음성의 주기성과 QSNR을 이용한 잡음환경에서의 음성검출 알고리즘

정주현, 송화전, 김형순
부산대학교 전자공학과

Voice Activity Detection Algorithm Using Speech Periodicity and QSNR in Noisy Environments

Ju Hyun Jeong, Hwa Jeon Song, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
{jeongju78, hwajeon, kimhs}@pusan.ac.kr

Abstract

Voice activity detection (VAD) is important in many areas of speech processing technology. Speech/nonspeech discrimination in noisy environments is a difficult task because the feature parameters used for the VAD are sensitive to the surrounding environments. Thus the VAD performance is severely degraded at low signal-to-noise ratios (SNRs). In this paper, a new VAD algorithm is proposed based on the degree of voicing and Quantile SNR (QSNR). These two feature parameters are more robust than other features such as energy and spectral entropy in noisy environments. The effectiveness of proposed algorithm is evaluated under the diverse noisy environments in the Aurora2 DB. According to out experiment, the proposed VAD outperforms the ETSI Advanced Frontend VAD.

Ⅰ. 서론

잡음환경에서의 음성검출(voice activity detection (VAD)) 알고리즘은 음성코딩, 음성인식, 음질향상과 같은 음성 응용분야에서 중요한 역활을 차지한다. 예를 들어 음성인식에서는 정확한 VAD를 통해 잡음환경에서의 음성인식성능을 향상시키고 음성구간에 대해서만 특징 파라미터를 추출함으로서 계산량을 줄일 수 있다. 또한 음성코딩에서는 음성이 아닌 부분을 무시함으로써 압축 효율을 높일 수 있다.

VAD에 널리 사용되는 특징 파라미터로는 에너지, cepstal distance[1], spectral entropy[2] 등이 있다. 에너지의 경우 신호 대 잡음비(SNR)가 낮은 환경에서는 성능이 떨어지며, cepstral distance와 spectral entropy와 같은 스펙트럼 영역에서의 특징 파라미터들은 비음성 구간의 스펙트럼이 음성의 스펙트럼과 다를 경우 신호 대 잡음비가 낮은 환경에서도 잘 동작하지만, 음성과 비슷한 스펙트럼 특성을 가지는 배경 잡음 환경에서는 성능이 크게 떨어진다.

음성의 특징인 주기성을 이용하면 음성과 비음성을 구분하가 위한 좋은 수단이 될 수 있다. 하지만 주변 잡음 역시 주기성을 가질 경우 음성을 판단하기 위한 파라미터로서 적합하지 않다. 반대로 에너지 기반의 파라미터는 잡음레벨에는 민감하지만 잡음의 종류에는 민감하지 않은 특성이 있다. 최근에 제안된 Quantile SNR(Quantile SNR)[3]은 에너지 기반의 특징 파라미터 중에서 비교적 좋은 성능을 나타낸다.

본 논문에서는 다양한 잡음환경에서 VAD가 동작하도록 하기 위해서 기존에 제안된 QSNR과 함께 음성의 주기성을 나타내는 파라미터로서 degree of voicing(DoV)을 사용했다. DoV를 구하기 위해서 피치 검출에 사용되는 YIN 알고리즘[4]을 이용하였다. 제안된 방법을 평가하기 위해 다양한 잡음환경을 포함하고 있는 Aurora2 DB를 사용하였으며 ETSI의 Advanced Frontend[5]에 포함된 VAD를 비교대상으로 삼았다.

본 논문의 구성은 다음과 같다. 2장에서는 음성 검출에 사용된 특징 파라미터들에 대해서 설명하고,

3장에서는 제안된 음성검출 알고리즘에 대해서 설명한다. 4장에서 실험 환경과 실험 결과에 대해서 살펴보고, 마지막 5장에서 결론을 맺는다.

Ⅱ. 음성 검출 특징 파라미터

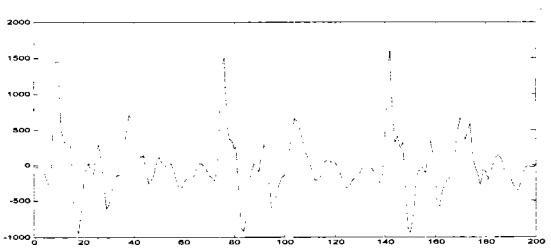
2.1 Degree of voicing

음성의 특성 중의 하나로 성대의 진동으로 인한 유성음의 주기성을 들 수 있다. 본 논문에서는 음성이 가지는 주기성을 이용하여 특징 파라미터를 추출하기 위해 피치 검출에 널리 사용되는 YIN 알고리즘을 이용하였다. YIN 알고리즘은 다음 식들로 표현될 수 있다.

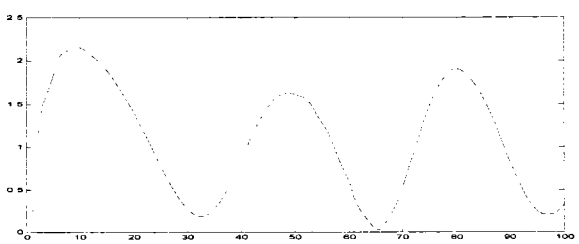
$$d_t(\tau)=\sum_{j=1}^{W}\left(x_t(j)-x_t(j+\tau)\right)^2 \tag{1}$$

$$D_t(\tau)=\frac{d_t(\tau)}{(1/\tau)\sum_{i=1}^{\tau}d_t(\tau)} \tag{2}$$

식 (1)에서 $d_t(\tau)$는 t번째 프레임의 difference function이다. x_n은 윈도우 안의 n번째 샘플이고, W는 윈도우 사이즈이며, τ는 lag를 가리킨다. 식 (2)에서 $D_t(\tau)$는 $d_t(\tau)$를 정규화한 것이다.



(a)유성음 파형



(b) $d_t'(\tau)$

그림 1. 음성 파형과 YIN 파라미터

그림 1에서 보는 바와 같이 유성음의 경우에는 $D_t(\tau)$가 피치간격마다 국부적인 최소값을 가지며, 주기성이 크면 클수록 그 값은 0에 가까워진다. 본 논문에서는 음성의 주기성 정도를 나타내는 파라미터로서 t번째 프레임의 DoV, V_t를 식 (3)과 같이 정의하였다.

$$V_t=\begin{cases}1.0-P_t/Th, & \text{if } P_t<Th\\ 0, & elsewhere\end{cases} \tag{3}$$

여기서 P_t는 음성의 주기성이 존재하는 범위 내에서 가장 작은 $D_t(\tau)$값이며, 이 값이 경계값 Th보다 클 경우에는 V_t를 0으로 둔다. V_t는 0과 1사이에 분포하며 음성의 주기성이 크면 1에 가까워지고, 주기성이 작으면 0에 가까워진다.

DoV는 SNR이 변하더라도 다른 특징 파라미터에 비해 상대적으로 강인한 특성을 보여준다. 하지만 Babble noise와 같은 주기성을 가지는 잡음이 더해지면 음성을 검출하는데 적합하지 못하다. 그래서 본 논문에서는 DoV와 에너지 기반의 파라미터인 함께 QSNR을 이용하였다.

2.2 QSNR

에너지를 기반으로 하는 파라미터는 잡음이 심한 환경에서는 성능이 떨어지나, 어느 정도의 SNR이 보장되는 환경에서는 높은 성능을 보여준다. 본 논문에서는 에너지 기반의 특징 파라미터 중에서 비교적 높은 성능을 보여주는 QSNR을 사용하였다. QSNR은 음성 로그 에너지의 분위수 정보를 이용하는 것으로 구하는 과정은 다음과 같다.

$$QSNR_t = Q_t(0.8)-N_t \tag{4}$$

$$N_t=\alpha N_{t-1}+(1-\alpha)Q_t(0.5) \tag{5}$$

식 (5)에서 $Q_t(0.8)$은 현재 프레임을 기준으로 앞과 뒤에 있는 21개의 프레임의 에너지 값들을 크기순으로 정렬했을 때 상위 80%에 해당되는 값이다. N_t는 배경잡음 레벨이며 시작점의 잡음레벨은 초기 20프레임의 에너지의 평균을 사용한다. 음성 검출 결과 음성이 아닌 것으로 판단되면 식 (5)와 같이 잡음레벨을 갱신할 수 있다. 식 (5)에서 $Q_t(0.5)$는 현재 프레임을 기준으로 앞과 뒤에 있는 21개의 프레임의 로그 에너지 값들 중에서 상위 50%에 해당되는 값으로 median 값을 의미한다.

III. 제안된 VAD 알고리즘

DoV의 경우에는 무성음구간과 짧은 휴지 구간에서 주기성을 찾을 수 없기 때문에 앞과 뒤의 10 프레임에 대한 moving average를 구해서 사용하였다. Babble noise와 같이 주기성을 가지는 잡음이 섞인 경우에는 DoV에 영향을 미치므로, 우선 음성 데이터를 차단 주파수가 500Hz인 저대역 필터로 통과시켰다.

음성을 판단하기 위한 특징 파라미터는 식 (6)와 같이 QSNR과 DoV를 곱한 값을 사용하였다.

$$Feature_t = QSNR_t \times V_t \tag{6}$$

QSNR의 경우에는 Dov와 동일한 범위를 유지하기 위해 실시간으로 0과 1사이로 정규화 하는 과정이 필요하다. QSNR은 잡음 레벨이 커지면 범위가 줄어들고 잡음레벨이 작으면 범위가 커진다. 그래서 식 (7)과 같이 잡음 레벨이 높을 경우에는 정규화 값을 낮추고 잡음레벨이 낮을 경우에는 정규화 값을 높이도록 하였다

$$T=\begin{cases}\dfrac{(T_0-T_1)}{(E_1-E_0)}\times(E-E_0)+T_1 & \text{if } E_0 < E \le E_1 \\ T_1 & \text{if } E < E_0 \\ T_0 & \text{if } E \ge E_1\end{cases} \tag{7}$$

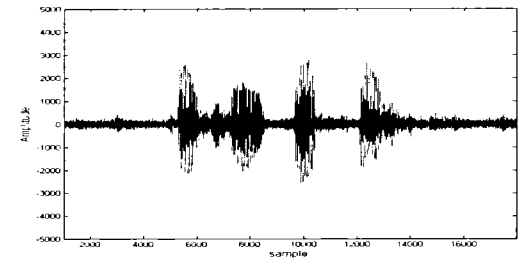
여기서 E_1와 E_0는 각각 SNR이 0dB과 20dB일 때의 잡음 레벨이고, E은 처음 20 프레임에서 구한 잡음 레벨이다. T_1와 T_0는 정규화 하기 위한 값의 범위이며, 이들 값들은 실험적으로 결정하였다. 식 (7)에서 구한 값을 이용하여 QSNR을 다음과 같이 정규화한다.

$$NQSNR_t = \frac{QSNR_t}{T} \tag{8}$$

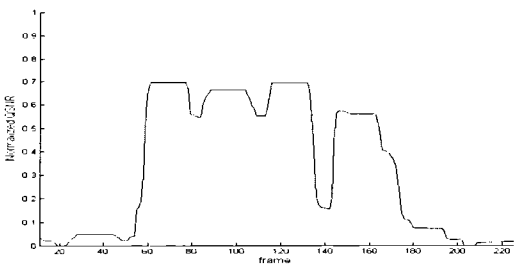
실시간으로 정규화를 하기 때문에 (8)식의 최대값이 1.0을 넘는 경우가 발생하는데 QSNR이 큰 값은 음성일 가능성이 높으므로 최대값을 1.0으로 제한하였다.

$$NQSNR_t = 1.0 \quad \text{if } NQSNR_t > 1.0 \tag{9}$$

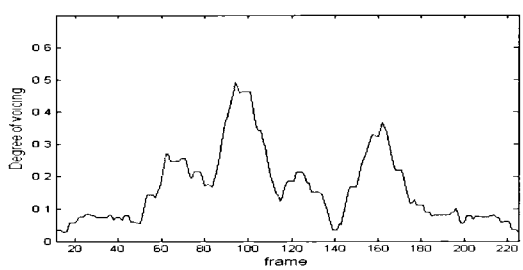
그림 2는 subway noise가 섞인 SNR이 10dB인 잡음음성에 대한 두 특징 파라미터를 보여주는 것으로서, 두 특징 파라미터를 서로 곱해서 사용하는 것이 음성을 판단하는데 있어 적합하다는 것을 보여준다.



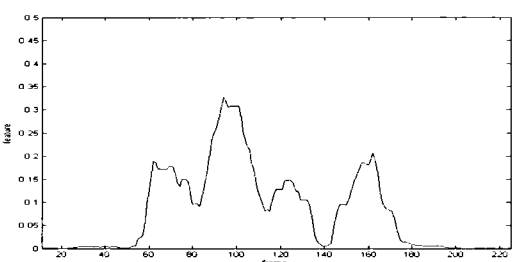
(a) 음성 과형



(b)정규화 된 QSNR



(c) Degree of voicing



(d) 제안된 특징 파라미터

그림 2. 음성 과형과 특징 파라미터

IV. 실험 결과

4.1 실험환경

실험에서 QSNR을 정규화 하기 필요한 T_1와 T_0은 각각 30과 10을 사용했으며, E_1와 E_0은 80과 50을 사용했다. 잡음 추정에 사용되는 α는 0.97을 사용했고, DoV를 구하는데 사용된 Th는 0.3을 사용하였다. 제안된 방법의 분석을 위해서 Aurora2 데이터베이스[6]가 사용되었다. Aurora2 데이터베이스는 1자리에서 7자리까지의 영어 연결숫자로 구성된 TI Digit에 다양한 잡음을 인위적으로 더한 것이다. 잡음환경은 8가지의 잡

음종류(subway, babble, car, exhibition, restaurant, street, airport, station)와 각각 7가지 잡음레벨(clean, 20dB, 15dB, 10dB, 5dB 0dB -5dB)로 구성되어 있으며, 훈련 데이터와 테스트 데이터로 구분되어 있다. 본 실험에서는 테스트 데이터 음성들을 사용하였다. 음성검출 결과를 비교하기 위해서 Aurora2 DB중에서 잡음이 섞이지 않은 clean speech에 대해서 HTK로 forced alignment된 결과를 사용하였고 제안된 알고리즘에 추가적인 hangover를 적용하였다.

4.2 실험결과

실험결과에 대한 성능평가 수단으로 [7]에서 사용된 receiver operating characteristic(ROC) 곡선을 이용하였다. ROC 곡선은 문턱값의 변경에 따른 false alarm rate(FAR)와 pause hit rate(PHR)의 변화를 보여준다. 그림 3에서 보는 바와 FAR가 증가하면 PHR가 감소하고, FAR가 감소하면 PHR이 증가하는 관계를 가지고 있다. 제안된 알고리즘이 FAR이 비슷한 경우에 Advanced Frontend의 VAD 보다 PHR관점에서 8% 더 좋은 성능을 보여줌을 확인할 수 있다.

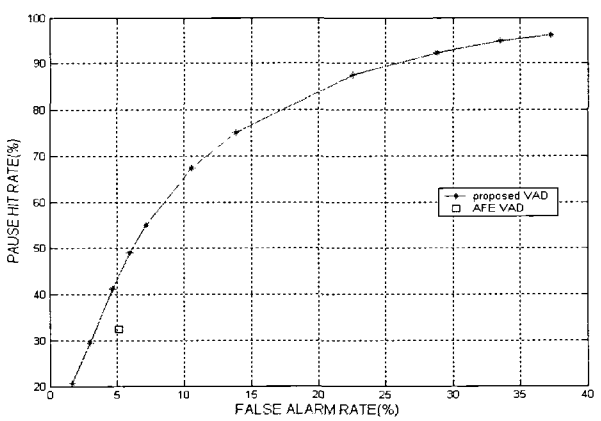


그림 3. ROC 곡선

V. 결론

본 연구에서는 잡음환경에서의 음성검출을 위한 알고리즘으로 QSNR과 degree of voicing을 결합하여 사용하는 방법을 제안하였다. 이 알고리즘은 별도의 훈련과정 없이 실시간으로 처리가 가능하다는 장점이 있다. 다양한 잡음환경을 가지고 있는 Aurora 2 데이터베이스에 대해 실험을 해본 결과 Advanced Frontend에 있는 VAD보다 좋은 성능을 보였다. 향후 다른 VAD 알고리즘들과 추가적인 성능비교를 하는 것이 필요하며, 제안된 VAD를 Advanced Frontend의 인식 시스템에 적용하여 인식성능을 살펴볼 계획이다.

이 논문은 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다.

참고문헌

[1] J. A. Haigh and J. S. Manson, "Robust voice activity detection using cepstarl feature," in Proc. IEEE TEN-CON, pp. 321-324, 1993.

[2] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in Proc. ICSLP, 1998.

[3] J. C. Segura, C. Benitez, A. de la Torre, A. Rubio, "Feature Extraction Combining Spectral Noise Reduction and Cepstral Histogram Equalization for Roubst ASR," in Proc. ICSLP, vol. 1, Sep, pp. 225-228, Sep, 2002.

[4] A. Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," JASA, vol. 111, no. 4, April, 2002.

[5] ETSI, "Speech processing, transmission and quality aspects(STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI ES 201 108 Recommendation, 2002.

[6] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," ISCA ITRW ASR2000, Paris, Sep. 2000.

[7] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre and A. Rubio, "A new voice activity detector using subband order-statistics filters for robust speech recognition," in Proc. ICASSP, May, 2004.