말소리와 음성과학 제2권 제4호 (2010)
pp. 137~143

상태변수 기반의 실시간 음성검출 알고리즘의 최적화

Optimization of State-Based Real-Time Speech Endpoint Detection Algorithm

김 수 환[1] · 이 영 재 · 김 영 일 · 정 상 배[2]
Kim, Suhwan · Lee, Youngjae · Kim, Young-Il, · Jeong, Sangbae

ABSTRACT

In this paper, a speech endpoint detection algorithm is proposed. The proposed algorithm is a kind of state transition-based ones for speech detection. To reject short-duration acoustic pulses which can be considered noises, it utilizes duration information of all detected pulses. For the optimization of parameters related with pulse lengths and energy threshold to detect speech intervals, an exhaustive search scheme is adopted while speech recognition rates are used as its performance index. Experimental results show that the proposed algorithm outperforms the baseline state-based endpoint detection algorithm. At 5 dB input SNR for the beamforming input, the word recognition accuracies of its outputs were 78.5% for human voice noises and 81.1% for music noises.

Keywords: endpoint detection, voice activity detection, speech recognition, speaker recognition

1. 서론

여러 가지 기계-인간 인터페이스 기법 중에서 음성에 의한 방법은 사용자에게 가장 많은 편의성을 제공할 수 있는 잠재력을 가지고 있음에 틀림없다[1]. 이러한 이유로 음성인식은 기계-인간 인터페이스 기법에서 매우 중요한 역할을 할 것으로 기대되고 있다. 이러한 음성인식은 지능형 로봇, 차량 네비게이션 등의 여러 가지 기기들에 적용되고 있다. 또한, 요즘 떠오르는 스마트폰의 출현으로 음성인식의 활용범위가 더욱 넓어져 사용자들에게 더욱 편리함을 제공하고 있다. 음성인식의 과정은 입력음성으로부터 음성구간만을 정확히 검출하는 부분, 특징추출을 수행하는 부분, 특징과 참조 패턴을 비교하는 부분 등으로 나눌 수 있다. 여기서, 음성검출의 목적은 배경잡음 구간으로부터 정확히 음성구간만을 추출해내는 데에 있다. 추정된 음성구간에서 너무 많은 잡음을 포함하고 있으면 정확한 음성인식을 기대하기 어려우며 패턴추출의 횟수 및 참조 패턴과의 비교 횟수가 많아져서 음성인식 응답 속도의 저하를 일으킨다. 또한, 잘못된 음성검출에 의해서 추정된 음성구간의 앞부분 혹은 뒷부분이 잘린다면 화자 발성의 음소 정보가 사라지게 되므로 음성인식률의 저하를 발생시킬 것이다. 기존 음성검출 기법에는 엔트로피를 기반으로 한 기법과 웨이브렛 계수 분산 및 부대역 진폭 분산을 기반으로 하는 기법 등이 있다[2], [3]. 가장 편리한 음성검출 기법은 평균배경 잡음의 에너지에 임의의 배수를 취하여 그 이상일 경우에 음성으로 판정하는 것이다. 잡음 조건이 열악할 경우에는 에너지 궤적만으로는 정확한 성능을 발휘할 수 없기 때문에 주파수 영역에서의 켑스트럼, 필터뱅크 에너지가 사용된다[4]. 대신 연산량이 많아져서 서버용 대용량 음성인식 시스템이 아닐 경우에는 구현하기가 힘들다.

본 논문에서 제안하고 있는 음성검출 알고리즘은 지능형 로봇을 위한 명령어 인식 시스템의 전처리기로의 채택을 목적으로 하고 있다. 일반적으로 지능형 로봇은 음성, 영상, 통신, 제어 등 다양한 기능을 탑재하고 있기 때문에 음성인터페이스에 많은 리소스를 할당하기 어려우므로 본 논문에서는 계산량이 적은 에너지 및 음성의 상태변수 모델링을 활용한 음성검출 알고리즘에 대해서 논하기로 한다. 지능형 로봇 환경에서의 음성인식은 주로 1 m 이상의 원거리 음원 취득이 목적이므로 다채

1) 경상대학교 전자공학과 edps2166@gnu.ac.kr
2) 경상대학교 전자공학과(공학연구원) jeongsb@gnu.ac.kr, 교신저자

접수일자: 2010년 10월 29일
수정일자: 2010년 12월 15일
게재결정: 2010년 12월 16일

널 마이크로폰을 활용하여 녹음을 하게 되며 사람 목소리, 음악 소리 등 비정상성 잡음이 강하게 수신되는 잡음환경을 고려한다. 따라서, 본 논문에서는 효과적인 음성검출을 위해서 참고문헌 [5]에서 Qi Li가 제안한 상태변수기반의 음성검출 알고리즘을 기반으로 하여 음성의 시작점 및 끝점 부근에 존재하는 단구간의 잡음 펄스를 효과적으로 제거할 수 있는 새로운 음성검출 알고리즘을 제안한다. 제안된 알고리즘의 성능은 지능형 로봇을 위한 원거리 음성인식률로 측정된다. 이를 위해서 다채널 기반의 빔포밍 알고리즘을 거친 후의 음성입력에 대해서 음성검출이 이루어진다. 본 연구에서 제안된 방식에는 다수의 파라미터가 사용되는데 이것의 최적화는 음성인식률 관점에서 전수조사 방식으로 수행된다.

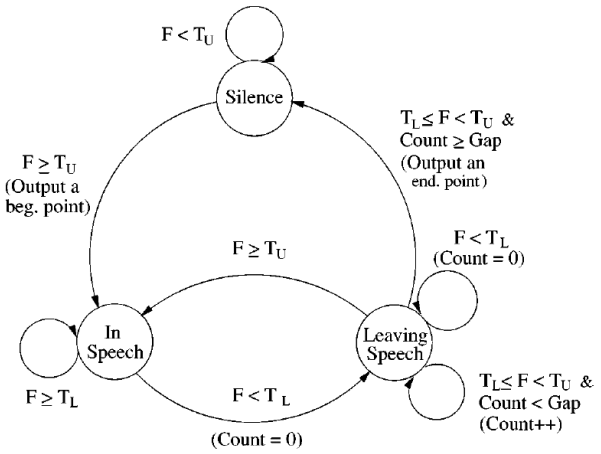

그림 1. Qi Li의 상태변수 기반 음성검출 알고리즘[5]
Figure 1. Qi Li's state variable-based endpoint detection algorithm[5]

본 논문의 구성은 다음과 같다. 제 2장에서 상태변수 기반 음성검출을 위한 베이스라인 알고리즘을 소개하고, 제 3장에서 제안된 방식에 대해서 소개한다. 제 4장에서는 음성검출 알고리즘의 실험 및 결과에 대해서 논하고 제 5장에서는 본 연구의 결론을 맺는다.

2. 베이스라인 음성검출 알고리즘

상태변수 기반의 알고리즘은 단구간 에너지의 크기, 지속시간, 휴지시간 등으로 요약될 수 있는 음성신호의 특성을 효과적으로 모델링할 수 있는 장점이 있다. 여러 가지 상태변수 기반 검출 방식 중에서 본 연구에서는 Qi Li가 제안한 상태변수 기반 방식을 베이스라인 알고리즘으로 선정하였다[5].

<그림 1>에 본 연구의 베이스라인 알고리즘을 나타내었다. <그림 1>에서 *F*는 단구간 로그 에너지를 사인형태의 함수로 필터링한 것을 의미한다[5]. 베이스라인 알고리즘에서 사용하고 있는 사인형태의 함수를 <그림 2>에 나타내었다. 참조 논문 [5]에서는 27차의 사인함수 필터를 사용하였다. T_L 및 T_U는 단구간 에너지가 상승 및 하강을 검출하기 위한 파라미터이다. Gap는 음성종료와 short-pause를 구분하기 위한 파라미터이다. 즉, 묵음 구간이 Gap 이하의 횟수만큼 검출되면 short-pause로 간주되며 그 이상이면 음성종료로 간주한다. <그림 1>로 주어진 음성검출 알고리즘에 임의의 음성파형을 입력하였을 때의 알고리즘 동작 형태를 <그림 3>에 나타내었다. <그림 3>에서 에너지 궤적의 변화량이 음성구간의 상승부 혹은 하강부에 비해서 크지 않은 묵음 또는 배경잡음 구간에서는 에너지 궤적을 사인형태의 함수에 필터링을 시켰을 때 그 출력 F 가 상대적으로 작은 값의 범위인 T_L 및 T_U 사이에 존재함을 알 수 있다. 반면, 음성구간에서는 <그림 3(b)>의 F 값이 <그림 3(c)>에서 에너지 궤적의 상승 변곡점에서 국부 최대값을, 하강 변곡점에서 국부 최소값을 갖게 되고 그 값의 크기가 배경잡음 구간에 비해서 매우 큼을 알 수 있다. 물론, 음성구간에서 에너지가 유지 되는 평탄한 구간에서

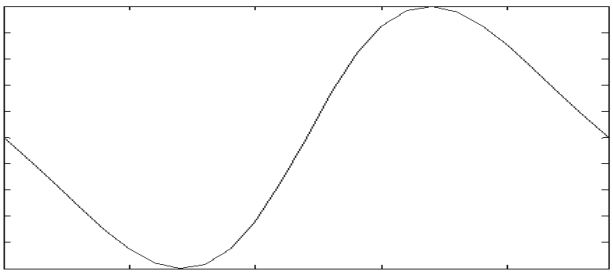

그림 2. 단구간 에너지 궤적의 필터링을 위한 사인형태 함수의 예시
Figure 2 Example of sinusoid-type function for the filtering of short-time energy contour

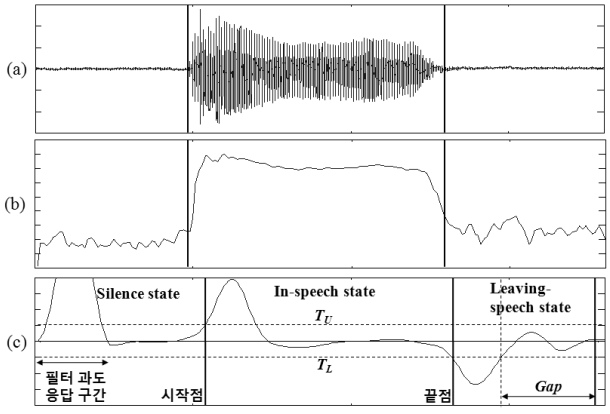

그림 3. (a) 입력 음성파형(한국인 남자 /아/ 발성, 수평축 1칸은 0.5초에 해당), (b) 단구간 에너지 궤적, (c) 사인형태의 함수로 필터링된 단구간 에너지 궤적
Figure 3. (a) Input speech waveform(/ah/ sound uttered by a Korean male, 0.5 sec per tick on horizontal axis), (b) Short-time energey contour, (c) short-time energy contour filtered by a sigmoidal function

는 F 값이 작아질 수 있다. 따라서, 단구간 에너지를 필터링한 F 값이 T_U 보다 커질 때 음성의 시작점을 선언할 수 있으며 T_L 보다 작아질 때를 음성의 종료점으로 선언할 수 있다. 물론, 음성 종료를 위해서는 F 값이 T_L과 T_U 사이에 존재하는 구간의 총 단구간 프레임 수가 Gap 이상 검출되어야 한다. <그림 3(b)>의 에너지 궤적에서 검출된 시작 및 종료점과 <그림 3(c)>의 필터링된 에너지 궤적에서 검출된 시작점 및 종료점의 동기가 맞지 않는 이유는 필터링 과정에서 생기는 시간 지연 때문으로 볼 수 있다.

3. 제안된 상태변수 기반 음성검출 알고리즘

3.1 베이스라인 음성검출 알고리즘 분석

<그림 1>에 표현된 베이스라인 음성검출 알고리즘은 음성검출을 위한 인간의 발성특성을 비교적 잘 표현하고 있으며 SNR(signal-to-noise ratio)이 높을 때 좋은 성능을 발휘할 수 있다. 베이스라인 음성검출 알고리즘의 단점은 다음과 같다. 첫 번째 단점으로, 입력 신호로부터 추출된 단구간 에너지를 <그림 2>의 사인 형태의 함수로 필터링을 수행하는데, <그림 3(c)>의 예에서 확인할 수 있듯이 필터링된 궤적값 F가 과도응답 구간을 보이는 구간이 존재한다는 것이다. 일반적인 음성검출 알고리즘의 경우 초기 100~200 ms 정도를 순수 잡음 구간이라고 가정하고 잡음의 통계치를 추출한다[9]. 즉, <그림 1>의 베이스라인 음성검출 알고리즘에서는 초기 배경잡음의 분산값을 분석하여 최적의 T_L, T_U 값을 결정할 수 있을 것이다. 그러나, 잡음으로 간주될 수 있는 구간에서 필터링된 궤적값이 과도 응답을 보일 경우에 잡음의 통계치 추출을 위해서 순수 잡음 구간의 영역을 음성 녹음 시작부터 더욱 긴 시간이라고 가정해야 한다.

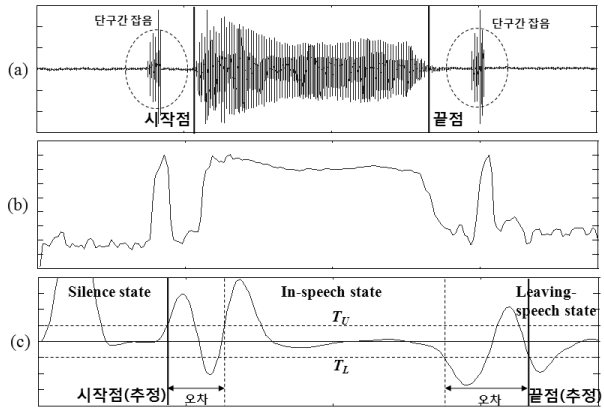

그림 4. (a) 단구간 잡음이 입력 음성의 시작점 끝점 부근에 포함된 경우의 예, (b) 단구간 에너지 궤적, (c) 단구간 에너지 궤적을 사인 형태의 함수로 필터링한 결과

Figure 4. (a) Input speech waveform including short-duration noises near the start and the end point of the waveform, (b) Corresponding short-time energy contour, (c) Corresponding filtered short-time energy contour

그렇게 될 경우에 사용자의 발성 특성에 따라서 순수 잡음 구간으로 가정된 초기 신호에 음성이 포함될 수 있는 가능성이 커지게 된다. 두 번째 단점으로는 음성이 아닌 짧은 구간에 존재하는 높은 에너지를 갖는 잡음이 입력될 경우에 오동작이 발생할 수 있다는 것이다. <그림 4>에서 오동작이 발생할 수 있는 음성입력의 예를 나타내었다. <그림 4>에서 알 수 있듯이 에너지가 충분히 큰 단구간 잡음이 입력될 경우에 필터링된 에너지 궤적이 음성의 것과 큰 차이를 보이지 않으므로 음성검출 오류를 발생시키게 된다. 이러한 오차의 가장 큰 이유는 음성으로 간주할 수 있는 펄스의 시간축에서의 지속시간 정보를 사용하지 않기 때문이다.

3.2 제안된 음성검출 알고리즘

표 1. 제안된 방식의 음성검출 개념 정리

Table 1. Summarization of proposed speech endpoint detection

시작점	어떤 임계치 이상의 단구간 에너지가 MPL(minimum pulse length) 회 이상 반복될 때
끝점	어떤 임계치 이하의 단구간 에너지가 HTH(hangover threshold) 이상 반복될 때
잡음	①단구간 에너지의 값이 TH 이하일 때 ②어떤 음향 펄스의 지속시간이 MPL 이하일 때 ③시작점과 끝점이 정상적으로 검출된 후 계산된 추정 음성 구간의 지속시간MSL(minimum speech length) 이하일 때

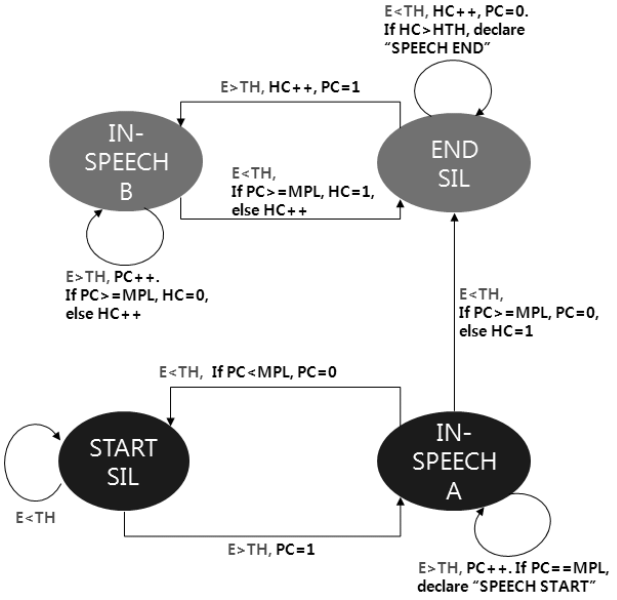

그림 5. 제안된 음성검출 알고리즘

Figure 5. Proposed speech endpoint detection algorithm

제안된 음성검출 알고리즘 역시 상태변수 기반의 알고리즘의 일종이며 적은 계산량을 사용하기 위해서 신호의 주파수 영역 분석 등의 기법은 사용하지 않는다. 즉, 음성의 실시간 검출

을 위해서 입력 신호의 단구간 에너지 궤적과 검출된 음향 펄스의 지속시간 정보를 이용한다. 제안된 알고리즘에서 음성의 시작점 및 끝점을 검출하기 위한 개념을 <표 1>에 정리하였다.

<표 1>의 개념을 바탕으로 구성한 음성검출 알고리즘을 <그림 5>에 나타내었다. <그림 5>에서 E는 단구간 에너지, TH는 단구간 에너지의 임계치, PC(pulse count)는 에너지가 큰 음향 펄스의 개수를 측정하기 위한 변수, HC(hangover count)는 음성의 종료점을 측정하기 위해서 음성 시작점 검출 후에 에너지가 낮은 배경잡음 구간의 연속적 개수를 세는 변수이다. 수식 (1)에서 단구간 에너지의 정의를 나타내었다.

$$E_t = \sum_{n=0}^{N-1} (w(n)x_t(n))^2 \tag{1}$$

여기서, $x_t(n)$은 t번째 단구간 입력 신호의 n째 표본값을 의미한다. $w(n)$은 추출된 에너지 궤적의 특성을 개선하기 위한 창함수이며, 본 연구에서는 해밍창 함수를 사용하였다. 음성 검출을 위한 단구간 에너지 임계치 TH는 수식 (2)를 사용하여 구해진다.

$$TH = NSG \times \left(\frac{1}{T_N} \sum_{t=0}^{T_N-1} E_t \right) \tag{2}$$

여기서, NSG(noise-to-speech gain)는 음성검출을 위한 평균 배경잡음 에너지의 배수, T_N은 순수 잡음구간으로 간주될 수 있는 초기 입력 프레임의 개수이다. 제안된 상태변수 기반 음성 검출 알고리즘의 설명은 다음과 같다. <그림 5>에서 START_SIL 및 IN_SPEECH_A 상태는 음성의 시작점을 판정하기 위한 상태이다. <표 1>에서 나타낸 바와 마찬가지로

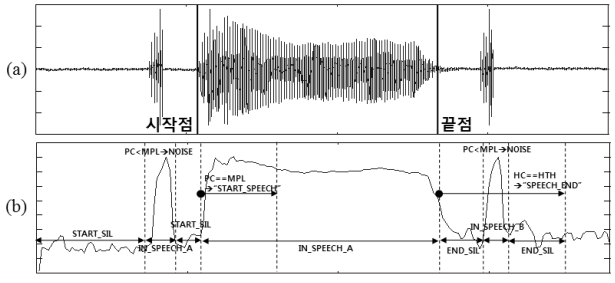

그림 6. (a)단구간 잡음이 입력 음성의 시작점 끝점 부근에 포함된 경우의 예, (b)단구간 에너지 궤적 및 제안된 검출 방식에서의 상태 천이

Figure 6. (a) Input speech waveform including short-duration noises near the start and the end point of the waveform, (b) Short-time energy contour and state transition in the proposed algorithm

IN_SPEECH_A 상태에서는 단구간 에너지의 크기가 임계치인 TH 이상인 프레임의 개수가 MPL 이상으로 존재하는지를 측정하며, 만약 음향 펄스의 지속시간이 MPL이 되는 순간 "SPEECH_START"를 선언한다. 만약, 음향 펄스의 지속시간이 MPL 이하가 되면 검출된 펄스의 지속시간이 너무 짧다고 판단하고 잡음 구간인 IN_SPEECH_A 상태에서 음성 시작을 다시 기다리게 된다. IN_SPEECH_B 및 END_SIL 상태는 음성의 종료점을 검출하기 위한 상태이다. IN_SPEECH_A 상태에서 음성의 시작점이 검출된 후에 단구간 음성 입력의 크기가 임계치 이하로 떨어지면 END_SIL로 상태 천이가 발생하게 되며 해당 단구간 입력 프레임을 short-pause 혹은 실제 음성종료의 후보점으로 간주하게 된다. 만약, short-pause일 경우에는 HC 변수가 임계치인 HTH 보다 작게 되며 IN_SPEECH_B 상태로 천이하게 된다. 만약, HC 변수가 HTH와 같아지면 즉시 "SPEECH_END"를 선언하게 된다. 음성의 종료 부근에서 발생하는 짧은 잡음성 펄스를 제거하는 로직은 IN_SPEECH_B 상태에서 자체 회전하는 부분에 있다. 즉, 음성의 종료 부근에서 잡음성 펄스가 들어왔다고 가정하면, 펄스의 지속시간이 임계치인 MPL 이하일 경우에는 비록 에너지가 높다 하더라도 잡음으로 간주하여 HC를 증가시키게 된다. <그림 4>에서 사용된 입력 신호를 제안된 방식의 알고리즘에 입력하였을 때 얻을 수 있는 상태변수의 천이 과정을 <그림 6>에 나타내었다. <그림 6>에서 알 수 있듯이 시작점 및 끝점 부근에 포함된 잡음성 펄스가 제안된 상태변수 모델링을 통해서 효과적으로 제거될 수 있음을 알 수 있다.

4. 실험 및 결과

4.1 성능 비교 대상의 알고리즘

제안된 음성검출 알고리즘의 성능 평가를 위해서 2장에서 소개한 베이스라인 알고리즘과 참고문헌 [10]에서 소개된 잡음환경에서의 강인한 끝점 검출 방식을 선택하였다. 참고문헌 [10]에서는 잡음에 강인한 끝점 검출을 위해서 에너지가 높은 몇 개의 에너지 밴드를 입력신호의 주파수별 에너지 분포를 분석하여 선택하여 스펙트럼 엔트로피를 측정한다. 만약, 측정된 엔트로피가 임계치 이상이면 음성구간으로 선언되고 그렇지 않으면 잡음 구간으로 선언된다. 참고문헌 [10]에서는 실시간 검출 방식을 제안하지 않았기 때문에 본 연구에서는 엔트로피 임계치를 넘는 프레임의 수가 SPC_TH 이상일 때 시작점을 선언하고, 엔트로피 임계치 이하의 프레임 수가 EPC_TH 이상일 때 끝점을 선언한다. 수식(3)에 엔트로피 임계치를 구하는 방법을 나타내었다[10].

$$T_s = \mu + \alpha \times \sigma \tag{3}$$

여기서, μ,σ는 초기 200 ms 에서 구한 스펙트럼 엔트로피의 평균 및 표준편차이며 α는 보정상수이다.

4.2 테스트 DB

제안된 음성검출 알고리즘의 성능은 다채널 마이크로폰 기반의 빔포밍 후의 결과에 대해서 측정이 되었다. 사용된 빔포밍 알고리즘은 주파수 영역의 GSC(generalized sidelobe canceller)이며, 채널의 수는 4, 채널별 주파수 분석을 위한 FFT 크기는 512였다. GSC 동작을 위한 적응 모드 제어는 선택된 2채널에서 계산된 정규화된 상호상관도로부터 음성부재확률을 측정하여 수행되었다. 주파수영역에서의 GSC 및 음성부재확률 기반의 적응모드제어기의 설계 방법은 참고문헌 [6]과 동일하다. 빔포밍을 위한 테스트 DB 구축을 위하여 PBW452에서 무작위로 추출된 1808개를 고성능 음향스피커로 재생시켜 수집되었다[8]. 녹음 장비는 (주)HCILab에서 제작한 8UMA0709가 사용되었고 녹음 장소로 전자통신연구원에서 제공한 12 m x 7 m x 3 m 크기의 실험실을 활용하였다. <그림 7>에 수집 환경을 나타내었다. 모든 DB는 표본화율 16 kHz, 표본해상도 16 bit를 사용하여 수집되었으며 알고리즘의 SNR 별 성능 측정을 위해서 목표 음성과 잡음을 별도로 녹음한 후에 잡음의 크기를 조절하여 인공적으로 가산하였다. 본 연구에서 고려된 입력 SNR은 -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB 였으며 이에 따라서 수집되는 총 DB의 개수는 2가지 잡음원이 고려될 때 총 4(채널) x 21696개였다. 잡음신호로는 여러 사람의 목소리가 포함되어있는 TV 오락프로그램 및 Westlife의 곡 Mandy를 고성능 음향스피커로 재생하였다. 잡음원, 녹음장비, 목표 음원이 이루는 각도는 45도였다. 다채널 마이크로폰은 지름 33 cm의 원에 90 도 간격으로 배치되었다. <그림 8>에 주파수영역 빔포밍 전후의 결과를 나타내었다. <그림 8>에서 수신되는 잡음신호가 비정상성일 때는 빔포밍 후에도 에너지 궤적이 잡음 구간에서 평탄하지 않을 것임을 예상할 수 있다. <그림 8(b)>에서 음성시작 앞부분에 에너지가 높은 짧은 구간을 확인할 수 있는데, 참고문헌 [5]의 베이스라인 알고리즘을 사용할 경우에 시작점 오류가 발생할 수 있다. 빔포밍 후에 얻을 수 있는 SNR 이득은 평균 9.2 dB 였으며 입력 SNR에 대해서 큰 편차를 보이지 않았다.

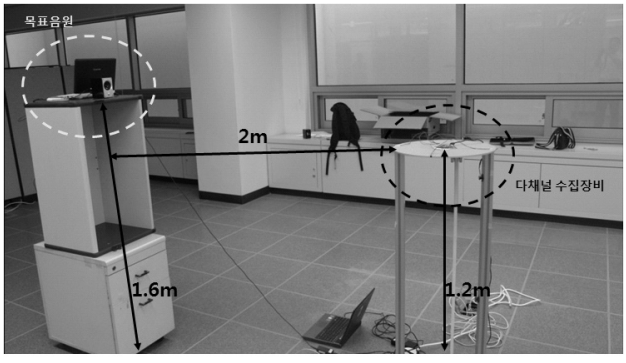

그림 7. 다채널 DB 구축 환경
Figure 7. Environment for multi-channel DB construction

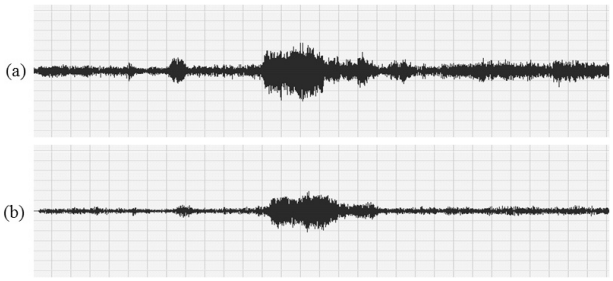

그림 8. (a) SNR 5 dB 입력(/과언이/, 사람 목소리 잡음), (b) 주파수영역 GSC의 결과
Figure 8. (a) Input with 5-dB input SNR(/gwaΛni/, human voice for target speech-corrupting noise), (b) Resulting waveform by frequency-domain GSC

4.3 파라미터 최적화

제안된 방식의 음성검출을 위해서 수식 (1)에서 나타낸 단구간 에너지 추출을 위한 프레임의 길이 N은 480, 수식 (2)에서 순수 잡음 구간으로 간주될 수 있는 프레임의 개수 T_N은 20으로 두었으며 다른 비교대상의 알고리즘에 대해서도 동일한 가정을 하였다. 음성검출을 위해서 NSG, MPL, HTH, MSL 등의 파라미터가 사용되는데 최적화의 대상이 되는 파라미터는 NSG와 MPL 이었다. HTH, MSL 파라미터는 음성신호의 특성을 고려하여 각각 25, 20을 할당하였다. 파라미터 최적화를 위한 성능지수로 음성인식률을 사용하였다. 음성인식기는 HTK를 활용하여 구성되었으며 잡음이 섞이지 않은 대용량 PBW DB에서 추정된 음소별 HMM을 빔포밍을 거친 10848개의 DB에 MLLR 기반의 적응 기법을 통해서 생성하였다. MLLR을 위한 음소 그룹의 개수는 36개였으며 음성인식을 위한 특징 파라미터로 매 10 ms 당 MFCC(mel-frequency cepstral coefficient) 12차 및 에너지, 그것의 속도, 가속도 성분을 추출하였다[7]. 분석 프레임의 크기는 30 ms 이며 과거 프레임 20 ms가 중첩된다. 이러한 HTK기반 음성인식 시스템을 기반으로 본 연구에서는 파라미터의 최적화를 위해서 전수조사(exhaustive search) 방식을 선택하였다. NSG는 1.2~2.5 사이에서 0.1 단위로 검색이 수행되며 MPL은 3에서 15사이에서 1 단위로 검색이 수행되었다. 전수조사의 입력으로 사용되는 DB의 개수가 너무 커지면 파라미터 검색에 많은 시간이 소요되므로 실제 빔포밍 기반의 음성인터페이스에서 알고리즘 동작 환경으로 고려될 수 있는 입력 SNR 5 dB의 데이터 총 1808개를 선택하였다. <그림 1>의 베이스라인 알고리즘에 사용되는 파라미터 T_L, T_U 역시 전수조사 기법으로 최적화되었다. 참고문헌 [10]의 방식에서 전수조사 최적화 대상의 파라미터는 4.1절에서 제시한 SPC_TH, EPC_TH, α가 해당된다. 그 결과를 <표 2>에 정리하였다.

표 2. 전수조사 기법으로 추정된 최적 파라미터
Table 2. Optimal parameter values by exhaustive search

베이스라인 알고리즘	T_L	20.2
	T_U	-17.5
참고문헌 [10]	SPC_TH	8
	EPC_TH	27
	α	2.2
제안된 알고리즘	NSG	1.3
	MPL	8

4.4 음성인식률 측정 결과

최적 파라미터값이 사용되었을 때의 음성인식률을 그림 9, 10에 정리하였다. 잡음원이 사람 목소리일 경우, 빔포밍 입력 SNR이 5 dB 일 때 수작업 음성검출에 대해서의 음성인식률은 80.3 %, 베이스라인 알고리즘에 대해서는 76.4 %, 참고문헌 [10]의 알고리즘에 대해서는 77.4 %, 제안된 방식에 의해서는 78.5 %의 성능을 보였다. 모든 입력 SNR에 대해서 제안된 방식이 베이스라인 알고리즘에 비해서 높은 성능을 보였으며 입력 SNR이 낮을수록 향상 폭이 더 큼을 확인할 수 있었다. 전체 테스트 DB에 대한 평균 음성인식률은 수작업 음성검출이 72.8 %, 베이스라인 방식이 66.9 %, 참고문헌 [10] 방식이 68.1 %, 제안된 방식이 69.3 %의 성능을 보였다. 잡음원이 음악일 경우, 빔포밍 입력 SNR이 5 dB 일 때 수작업 음성검출에 대해서의 음성인식률은 79.8 %, 베이스라인 알고리즘에 대해서는 82.4 %, 참고문헌 [10]의 알고리즘에 대해서는 81.0 %, 제안된 방식에 의해서는 81.1 %의 성능을 보였다. 입력 SNR 5 dB에 대해서는 베이스라인 음성인식률이 제안된 방식보다 1.3 % 정도의 상승을 보였으나 나머지 입력 SNR에 대해서는 제안된 방식이 더 나은 성능을 보였다. 전체 테스트 DB에 대한 평균 음성인식률은 수작업 음성검출이 72.6 %, 베이스라인 방식이 68.3 %, 참고문헌 [10]의 방식이 69.1 %, 제안된 방식이 69.9 %의 성능을 보였다. <그림 10>에서 수작업 검출의 인식률이 알고리즘에 의한 인식률보다 더 낮게 나오는데 실제 수작업 끝점 검출 오류는 발생하지 않았을 때의 결과이다. 이러한 현상은 잡음환경 하에서의 음성인식 성능 측정 실험에서 빈번하게 발생되는데, 음성인식 어휘 네트워크에서 HMM의 구성을 'SILENCE-WORD-SILENCE' 형태로 두는 과정에서 배경잡음에 의한 스코어 변동 혹은 SILENCE와 인접해 있는 음소 HMM의 추정 오류 때문에 생기는 것으로 판단된다. 실험결과를 종합적으로 평가할 때 본 연구에서 제안된 음성의 상태변수 모델링 방식이 두 가지 비교 대상 알고리즘에 비해서 잡음환경에서 더 좋은 성능을 나타냄을 확인할 수 있다. 참고문헌 [10]의 음성검출을 위한 엔트로피 궤적은 차량주행 잡음, 백색 잡음 등의 정상성 잡음환경일 때 일반 단구간 에너지 궤적보다 더 낮은 특성을 보인다. 그러나 본 연구에서와 같이 사람 목소리, 음악 등 비정상성 잡음 환경에서 원거리 음성입력이 이루어질 경우에는 알고리즘의 유용성이 크지 않음을 알 수 있다.

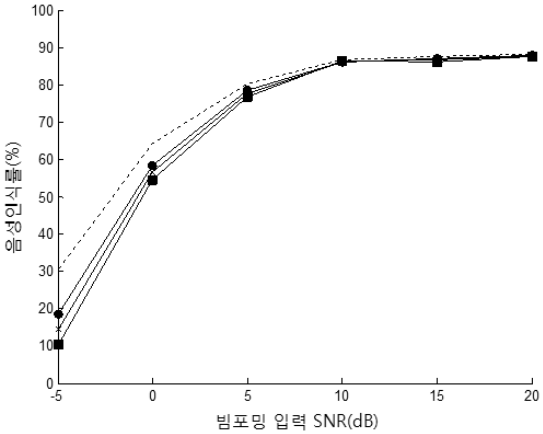

그림 9. 잡음원이 사람 목소리일 때 빔포밍 입력 SNR에 따른 음성인식률(점선: 수작업 EPD, ■: 베이스라인 알고리즘, x: 참고문헌 [10], ●: 제안된 알고리즘)
Figure 9. Speech recognition rates according to beamformer input SNR for human voice noises(dotted: manual EPD, ■: baseline algorithm, x: reference [10], ●: proposed algorithm)

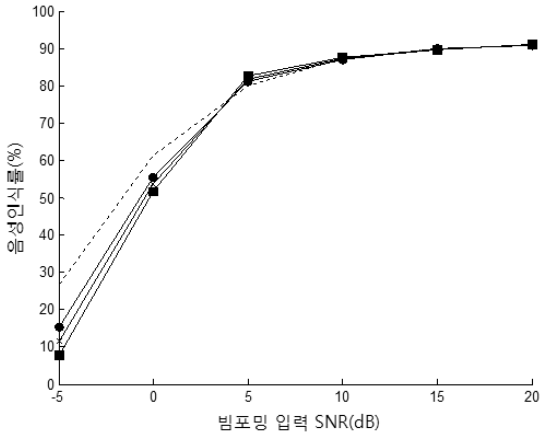

그림 10. 잡음원이 음악일 때 빔포밍 입력 SNR에 따른 음성인식률
Figure 10. Speech recognition rates according to beamformer input SNR for music noises

5. 결론

본 논문에서는 단구간 펄스성 잡음 입력에 강인한 상태변수 모델링 기반의 실시간 음성검출 방식에 대해서 제안하였으며 파라미터의 최적화 기법으로 음성인식률을 성능 지수로 하는 전수조사 기법을 사용하였다. 테스트 DB로는 주파수영역 GSC 알고리즘으로 잡음이 어느 정도 제거된 파형을 사용하였고 최적 음성인식률 측정 결과에서는 제안된 방식이 기존의 상태변수 모델링 방식보다 더 우수한 성능을 보임을 확인할 수 있었다. 향후 연구 계획으로는 주파수 영역에서 추출된 켑스트럼,

엔트로피 등 다양한 특징벡터를 활용하여 검출된 음향 펄스가 실제 목표 음원에서 발생한 인간의 음성인지를 판별하는 알고리즘을 접목하는 것이다. 현재, 지능형 홈 환경에서의 로봇 관련 국책과제가 전자통신연구원에서 다년간 수행되고 있는데 본 연구 결과가 인간과 로봇간의 원거리 음성인식 알고리즘의 전처리기로서 활용될 수 있을 것으로 기대된다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT성장동력 기술개발사업의 일환으로 수행하였음.[2008-F-037-01, u-로봇 HRI 솔루션 및 핵심소자 기술개발]

참고문헌

[1] Rabiner, L., Juang, B. (1993). *Fundamentals of Speech Recognitions*, Prentice Hall.

[2] Shen, J., Hung, J., and Lee, L. (1998). "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments," *Int. Conf. on Spoken Lang. Processing*, CD-ROM, Sydney.

[3] Zhang, X., Zhao, Z., and ZhaoA, G. (2006). "Speech Endpoint Detection Method Based on Wavelet Coefficient Variance and Sub-Band Amplitude Variance", *First International Conference on Innovative Computing, Information and Control*, Vol. 3, pp. 83-86.

[4] Rabiner, L., Schafer, R. (1978). *Digital Processing of Speech Signals*, Prentice Hall.

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

[6] Han, S., Hong, J., Jeong, S., and Hahn, M. (2010). "Robust GSC -based speech enhancement for human machine interface", *IEEE Trans. Consumer Electronics*, Vol. 56, No. 2, pp. 965-970, May.

[7] http://htk.eng.cam.ac.uk

[8] http://www.sitech.or.kr

[9] ETSI ES 202 212(2005). speech processing, transmission, and quality aspects(STQ), v.1.1.2.

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

• **김수환 (Kim, Suhwan)**
경상대학교 공과대학
경남 진주시 가좌동 900번지
Tel: 055-751-5357 Fax: 055-759-7814
Email: edps2166@gnu.ac.kr
관심분야: 음성신호처리
현재 경상대학교 전자공학과 석사과정

• **이영재 (Lee, Youngjae)**
경상대학교 공과대학
경남 진주시 가좌동 900번지
Tel: 055-751-5357 Fax: 055-759-7814
Email: clever1999@gnu.ac.kr
관심분야: 음성신호처리
현재 경상대학교 전자공학과 석사과정

• **김영일 (Kim, Young-Il)**
경상대학교 공과대학(공학연구원)
경남 진주시 가좌동 900번지
Tel: 055-751-5352 Fax: 055-759-7814
Email: yi@gnu.ac.kr
관심분야: 음향공학, 음성신호처리
현재 경상대학교 전자공학과 교수

• **정상배 (Jeong, Sangbae),** 교신저자
경상대학교 공과대학(공학연구원)
경남 진주시 가좌동 900번지
Tel: 055-751-5357 Fax: 055-759-7814
Email: jeongsb@gnu.ac.kr
관심분야: 음성신호처리
현재 경상대학교 전자공학과 조교수