 경향을 보이기 때문에 DTD 결과만을 음성검출에 바로 사용하기에는 부족하다. 일반적으로 음성검출에 있어서, EPD

는 순간잡음에 강인하고 반향을 음성으로 오검출할 확률이 높으며, DTD는 반향에 강인하지만 순간잡음을 음성으로 오검출할 확률이 높다. 본 논문에서는 위와 같은 두 검출기의 특성을 고려하여, 기존 EPD 결과와 DTD 결과의 AND 연산을 통해 반향에 강인한 EPD를 구성하였다.

II 장에서는 반향환경의 음성인식 시스템과 동시통화검출 알고리즘에 대하여 설명하였고, III 장에서는 기존의 음성 구간 검출 알고리즘들에 대하여 설명하였다. IV 장에서는 본 논문에서 제안한 동시통화검출 기반의 음성끝점 검출 알고리즘에 대해 설명하였고, V 장에서 모의실험을 통해 기존의 알고리즘과 제안하는 알고리즘의 결과를 비교하여 성능을 평가하였다. 마지막으로 V 장에서 본 논문의 결론을 맺는다.

II. 반향환경의 음성인식 시스템 및 동시통화검출기

2.1 반향환경의 음성인식 시스템

일반적인 반향제거 및 음성인식 통합 시스템을 그림 2에 나타내었다. AEC 시스템은 반향 제거를 위해 스피커부터 마이크까지의 반향경로 (echo path) h 를 추정하여, 추정된 반향경로 $\hat{h}$ 를 적응적으로 갱신한다. 스피커 출력 전 신호를 추정된 반향경로에 통과시켜서 추정 반향을 구하고, 추정 반향을 마이크 입력 신호에서 차감하여 반향을 제거한다. 일반적으로 적응 필터 차수가 실제 반향경로의 room response 보다 짧아서 반향경로 추정이 완벽하게 이루어지지 못한다. 이로 인해 반향이 완전히 제거되지 못하고 남게 되어, 잔여반향 (Residual Echo)이 생긴다 [11].

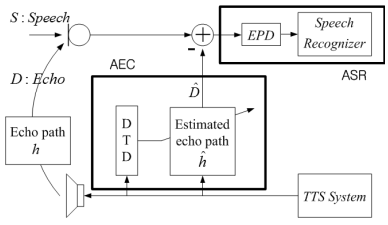

그림 2. 반향제거 및 음성인식 통합시스템
Fig. 2. Integrated system for Acoustic Echo Cancellation (AEC) and Automatic Speech Recognition (ASR).

AEC 시스템에서 동시통화 (Double Talk, DT)는 마이크로폰 입력에 사용자의 음성과 반향이 입력되는 경우, 혹은 반향 없이 음성만 존재하는 경우를 말한다. DT 구간의 경우, 적응필터의 desired signal (마이크 입력신호)에 사용자 음성이 존재한다. 반향 관련 성분만을 갖고 있는 reference signal (스피커 출력 전 신호)로 반향경로를 추정 시, 수렴 속도에 문제가 생기거나 필터 결과가 발산하게 된다. 따라서 DT 구간을 검출하여 DT 구간 동안 적응필터의 갱신을 중지시켜야 한다. 대부분의 AEC가 적응필터를 기반으로 구성되기 때문에 DTD는 필수적으로 사용 된다 [6-11].

2.2 동시통화검출 알고리즘

동시통화검출 알고리즘은 검출 방식에 따라, 스피커 출력신호와 마이크 입력신호의 에너지를 기반으로 한 알고리즘과 두 신호의 상호상관 기반으로 한 알고리즘으로 크게 나뉘어 각각 널리 쓰이고 있다 [12-13]. 두 종류의 알고리즘 중 대표적인 알고리즘은 다음과 같다. 여기서 $x(t)$는 스피커 출력신호이고, $y(t)$ 는 마이크 입력 신호이다.

2.2.1 Geigel 알고리즘 [12]

기본적으로 마이크 입력에 반향 되어 돌아오는 신호보다 사용자의 음성이 더해진 신호의 에너지가 크다는 성질을 기반으로 한다. k 는 실험적으로 설정된 상수 값이다.

$$D(n) = \frac{|y(n)|}{\max[x(n),...,x(n-k+1)]} \tag{1}$$

사용자의 음성이 더해지면 $|y(n)|$ 값은 커지게 되고 $D(n)$ 의 값이 커지게 된다. $D(n)$ 가 일정 문턱 값을 넘으면 동시통화 구간으로 판단한다.

2.2.2 상호상관 알고리즘 [13]

$$D(n) = \max_i [\frac{E[x(n-i)y(n)]}{E[x^2(n)]}] \tag{2}$$

스피커 출력신호 $x(n)$와 사용자 음성이 입력되지 않은 반향신호 $y(n)$는 서로 상관성이 높은 신호이다. 만약 $y(n)$에 사용자 음성이 들어간다면 그 부분에서 상관성이 갑자기 떨어지게 되는 성질을 이용한 것이다. 위 식에서 분자 부분이 상호상관 부분에 해당되며 분모부분은 정규화를 위한 성분이다. i는 delay에 해당되는 변수로 i를 변화시키면서 최대가 되는 값을 찾게 된다. 동시통화가 발생하였을 경우 $D(n)$의 값은 작아지게 된다. $D(n)$이 일정 문턱 값 이하로 떨어지면 동시통화 구간으로 판단한다.

III. 기존의 음성 구간 검출 알고리즘

AEC 시스템을 통해 반향이 제거된 신호는, 자동음성인식 (Automatic System Recognition, ASR) 시스템으로 입력된다. 본격적인 음성 인식 과정을 수행하기에 앞서, 음성의 시작과 끝을 알려주는 EPD 과정이 필요 하다 [14]. EPD를 통해서 ASR 시스템 입력 신호 중에 음성이 존재하는 구간만 선별하여 음성 인식을 수행하게 된다. Rabiner와 Sambur가 제안한 에너지와 영교차율을 이용한 EPD 알고리즘이 수식적으로 간단하여 널리 쓰이고 있다[15]. 하지만 위의 알고리즘은 순간잡음도 음성으로 검출하는 문제점을 보이는데 이를 보안하고자 순간잡음에 강인한 엔트로피 기반의 방법이 널리 쓰이고 있다[16].

3.1. 프레임 에너지와 영교차율 기반의 음성 구간 검출 [15]

조용한 환경에서 가장 효과적으로 사용할 수 있는 방법은 프레임 에너지 기반에 영교차율을 고려한 음성 구간 검출 기법이다. 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 프레임에너지와 영교차율만 이용한 음성 구간 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.

3.2 스펙트럼 엔트로피 기반의 음성 구간 검출 [16]

음성과 잡음은 주파수 대역에서 데이터 분포 형태가 다르다. B. F. Wu 는 이를 바탕으로 그 엔트로피를 계산하여 음성과 잡음으로부터 음성을 구분해 내는 방법을 제안하였다. 이 방법에서는 음성과 잡음의 엔트로피가 다르게 나타나도록, 주파수 대역을 일정한 간격으로 마스킹 하여 주파수 대역별로 분리 한 후, 엔트로피를 구하였다. 음악 소리와 같이 사람의 음성과 유사하게 넓은 대역에 에너지가 분포하는 경우 음성 검출이 어렵다는 단점이 있지만, 잡음과 같이 특정대역에 에너지가 집중된 경우 그 에너지 크기에 상관없이 강인하게 음성 구간을 검출해 낸다.

IV. 제안하는 동시통화검출을 이용한 음성 끝점 검출

앞에서 언급된 잡음에 강인한 선행 연구에서도, 그림 3의 예시와 같이 잔여반향의 크기가 사용자의 음성의 크기와 비슷하거나 더 큰 수준의 경우 안정적인 EPD 수행이 어렵게 된다. 에너지에서 음성과 잔여반향이 차이가 없고, 반향이 음성신호이므로 신호의 특성이 사용자의 음성과 차이가 없다. 따라서 앞서 언급된 엔트로피나 그 밖의 하모닉스 기반의 강인한 알고리즘 [16-17]을 통해서도 반향에 강인한 음성끝점검출이 어렵게 된다.

앞서 언급되었듯이 DTD는 마이크 입력신호만으로 음성을 검출하는 EPD와는 달리, 반향의 original source라고 할 수 있는 스피커 출력신호와 마이크 입력신호의 비교를 통해서 음성 구간을 찾는다. 따라서 앞선 방법들보다 반향에 강인하게 음성 구간을 검출하는 결과를 기대할 수 있다. DTD를 음성검출에 사용하기 전에, 순간적인 DTD 오검출 값의 영향을 줄이기 위해서 DTD 결과에 식 (3)과 같은 smoothing 연산을 수행한다.

$$\lambda\, DTD(n) + (1-\lambda)\, DTD(n-1) = DTD_{smooth}(n) \tag{3}$$

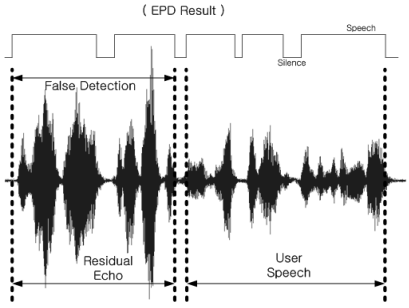

그림 3. 잔여반향으로 인한 잘못된 EPD 결과
Fig. 3. False end point detection caused by residual echo.

$$\begin{aligned} &\text{if} \quad DTD_{smooth}(n) > T_{up}, \\ &\quad DTD_{result}(n-M), DTD_{result}(n-(M-1)) \\ &\quad \ldots DTD_{result}(n) = 1 \\ &else \quad DTD_{result}(n) = 0 \end{aligned} \tag{4}$$

식 (3)에서 $DTD(n)$ 은 n 번째 프레임 (DTD 알고리즘에 따라 n 번째 샘플)의 DTD 결과이고, λ 는 0에서 1사이의 값을 갖는 mixing paramter 이다. 시작점에 margin을 주기 위해 식 (4)에서 표현 되었듯이, 식 (3)을 통해 smoothing이 된 DTD 결과인 $DTD_{smooth}(n)$가 문턱값 T_{up}를 넘으면, 적절한 프레임 margin 값 M을 주어 해당 프레임에서 M 번째 이전 프레임부터 DTD_{result}에 1 값을 준다. 그 외의 프레임에는 DTD_{result}에 0 값을 준다. 본 논문에서 M, λ 값은 반복 실험을 통해 최적 값을 찾아 사용하였다.

잔여반향과 음성 그리고 순간잡음이 동시에 발생하지 않고 개별적으로 존재하는 AEC 출력 신호에 대한 DTD_{result} 결과와 EPD 결과를 그림 4에 나타내었다. 추가적으로 로봇과 사용자의 발화가 겹치는, 즉 잔여반향과 음성이 같은 시간에 입력되는 경우에 대한 실험 결과를 그림 5에 나타내었다. 그림 4, 5의 실험을 위해 B. F. Wu의 EPD [16]를 사용하였고, H. Ye

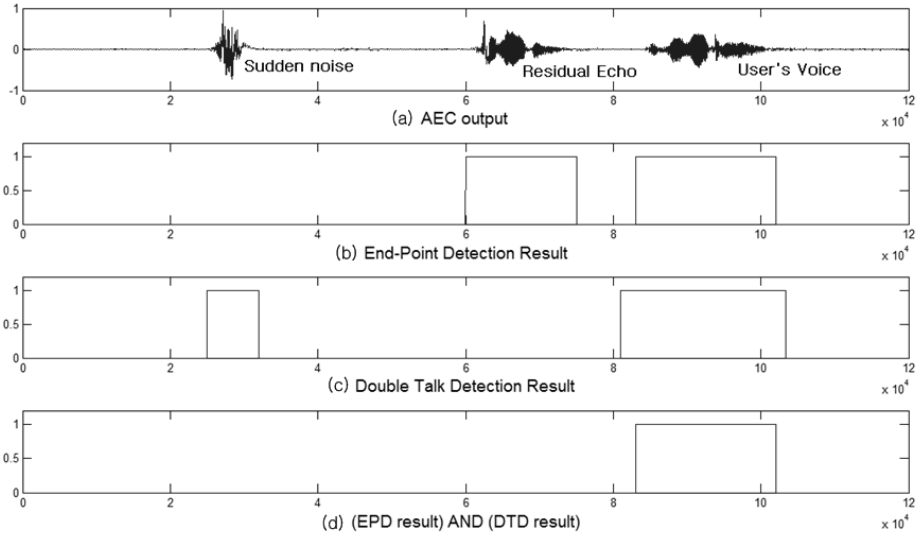

그림 4. (a) AEC 출력신호, (b) 음성끝점검출 결과 (c) 동시통화검출 결과 (d) 음성끝점검출 결과와 동시통화검출 결과의 AND 연산 결과
Fig. 4. (a) AEC output (b) End-point detection result (c) Double tak detection result (d) AND operation result of EPD result and DTD result.

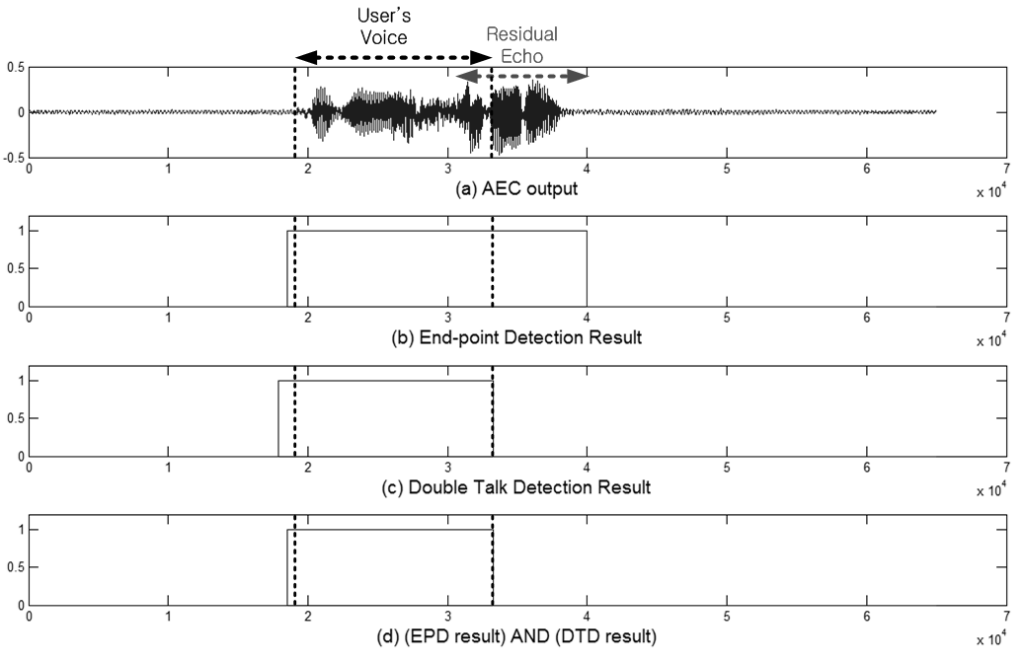

그림 5. (a) AEC 출력신호, (b) 음성끝점검출 결과 (c) 동시통화검출 결과 (d) 음성끝점검출 결과와 동시통화검출 결과의 AND 연산 결과

Fig. 5. (a) AEC output (b) End-point detection result (c) Double tak detection result (d) AND operation result of EPD result and DTD result.

의 DTD [13]를 사용하였다.

그림 4 (b)의 끝점검출 결과에서 보이듯이, EPD는 순간잡음에 대해서는 강인하지만 잔여반향을 사용자 음성처럼 검출하였다. DTD는 스피커출력 신호와 마이크 입력신호를 비교하기 때문에 그림 4 (c)의 결과처럼 잔여반향은 동시통화구간으로 검출하지 않고 순간잡음과 사용자 음성을 동시통화구간으로 검출하였다. 그림 5 (b,c)의 결과를 통해 사용자 음성과 잔여반향이 같은 시간에 입력되는 경우에도, EPD가 사용자 음성과 잔여반향을 모두 검출한 것에 비해 DTD는 사용자 음성 구간만 검출한 것을 확인 할 수 있다.

사용자의 음성구간을 검출하는데 있어서 그림 4-5 결과가 보여주듯이, EPD는 순간잡음에 강인하지만 반향을 사용자 음성으로 오검출 하는 경우가 많고, DTD는 반향에 강인하지만 순간잡음을 오검출 하는 경우가 많다. 따라서 본 논문에서는 검출기의 장단점을 보완하기 위해, DTD 결과를 EPD 결과와 AND 연산하여 음성 끝점 검출에 사용한다. 식 (4)를 통해 구한 smoothing된 DTD인 $DTD_{result}(n)$과 $EPD(n)$ 결과를 식 (5)처럼 AND 연산하여, 최종적으로 제안하는 음성끝점검출 결과, $EPD_{DTD}(n)$ 를 얻을 수 있다. 그림 4 (d)와 그림 5 (d)를 통해 제안하는 음성끝점검출기의 결과가 반향과 순간잡음에 강인하다는 것을 확인할 수 있다. 제안하는 시스템은 그림 6과 같다.

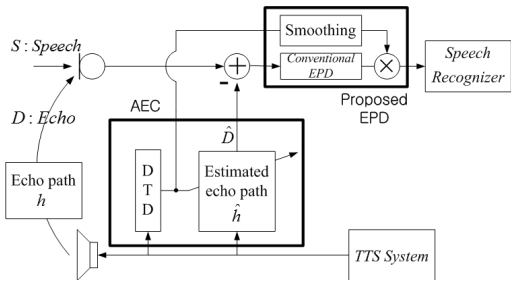

그림 6. 제안하는 반향제거 및 음성인식 통합시스템

Fig. 6. Proposed integrated system for Acoustic Echo Cancellation (AEC) and Automatic Speech Recognition (ASR).

$$DTD_{result}(n) \otimes EPD(n) = EPD_{DTD}(n) \quad (5)$$

V. 모의실험 및 결과고찰

본 논문에서 제안한 DTD에 기반 한 EPD의 성능을 평가하기 위해 그림 7과 같이 두 가지 실험 상황을 가정하였다. 첫 번째는 로봇과 사용자가 한번 씩 대화를 주고받는 경우와 같이 음성과 반향 구간이 겹치지 않는 일반적인 상황, 두 번째는 로봇의 발화 중에 사용자가 발화 하는 경우 또는 로봇의 스피커를 통해 음악이나 TV소리가 출력되는 중에 사용자가 발화 하는 경우와 같이 음성과 반향 구간이 겹친 상황이다. 두 실험 상황에 대해 고립 단어 인식을 수행하여, 제안한 EPD가 적절하게 음성 구간을 검출하였는지를 단어 인식률로 평가하였다. HTK 3.2 기반의 인식기를 사용하였고, ETRI 헤드셋 한국어 DB를 통해 72명의 DB를 학습용으로 (사람당 452 단어 발화), 그 외의 32명의 화자의 DB를 평가용으로 사용하였으며, 해당 DB의 임의 단어발화를 반향으로 사용하였다 [18-20]. 정상잡음 환경을 가정하여 반향 외에 배경잡음이 효과적으로 제거된 상황을 설정하고, 마이크와 스피커의 거리는 10 cm, 마이크와 사용자의 거리는 1 m를 가정하였으며, SER이 -5 dB, -10 dB인 상황으로 실험 데이터를 합성하였다. EPD는 순간잡음에 강인한 B. F. Wu의 EPD 방법 [16]을 사용하였으며, DTD는 에너지 기반의 대표적 알고리즘인 Geigel DTD [12]와 상호상관도 기반의 대표적 방법인 H. Ye의 DTD [13], 두 종류를 사용하여 실험해 보았다.

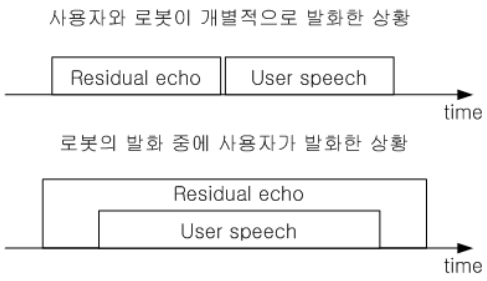

그림 7. 실험 상황 예시
Fig. 7. Example of experimental environment.

기존 EPD 결과와 제안하는 EPD 결과를 그림 8에 나타내었다. 기존의 EPD는 입력신호의 묵음 (silence) 구간을 효과적으로 제거 했지만, 잔여반향 구간을 음성구간으로 오검출하였다. 제안하는 EPD는 잔여반향 구간과 묵음 구간을 음성구간으로 오검출하지 않고 음성 구간만을 검출하였다. 두 가지 실험 상황에 대한 고립 단어 인식 실험 결과는 표 1과 같다. 사용자와 로봇이 개별적으로 발화한 경우, 기존의 방

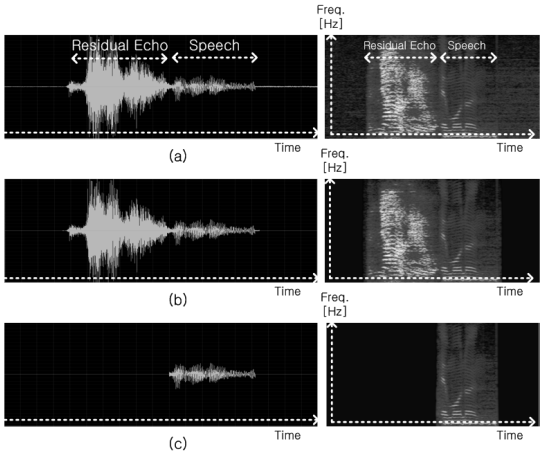

그림 8. 음성 구간 검출 결과 예시
(a) 음성 구간 검출 전 신호의 시간축 파형 (左)과 스펙트로그램 (右), (b) 기존 음성 끝점검출 결과의 신호의 시간축 파형 (左)과 스펙트로그램 (右), (c) 제안하는 음성 끝점검출 결과의 신호의 시간축 파형 (左)과 스펙트로그램 (右)
Fig. 8. Experimental results of speech end point detection.

표 1. 고립어 단어 인식 실험 결과
Table 1. Experimental result of isolated word recognition.

	SER [dB]	Ground truth [%]	Conventional Method	Proposed Method	
			EPD only [%]	EPD with Geigel DTD [%]	EPD with J. Benesty [%]
사용자, 로봇 개별 발화	-5	87.93	15.96	73.16	73.28
	-10	87.93	14.34	64.31	64.27
사용자, 로봇 동시 발화	-5	23.51	1.94	17.50	17.81
	-10	19.85	2.39	14.62	15.55
Average		57.24	8.66	42.40	42.73

법으로는 정확한 음성 시작점을 찾지 못하여 잔여 반향을 사용자의 음성처럼 인식 해버리기 때문에 인식률이 저하 된다. 제안하는 EPD 방법은 기존 EPD에 비해 정확한 음성 시작점을 검출해 인식률이 평균 54 % 정도 개선되었다. 로봇의 발화 중에 사용자가 발화한 상황 같은 경우는 사전 정보를 통해 이상적으로 사용자 음성 구간을 검출한 ground truth의 결과도 높지 않은 상황이다. 기존의 EPD만을 사용하였을 때는 인식결과가 ground truth에 비해 현저히 낮았지만, DTD를 기반으로 EPD의 경우 어느 정도 ground truth의 결과에 근접하였음을 확인 할 수 있다.

본 논문에서 제안하는 방법이 동시통화 구간에서 반향을 제거하지는 않기 때문에 로봇의 발화 중에 사용자가 발화한 상황에서는 큰 성능 향상을 보이지 않는다. 하지만 기존 EPD에 비해 일관되게 성능향상을 보였고, 사용자와 로봇이 개별적으로 발화한 상황에서는 제안하는 방법이 기존 방법에 비해 50 % 이상의 인식성능 향상을 보이기 때문에 음성인식 로봇 시스템에서 기존 방법에 비해 효과적임을 확인할 수 있다. 또한 에너지 기반의 DTD와 상호상관을 기반으로한 DTD를 사용하였을 때 모두 비슷한 수준의 성능향상을 보이므로 제안하는 방법에 다양한 DTD 알고리즘이 적용 가능함을 확인 하였다.

VI. 결 론

본 논문에서는 반향이 큰 열악한 환경 (SER -5dB 이하)의 양방향 대화 로봇의 EPD 성능 개선에 관해 연구하였다. 기존의 AEC 적응 필터의 발산 방지를 위한 DTD의 결과가 음성의 존재 여부를 판단한다는 것에 착안하여, DTD의 결과를 음성 끝점검출에 사용하였다. 반향에 강인한 DTD와 순간잡음에 강인한 EPD의 특성을 보완하고자, 각 검출기의 결과를 AND 연산하여 반향에 강인한 음성끝점검출기를 구성하였다. 대부분의 AEC 시스템이 기본적으로 DTD를 사용하기 때문에 별도의 알고리즘 추가 없이 제안하는 방법을 사용 할 수 있다. 고립 단어 인식 모의실험에서, 사용자와 로봇이 겹치지 않게 발화할 때 평균 54 % 정도의 인식성능 개선을 나타내었고, 로봇과 사용자의 발화가 겹치더라도 기존 방법에 비해 14 % 정도의 인식성능이 향상되었다. 실험을 통해서 널리 쓰이는 에너지 기반과 상호상관 기반의 DTD 모두 제안하는 방법에 적용 가능함을 확인하였다. 향후에는 본 방법에 최적화된 EPD와 DTD를 연구하여 음성 끝점검출 성능을 개선하고, DTD의 성능 개선을 위해 EPD를 사용하는 연구도 진행할 계획이다.

감사의 글

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

참고문헌

1. R. Martin, “Spectral Subtraction Based on Minimum Statistics”, *Proc. EUSIPCO 94*, pp. 1182-1185, Apr. 1994.
2. S. Boll, “Suppression of Acoustic Noise in Speech using Spectral Subtraction”, *IEEE Transactions on Speech and Audio Processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.
3. R. McAulay, M. Malpass “Speech Enhancement using a Soft-decision Noise Suppression Filter”, *IEEE Transactions on Speech and Audio Processing*, vol. 28, no. 2, pp. 137-145, Apr. 1980.
4. J. Chen, J. Benesty, H. Yiteng, S. Dolco “New Insights into the Noise Reduction Wiener filter”, *IEEE Transactions on Audio, Speech and Audio Processing*, vol. 14, no. 4, pp. 1218-1234, July. 2006.
5. Y. Ephraim, “Statistical-model-based Speech Enhancement Systems”, *Proc. IEEE*, vol. 80, no. 10, pp. 1526-1555, Oct. 1992.
6. C. Paleologu, S. Ciochina, J. Benesty, “An Efficient Proportionate Affine Projection Algorithm for Echo Cancellation ,” *IEEE Signal processing letter*, vol. 17, issue 2, 165-168, 2010.
7. A. Mader, H. Puder, G. U. Schmidt, “Step-size control for acoustic echo cancellation filters -an overview”, *Signal Processing*, vol. 80, issue 9, pp. 1697-1719, Sept. 2000.
8. C. Paleologu, S. Ciochina, J. Benesty, “Variable step-size NLMS algorithm for under-modeling acoustic echo cancellation”, *IEEE Signal Processing Letters*, vol. 15, pp.5-8, Sept. 2008.
9. T. V. Waterschoot, R. Geert, V. Piet, M. Marc “Double-Talk-Robust Prediction Error Identification Algorithms for Acoustic Echo Cancellation”, *IEEE Transactions on Signal Processing*, vol. 55, issue 3, pp. 846-858,

Mar. 2007.
10. A. Mader, H. Puder, G. U. Schmidt, “Step-size control for acoustic echo cancellation filters -an overview”, *Signal Processing*, vol. 80, issue 9, pp. 1697-1719, Sept. 2000.
11. S. Gustafsson, R. Martin, P. Jax, P. Vary “A psychoacoustic approach to combined acoustic echo cancellation and noise reduction”, *IEEE Transactions on Audio, Speech and Audio Processing*, vol. 10, issue 5, pp.245-256, Jul. 2002.
12. D. Duttweiler, “A twelve-channel digital echo canceler,” *IEEE trans. Commun*., vol.26, no. 5, pp. 647-653, May. 1978.
13. Hua Ye, Bo-Xiu Wu, “A new double-talk detection algorithm based on the orthogonality theorem”, *IEEE trans. communications*, vol. 39, issue 11, 1542-1545, Nov. 1991.
14. 박진수, 이윤재, 이인호, 고한석 “ 스펙트럼 패턴 기반의 잡음 환경에 강인한 음성의 끝점 검출 기법”, *말소리와 음성과학*, 1권, 4호, 2009.
15. L. R Labiner, M. R. Sambur, “Voiced-unvoiced-silence detection using the Itakura LPC distance measure”, *Proc. ICASSP,* pp. 323-326, 1977.
16. B. F. Wu, K. C Wang, “Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse ”, *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 762-775, Sept. 2005.
17. T. Fukuda, O. Ichikawa, M. Nishimura, “Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection”, *IEEE trans. selected topics in signal processing*, vol. 4, issue 5, pp. 834-844, Oct. 2010.
18. Hidden Markov model Toolkit 3.2, *http://htk.eng.cam.ac.uk*
19. ETRI Headset Korean DB, http://voice.etri.re.kr
20. L. Yeonja, L. Youngjik, “Implementation of the POW (phonetically optimized words) algorithm for speech database”, *Proc. ICASSP'95,* pp.89-92, May. 1995.

저자 약력

▸ **문 성 규 (Sung-Kyu Moon)**

2011년: 단국대학교 전자컴퓨터공학부 (공학사)
2011년~현재: 고려대학교 영상정보처리협동과정 석·박사 통합과정 재학중
〈관심 분야〉 음향 신호처리

▸ **박 진 수 (Jin-Soo Park)**

2008년: 경희대학교 전자정보학부 전자공학과 (공학사)
2008년 ~ 현재: 고려대학교 바이오마이크로세스템기술 협동과정 석·박사 통합과정 재학중
〈관심분야〉 음성·음향 신호처리, 잡음 제거, 음성 구간 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
〈관심분야〉 영상 및 음성 신호처리, 패턴 인식, 데이터 융합