한국음향학회지 제28권 제5호 pp. 447-453 (2009)

잡음 환경에서 심리음향모델 기반 음성 에너지 최대화를 이용한 음성 검출 방법

Voice Activity Detection Method Using Psycho-Acoustic Model Based on Speech Energy Maximization in Noisy Environments

최 갑 근*, 김 순 협*
(Gab-Keun Choi*, Soon-Hyob Kim*)

*광운대학교 대학원 컴퓨터공학과
(접수일자: 2009년 5월 23일; 수정일자: 2009년 6월 22일; 채택일자: 2009년 6월 26일)

이 논문은 음성 에너지를 최대화 하여 낮은 SNR환경에서 음성 존재 여부를 판단하고 정확한 끝점을 검출하는 방법에 대한 것이다. 전통적인 VAD (Voice Activity Detection) 알고리듬은 잡음의 추정치를 이용해 음성과 비음성 구간을 선택하여 낮은 SNR환경이나 비안정 잡음환경에서는 정확하지 못한 문턱값으로 인해 부정확한 끝점검출을 하였다. 또한 잡음의 시간적 변화를 반영하기 위해 비교적 큰 분석 구간을 두어 계산량이 증가함에 따라 실제 응용에 적합하지 않은 단점이 있다. 이 논문은 잡음환경에서 정확한 음성 구간의 검출을 위해 심리음향 모델에 기반 한 바크 스케일 필터 뱅크를 이용하여 주어진 프레임에서 음성 에너지를 최대화 시키고 잡음을 억제하는 SEM-VAD (Speech Energy Maximization-Voice Activity Detection) 방법을 제안하였다. 다양한 잡음환경, SNR 15 dB, 10 dB 5 dB 0 dB 상황에서 실험한 결과 SNR의 변화에 안정적인 문턱값을 얻었고, 음성 검출을 위한 실험에서 자동차 잡음 환경에 대한 PHR (Pause Hit Rate)은 모든 잡음 환경에서 100%의 정확도를 보였고, FAR (False Alarm Rate)는 SNR 15 dB와 10 dB에서는 0%, SNR 5 dB에서 5.6% SNR 0 dB에서 9.5%의 성능을 보였다.
핵심용어: 음성검출, 음성인식
투고분야: 유향 신호처리 분야 (1)

This paper introduces the method for detect voices and exact end point at low SNR by maximizing voice energy. Conventional VAD (Voice Activity Detection) algorithm estimates noise level so it tends to detect the end point inaccurately. Moreover, because it uses relatively long analysis range for reflecting temporal change of noise, computing load too high for application. In this paper, the SEM-VAD (Speech Energy Maximization-Voice Activity Detection) method which uses psycho-acoustical bark scale filter banks to maximize voice energy within frames is introduced. Stable threshold values are obtained at various noise environments (SNR 15 dB, 10 dB, 5 dB, 0 dB). At the test for voice detection in car noisy environment, PHR (Pause Hit Rate) was 100% accurate at every noise environment, and FAR (False Alarm Rate) shows 0% at SNR 15 dB and 10 dB, 5.6% at SNR 5 dB and 9.5% at SNR 0 dB.
Keywords: Voice Activity Detection, Speech Recognition
ASK subject classification: Acoustic Signal Processing (1)

I. 서 론

잡음환경에 강인한 음성인식을 위해서는 잡음에 영향을 받지 않고 정확한 끝점을 검출해 내는 끝점 검출기가 반드시 필요하다. 끝점 검출기는 인식될 음성의 시작과 끝을 비음성구간과 구분하여 주게 된다. 따라서 신뢰성 있는 음성검출기는 비음성 선택으로 인한 인식 오류를 감소시킨다. 음성 인식률 향상을 위한 음성 검출기는 정확성 향상은 물론 음성인식 시스템의 속도까지 고려해야 한다. 현재 사용되어지고 있는 잡음에 강인한 음성 인식기는 ETSI ASDR (Advanced Distributed Speech Recognition)과 같이 잡음 제거 기반 Wiener-filter 기술을 사용 한다 [1].

음성 검출은 크게 신호의 에너지 (Energy), 영교차율

책임저자: 최 갑 근 (cocomm@kw.ac.kr)
139-701 서울시 노원구 월계동 447-1 광운대학교 컴퓨터공학과
(전화: 02-940-5123; 팩스: 02-941-8919)

(ZCR:Zero Crossing Rate), LPC 파라미터 등과 같은 특징들을 이용하는 방법과, Likelihood Ratio 등과 같은 통계적인 특성에 기반 한 방법들로 나눌 수 있다 [2]. Rabiner, L. R.등에 의해 소개된 ZCR과 Energy를 이용한 끝점 검출 방법은 시간 영역에서 비교적 적은 연산량을 요구하여 일반적으로 사용되어진다 [2]. 그러나 신호의 에너지가 낮은 SNR에서는 성능이 급격히 저하되며, 영교차율은 잡음의 종류에 따라 무성음과 구분되지 않는 단점이 있어 다른 방법들과 같이 쓰이는 보조 역할을 한다. 한편 통계적인 특성을 기반으로 하는 방법들은 많은 연구들에서 좋은 성능을 보여주고 있지만, 계산량이 많거나 잡음의 통계특성이 음성신호와 비슷한 경우 성능이 저하되는 경우가 있어 이를 보완한 여러 가지 방법들이 제안되고 있다 [5][9].

근래에는 잡음환경에서 신호의 에너지나 영교차율 등으로 음성 검출이 어려운 문제점을 개선하기 위해 통계적인 특성에 기반한 방법들이 많이 사용되고 있다. 특히 백색잡음 (White Noise)등에서 성능이 좋은 통계 특성을 나타내는 분산 값을 이용한 방법은 특정 프레임내의 에너지에 대해 기대 값의 변화량을 이용하는 방법으로 프레임 내의 전체 대역에너지가 일정한 에너지 패턴을 보이는 잡음추정치에 비해 일정 대역에 몰려있는 음성은 그 분산 값이 서로 상이하게 다른 점을 이용한 방법이다 [6].

통계적인 특성을 이용한 또 다른 방법으로는 엔트로피를 이용한 방법이 있다. 엔트로피를 이용한 방법 중 특히 Reveney. 등은 심한 잡음환경의 음성신호에 백색잡음을 부가하는 과정을 거쳐, 인위적인 백색잡음이 더해진 신호에서 잡음에 둔감한 특징 추출이 가능하도록 하였으나, 이 방법은 계산량이 많아 실시간에 적합하지 않은 단점이 있다 [3][4][7].

본 논문에서는 기존의 음성검출 및 음성강화 알고리듬들의 대부분이 잡음을 추정하고 잡음의 변화를 시간의 변화에 따라 추적하는 학습알고리듬을 사용하여 잡음추정의 오차로 인한 알고리듬의 성능저하와 잡음추정의 갱신으로 인한 연산량 증가를 개선하기 위해 주어진 프레임 내에서 음성에너지를 최대화 시키고 잡음을 억제하는 SEM-VAD (Speech Energy Maximization-Voice Activity Detection) 방법을 제안하였다. 제안된 알고리듬은 현재 프레임과 이전 프레임을 비교 하거나 100 ms 이상의 긴 시간을 분석구간으로 사용하여 실제 환경에 응용하기에 적합하지 않은 것을 개선하기 위해 주어진 프레임에서 음성 에너지의 최대화를 통해 성능을 개선하였고, 실시간 응용에 적합하고 낮은 SNR환경에서의 성능개선을 보기 위해 15 dB ~ 0 dB로 변화하는 잡음 음성 DB인 NOIZEUS를 이용하여 음성 검출 실험을 하였다.

본 논문의 구성은 다음과 같다. 제 2장에서 심리음향 모델에 기반 한 음성 에너지 최대화를 설명하고 제 3장에서 주어진 프레임 내에서 음성을 잘 표현하도록 바크 스케일 필터 뱅크를 이용하여 PSR (Peak to Side-lobe Ratio)을 계산하고, 음성에너지를 최대화 시켜 잡음억제와 음성 검출이 이루어지는 과정을 설명한다. 제 4장에서는 제안한 음성검출 방법의 성능을 검증하기 위해 실시한 실험환경과 결과를 설명하고, 제 5장에서 결론을 맺는다.

II. 심리음향 모델에 기반한 음성에너지 최대화

인간의 청각특성은 음성대역인 350 Hz~3.4 kHz 대역에 민감하다. 이에 따라 음성은 이대역에 중요한 정보를 많이 갖고 있다. 이러한 이유는 귀의 구조와 관계가 있으며 이와 같이 인간의 귀에 음으로서 들리기 위해서는 음파의 세기와 주파수에 어느 한계가 있다.

2.1. 최소가청한계 (threshold of hearing)

주파수 측에서 들리는 범위는 최저 가청한계 (lower limit of hearing)와 최고 가청한계 (upper limit of hearing)사이이며, 음의 세기에 대한 범위는 최소가청한계 (threshold of hearing)와 최대 가청한계 (threshold of feeling)사이이다. 절대 최소가청한계 (absolute of hearing)는 잡음이 없는 환경에서 사람이 소리를 감지할 수 있는 최소가청한계를 말한다. 이 한계는 1940년 H. Fletcher [10]에 의해 보고되었다. 이를 함수로 나타내면 비선형 함수로 다음과 같이 근사화 할 수 있다.

$$T_q = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000 - 3.3)} + 10^{-3}(f/1000)^4 \quad (1)$$

여기서 $T_q(f)$는 주파수 영역에서 최소가청한계를 나타낸다.

2.2. 크리티컬 밴드 (critical band)

크리티컬 밴드 (critical Band)는 마스킹 효과가 일어나는 주파수의 폭을 말하며 주파수 크기에 따라 다르게 변한다. 500 Hz까지는 약 100 Hz의 대역을 가지고 500 Hz

이상에서는 중심주파수의 약 20%대역을 가진다.

크리티컬 밴드 대역 ($BW_c(f)$)은 식 (2)와 같이 함수식으로 나타낼 수 있다 [10].

$$BW_c(f) = 25 + 75(1 + 1.4(f/1000)^2)^{0.69} \ (Hz) \quad (2)$$

크리티컬 밴드의 넓이를 바크 (Bark)라고 하며, 주파수단위 ($z(f)$)를 바크 단위로 변환 하는 식은 다음과 같다 [8].

$$z(f) = 13\arctan(0.0076f) + 3.5\arctan[(\frac{f}{7500})^2] \ (\text{Bark}) \quad (3)$$

바크 스케일 (Bark scale) 주파수는 음성의 중요 요소가 있는 대역에 대한 표현이 우수하다. 본 논문은 입력신호 $x(i)$에 대해 FFT를 취해 얻어진 부밴드 즉 프레임내의 각 주파수 빈 (Frequency bin)에 대해 바크 스케일로 변환 처리 하였다. 아래 표 1은 바크 스케일 대역과 주파수 빈 (Frequency bin)을 표시한다.

표 1에서 보는바와 같이 음성의 주요 특징이 밀집되어 있는 대역에서 바크 스케일은 밴드의 폭이 매우 촘촘하다. 바크스케일의 이러한 비선형적 밴드 특성은 음성에너지를 표현하기에 적합하게 구성되어 있어 음성에너지 최대화에 도움이 된다.

표 1. 부밴드 주파수 범위
Table 1. Sub-band frequency range.

부밴드	Bin 갯수	주파수 범위 (Hz)
1	3	0 ~ 99
2	3	100 ~ 199
3	3	200 ~ 299
4	3	300 ~ 399
5	4	400 ~ 509
6	4	510 ~ 629
7	4	630 ~ 769
8	5	770 ~ 919
9	5	920 ~ 1079
10	6	1080 ~ 1269
11	7	1270 ~ 1479
12	8	1480 ~ 1719
13	8	1720 ~ 1999
14	11	2000 ~ 2319
15	12	2320 ~ 2699
16	14	2700 ~ 3149
17	18	3150 ~ 3699
18	10	3700 ~ Fs/2

최대화된 음성은 잡음에 대해 상대적으로 SNR이 높아지므로 음성과 비음성을 구분하기에 유리해 진다. 본 논문은 이와 같이 최대화된 음성을 중심으로 문턱값에 의해 음성과 비음성 구간을 결정한다. 그 과정은 3장에서 자세히 설명한다.

III. 음성에너지 최대화를 이용한 잡음제거와 음성검출과정

기존의 음성 검출 및 음성 강화 알고리듬들은 대부분 잡음을 추정하고 잡음의 변화를 시간의 변화에 따라 추적하는 학습 알고리듬을 사용하였다. 하지만 이와 같은 방법들은 잡음의 변화량을 계산하기 위해 현재 프레임과 이전 프레임을 비교거나 100 ms 이상의 긴 시간을 분석구간으로 사용하여 실제 환경에 응용하기에 적합하지 않은 점이 있다 [9]. 본 논문에서는 주어진 프레임 내에서 음성 에너지를 최대화 시키고 잡음을 억제하는 SEM-VAD (Speech Energy Maximization-Voice Activity Detection) 방법을 제안한다. 제안된 방법은 주어진 프레임에서 모든 것을 처리하기 때문에 분석 구간이 상대적으로 적어 실제 응용에서 유리하다.

제안된 알고리듬을 처리하기 위해 입력신호에 대한 단구간 퓨리에 분석은 음성신호 $x(t)$를 식 (5)와 같이 정의하고 주파수 영역에서 처리하기 위하여 식 (7)을 이용하여 DFT (Discrete fourier Transform) 처리하여 식 (6)과 같이 주파수 성분을 $[X_k]$로 표시한다.

$$\begin{aligned} x(t) &= [x_i] \\ &= [x_0, x_1, x_2, \ldots, x_{N-1}] \end{aligned} \quad (5)$$

$$\begin{aligned} [X_k] &= DFT[X_k] \\ &= [X_0, X_1, X_2, \ldots X_{N/2}] \end{aligned} \quad (6)$$

$$X_i = \left| \sum_{k=0}^{N-1} x_k e^{-j(2\pi mk/N)} \right| \quad (7)$$

단 구간 퓨리에 변환된 X_k는 음성에너지 최대화를 위해 선형 주파수를 인간의 청각모델에 기반한 비선형 주파수 크기인 바크스케일로 식 (3)에 의해 변환된다. 여기서 k는 주파수 빈 (frequency bin)의 인덱스 (index)이다.

식 (3)에 의해 바크 스케일로 변환된 밴드는 표 1을 참조하여 바크 스케일 밴드 대역에 맞도록 주파수 빈 (Frequency bin)을 할당한다. 주파수 빈 (Frequency bin)의 주파수 대역은 FFT 크기 256을 기준으로 8 kHz 샘플링을

하였을 때 31.25 Hz를 가진다. 따라서 바크 스케일 밴드 범위에 맞도록 주파수 빈 (Frequency bin)의 개수를 조정한다.

$$\mu(l)=\frac{1}{n}\sum_{b_i=1}^{n-1}|B(b_i,l)| \tag{8}$$

식 (8)은 주어진 프레임 내에서 각 바크 주파수 인덱스들의 평균에너지를 구하기 위한 것으로 l은 주어진 현재 프레임 인덱스이고 b_i는 바크 스케일 인덱스이다.

음성은 모음에서 피치 주파수 (pitch frequency)를 가진다. 이 피치 주파수는 기본 주파수라고도 하며 이 기본 주파수는 음성 영역의 전 대역에 걸쳐 에너지가 가장 큰 특징을 가지고 있어 최대 에너지로 보며, 최소 에너지는 음성 신호와 무관한 잡음 신호로 본다. 따라서 잡음 신호는 주어진 프레임 l에서 최소 에너지로 본다. 본 논문은 음성 에너지의 이와 같은 특징을 이용하여 음성 에너지의 최대화를 위해 주어진 프레임에서 최소 에너지를 식 (9)와 같이 계산한다.

$$P_{\min}(l)=\min\{B(b_i,l),\ldots,B(b_{i-1},l)\} \tag{9}$$

식 (10)에서는 주어진 프레임 l에서 음성 에너지 최대화를 위해 식 (9)에서 계산된 $P_{\min}(l)$ 값을 중심으로 바크 필터 뱅크를 통과한 스펙트럴 피크들의 관계를 조사한다. 음성 에너지는 잡음에 비해 상대적 에너지가 크고, 잡음은 상대적 에너지가 작은 것으로 보고 각 바크 필터 뱅크에 대한 음성 에너지와 잡음 에너지의 비율을 계산하기 위해 식 (9)에서 계산한 $P_{\min}(l)$ 값을 사용하여 각 바크 필터 뱅크의 출력에 대한 에너지의 비율을 식 (10)과 같이 계산한다.

$$PSR(b_i,l)=\frac{B(b_i,l)-P_{\min}(l)}{\mu(l)} \tag{10}$$

식 (10)에서는 음성 에너지와 잡음 에너지의 비율을 계산하였다. 이 비율은 각 바크 필터 뱅크 출력에 대한 음성 에너지와 잡음 에너지의 비율을 나타내며 이 비율은 식 (11)에서 바크 필터 뱅크와 곱해지게 된다. 결과적으로 그림 2의 (b)에서 보여 지는 바와 같이 입력 신호에 대해 식 (10)에서 계산된 PSR (Peak to Side-lobe Ratio)이 곱해지게 되어 음성 에너지는 잡음 에너지에 비해 상대적으로 더욱 커지게 된다.

$$\overline{B}(b_i,l)=B(b_i,l)\times PSR(b_i,l) \tag{11}$$

식 (12)와 식 (13)은 음성 에너지가 최대화된 상태의 바크 인덱스 (bark index) 주파수들에 대해 평균값과 표준편차 값을 계산한다. 식 (14)는 계산된 표준 편차와 평균을 이용하여 바크 인덱스별 음성 에너지에 대한 에너지 비율을 계산하여 식 (11)의 결과에 곱해지게 된다. 음성 에너지는 그림 2에서 보여 지는 바와 같이 최대화된 상태의 에너지 비율을 곱하게 되어 식 (11)의 결과에서 SNR보다 식 (14)의 결과에서 SNR이 매우 커지게 된다.

$$\hat{\mu}(l)=\frac{1}{n}\sum_{b_i-1}^{n-1}|\overline{B}(b_i,l)| \tag{12}$$

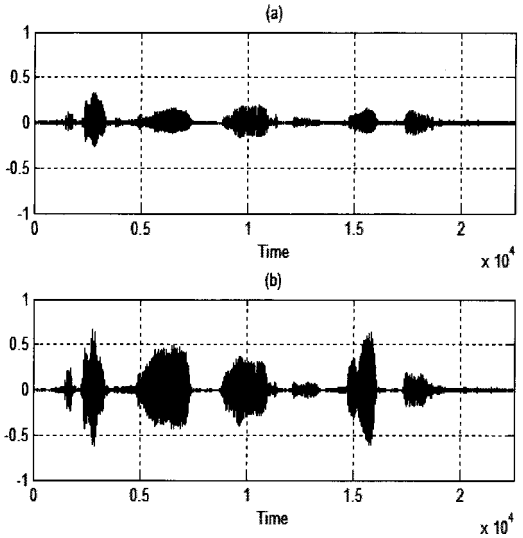

그림 1. (a) 입력신호, (b) 입력신호에 PSR을 곱한 결과
Fig. 1. (a) Input signal, (b) Product of input signal and PSR.

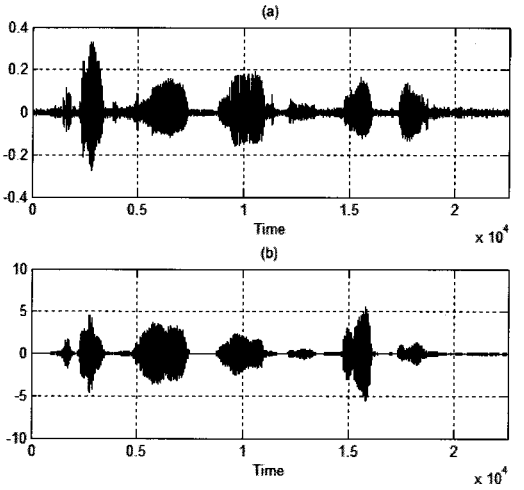

그림 2. (a) SNR 15 dB에서 입력신호 (b) 식 (14)에 대한 결과
Fig. 2. (a) Input signal at SNR 15 dB, (b) Result of Eq. (14).

$$\hat{\sigma}(l)=\left(\frac{1}{n}\sum_{b_i=1}^{n}(\overline{B}(b_i,l)-\hat{\mu}(l))^2\right)^{\frac{1}{2}} \tag{13}$$

$$\hat{B}(b_i,l)=\overline{B}(b_i,l)\times\frac{\overline{B}(b_i,l)-\hat{\mu}(l)}{\hat{\sigma}(l)} \tag{14}$$

식 (14)에 의해 최대화된 음성 에너지만 남고 억제된 잡음들 중 아직 남아 있는 잔존 잡음은 문턱값을 기준으로 음성을 검출하는 시스템에서 부정확한 음성 검출을 실시할 가능성이 있다. 따라서 본 논문에서는 음성 에너지가 최대화된 상태의 바크 크기 주파수의 각 인덱스를 이용하여 식 (12)와 식 (13)에서 평균과 표준편차 값을 구하고 이를 이용하여 $N_{mask}(l)$을 식 (15)와 같이 구한다. 각 프레임에서 표준편차는 음성 에너지의 분포를 표현할 때 유용하다. 음성 에너지의 중요한 성분들은 100 Hz~600 Hz 대역에 집중되어 있는 특성이 있어 모든 가청 영역에 걸쳐 에너지 크기의 편차가 크다. 그러나 잡음의 경우 특히 부색 잡음은 에너지 크기가 모든 가청 영역에 비교적 고르게 분포되어 편차가 작게 된다. 따라서 음성이 존재하는 프레임과 존재하지 않는 프레임의 표준편차 값은 차이가 있고 음성에너지를 최대화한 상태에서의 평균값 역시 차이가 있다. 본 논문에서는 비음성 구간에 대한 구분을 더욱 명확히 하기 위해 식 (15)와 같이 주어진 프레임 l에서 표준편차와 평균값을 더한다. 따라서 $N_{mask}(l)$은 그림 3에서 보여 지는 바와 같이 음성 검출을 위한 문턱값의 기준이 되며 식 (16)에서 N_{thresh}와 비교하여 음성의 존재 여부를 결정한다. N_{thresh}는 실험값으로 본 논문에서는 0.2를 사용하였으며 실험 결과 낮은 SNR상황 (0 dB)과 높은 SNR상황 (15 dB)에서 음성을 검출하기 위한 적절한 문턱값으로 확인되었다.

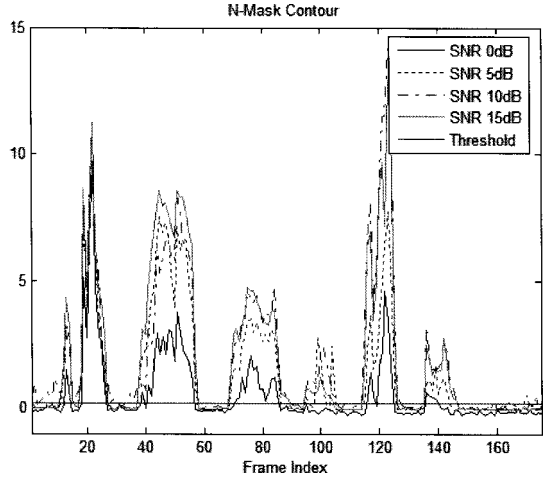

그림 3. 음성에너지 최대화에 대한 N-Mask Contour
Fig. 3. N-Mask Contour for Speech Energy Maximization.

$$N_{mask}(l)=\hat{\sigma}(l)+\hat{\mu}(l) \tag{15}$$

$$\begin{aligned}&\text{if}\quad N_{mask}(l)<N_{thresh}\ ,\quad speech\quad absent\\&else\quad N_{mask}(l)>N_{thresh}\ ,\quad speech\quad present\end{aligned} \tag{16}$$

결과적으로 $N_{mask}(l)$ 값은 주어진 프레임내의 표준 편차와 평균값의 합에 대한 값으로 음성이 존재하는 영역과 음성이 존재하지 않는 영역에 대한 값의 차이를 출력하게 된다. 이 값들은 문턱값과 비교하여 음성과 비음성으로 구분할 수 있고, 특별히 비음성으로 채택된 구간은 음성 구간과의 구분을 명확히 하기 위해 바크 필터 뱅크 에너지에 0을 주어 끝점 검출 성능을 향상 시킬 수 있다.

IV. 실험 및 성능분석

제안된 알고리듬의 성능분석을 위해 Texas Dalls 대학에서 개발한 음성 Corpus, NOIZEUS를 사용하였다. NOIZEUS Corpus는 white gaussian noise, babble noise 등을 포함하며 잡음 환경별로는 street, airport, car noise등으로 구분되며 음성향상 (Speech enhancement) 알고리듬의 성능 검증용으로 사용되는 Corpus이다 [11]. 또한 SNR변화에 대한 성능을 검증하기 위해 잡음환경 (15 dB, 10 dB, 5 dB, 0 dB)을 구분하여 실험하였고, 끝점검출 성능 평가를 위해서는 비음성 구간에 대한 비음성 구간 적중률 (Puase Hit Ratio)과 음성 구간에 대한 비음성 구간 오보율 (False Alarm rate)을 사용하였다. 음원은 8 kHz 샘플링 레이트 (Sampling Rate), 16비트 (Bit)를 사용하였으며 FFT 크기는 256 샘플 (Sample), 1/2 오버래핑 (Overlapping) 구간을 이용하였고 윈도우는 해밍 윈도우 (Hamming Window)를 사용하였다.

실험 결과 그림 4의 (b)에서 보여 주는 바와 같이 SNR 15 dB에서 입력신호 $x(i)$의 음성 에너지는 식 (14)에 의해 계산된 결과 음성 에너지가 증가한 결과를 볼 수 있다. 또한 식 (16)에 의해 잡음 구간으로 채택된 부분에 대해서는 에너지를 0으로 주어 음성 구간과 비음성 구간이 확연히 분리되는 것을 볼 수 있다. 여기서 음성 에너지에 비해 잡음 에너지가 작은 것은 음성 에너지에 비해 상대적으로 커지지 않은 결과이다. 표 2는 잡음환경에 따라 음성검출 성능을 평가한 결과이다. 성능의 척도로 사용되어진 비

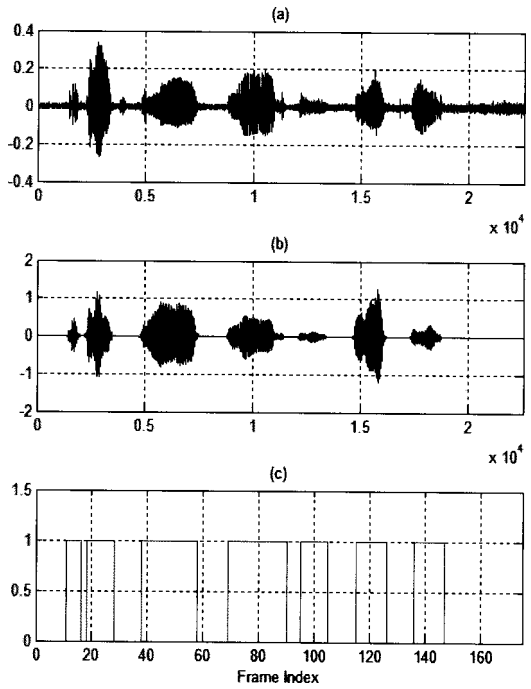

그림 4. (a) SNR 15 dB 입력신호 (b) 식 (16) 계산 결과 (c) 식 (16)에 대한 VAD 결과

Fig. 4. (a) Input signal at SNR 15 dB (b) Result of calculation for Eq. 16 (c) Result of VAD for Eq. 16.

표 2. SNR 15 dB~0 dB에 대한 Pause Hit Rate와 False Alarm Rate

tabte 2. Puase Hit Rate and False Alarm Rate for SNR 15 dB~0 dB.

Noise	SNR (dB)	VAD Result (%)	
		PHR	FAR
Car	0	100	9.5
	5	100	5.6
	10	100	0
	15	100	0
airport	0	91	12
	5	93	9
	10	98	5
	15	98	5
Street	0	95	10
	5	95	7
	10	98	4
	15	98	4

음성 구간 적중률 (Pause Hit Rate)은 car 잡음 환경의 모든 잡음 구간에서 100%의 정확도를 보였으며 음성 구간에 대한 비음성 구간 오보율 (False Alarm Rate)은 SNR 15 dB 와 10 dB에서는 0%로 좋은 성능을 보였으나 낮은 SNR 구간인 5 dB와 0 dB에서 각각 5.6% 9.5%의 성능을 보였다. 이와 같은 결과는 NOIZEUS Corpus의 car 잡음 환경이 백색잡음 형태의 특성을 띄고 있는 결과로 판단된다. 그 외의 steet와 airport의 잡음 형태는 babble 잡음의 특성을 보여 오보율 (False Alarm Rate)성능이 떨어지는 결과를 보인다. 그 이유는 입력신호의 음성에너지 자체가 매우 낮기 때문에 상대적으로 낮은 문턱값 설정으로 발생한 문제이나 N_{mask}값의 적절한 조정을 통해 개선할 수 있다. 그림 3은 주어진 프레임에서 음성 최대화를 실시한 이후에 평균과 표준 편차를 더한 값인 N_{mask}를 이용하여 음성과 비음성 구간을 표현하고 있다. 그림 3에서 볼 수 있듯이 N_{mask}값이 0근처에 머물면 비음성으로 0이상의 값에 존재하면 음성으로 구분할 수 있으며 낮은 SNR에서도 잡음 구간은 0근처를 유지하고 있고 낮은 SNR과 높은 SNR의 비음성구간이 0근처에서 일치하여 문턱값 설정이 용이한 것을 볼 수 있다. 따라서 현재의 문턱값인 N_{thresh}는 0.2로 설정되어 있으나 N_{mask}값을 기준으로 적절한 문턱값을 조정하면 낮은 SNR에서의 오보율 (False Alarm Rate)을 개선시킬 수 있을 것으로 본다.

V. 결 론

본 논문은 심리 음향 모델에 기반을 둔 바크 스케일 필터 뱅크를 이용하여 주어진 프레임에서 음성 에너지를 최대화 하고 음성 에너지가 최대화된 평균과 표준 편차 값을 이용하여 음성과 비음성 구간을 결정하여 음성구간을 검출하는 음성 에너지 최대화 음성검출 (Speech Energy Maximization-Voice Activity Detection) 방법을 제안하였다. 실험을 통해 제안된 음성검출 방법의 성능을 보면 낮은 SNR에서도 음성검출에 효과적임을 볼 수 있다. 제안된 방법은 음성 에너지를 최대화 하여 상대적으로 잡음 레벨을 저하시키는 방법을 사용하여 에너지를 기반으로 한 방법들 중 비교적 연산량이 적어 실제응용에 적합하다. 특히 무색잡음 특성을 갖는 자동차 잡음 환경에서 안정적이고 뛰어난 성능을 나타내며 유색잡음 환경에서도 음성 에너지만을 강화시키는 특징으로 비교적 좋은 성능을 가지는 것을 알 수 있었다. 제안된 음성 검출 방법은 음성 에너지가 중요한 특징으로 처리되는 음성인식 분야에서 인식률 향상에 기여할 수 있을 것으로 기대한다. 그리고 아직 낮은 SNR에서 오보율 (False Alarm Rate) 성능이 개선의 여지가 있으나 문턱값을 갱신하는 방법을 개선하면 보다 나은 성능을 가질 수 있을 것이다.

감사의 글

본 연구는 광운대학교 2009학년도 교내 학술 연구비 지원으로 이루어졌습니다.

참 고 문 헌

1. ETSI standard doc, ETSI ES 202 050 v1.1.1.
2. Rabiner, L. R. and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297–315, 1975.
3. Tuske, Zoltan and Mihajlik, Peter and Tobler, Zoltan and Fegyo, Tibor "Robust voice activity detection based on the entropy of noise-suppressed spectrum", in *Proc. of INTERSPEECH*, pp. 245–248, Sep. 2005.
4. E. Kosmides and E. Dermatas and G. Kokkinakis, "Stochastic endpoint detection in noisy speech", *SPECOM Workshop*, pp. 109–114, May. 1997.
5. S. Rangachari and P.C. Loizou "A noise-estimation algorithm for highly non-stationary environments", *Speech Communication*, vol 48, no 2, pp. 220 - .231, 2006
6. 김득수, "분산을 이용한 피치 및 유성음 구간 검출", *정보과학회 논문지*, 1권, 1호, 40 - 44쪽, 2004.
7. P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions', in *Proc. Eurospeech*, pp. 1887–1890, Sep. 2001.
8. E. Zwicker and H. Fastl, *Psycho-acoustics Facts and Models*, Springer-Verlag, Berlin, 1990.
9. David Kozel and Constantin Apostoaia, "Colored Noise Reduction Using Bark Scale Spectral Subtraction, Statistics, and Multiple Time Frames", in *Proc. IEEE International Conference Electro/Information Technology*, pp. 416–421, May, 2007.
10. Fletcher, "Auditory Patterns" *Re. Mod. Phys.*, Vol. 12, pp. 47–65, Jan. 1940.
11. *University of Texas Dallas Speech Copus NOIZEUS*, http://www.utdallas.edu/~loizou/speech/noizeus/, 2007.

저자 약력

•최 갑 근 (Gab-Keun Choi)

1999.2 : 광운대학교 정보과학원 졸업 (이학사)
2002.2 : 광운대학교 컴퓨터공학과 공학석사
2006.2 : 광운대학교 컴퓨터공학과 박사수료
2007.1~현재 : 에이팻 주식회사 IT 사업부장
※주관심분야 : 음성인식, 음성신호처리

•김 순 협 (Soon-Hyob Kim)

한국음향학회지 제27권 제7호 참조