잡음 환경에서의 음성 검출 알고리즘 비교 연구

양 경 철 , 육 동 석*
고려대학교 컴퓨터학과

A Comparative Study of Voice Activity Detection Algorithms in Adverse Environments

Kyongchul Yang, Dongsuk Yook
Department of Computer Science and Engineering, Korea University
E-mail : yankc@voice.korea.ac.kr, yook@voice.korea.ac.kr

Abstract

As the speech recognition systems are used in many emerging applications, robust performance of speech recognition systems under extremely noisy conditions become more important. The voice activity detection (VAD) has been taken into account as one of the important factors for robust speech recognition. In this paper, we investigate conventional VAD algorithms and analyze the weak and the strong points of each algorithm.

I. 서론

음성 인식 시스템이 실생활에 적용되는 분야가 넓어지면서 음성 인식 성능이 유무선 전화망과 같이 채널의 왜곡을 받거나 자동차 환경과 같은 잡음 환경에 놓여도 좋은 성능이 나타나기를 바라는 요구가 늘어나고 있다. 특히, 주변 잡음 정도가 높아 signal-to-noise ratio (SNR)가 낮은 환경에서도 좋은 성능을 보이는 음성 인식 시스템의 필요성이 크게 부각되고 있다. 음성 검출은 잡음 환경에서의 음성 인식 시스템 성능에 큰 영향을 미치는 것으로 알려져 있으며, 현재까지 지속적인 연구가 되어 오고 있다.

음성 검출 알고리즘으로 가장 많이 알려져 있는 알고리즘은 에너지와 영교차율을 이용한 알고리즘으로 SNR이 낮지 않은 환경에서 간단한 수식을 이용하여 적용이 가능해 여러 응용 분야에 사용되었다[1][2]. ITU-T G.729B에서는 line spectral frquency, 전체 주파수 대역의 에너지, 낮은 주파수 대역 에너지, 그리고 영교차율을 이용한 음성 검출 표준화 알고리즘 안을 내놓았으며, 전화망에서 묵음 구간을 효과적으로 데이터를 압축하는데 사용되었다[3].

이후 SNR이 낮은 다양한 잡음 환경에서 효과적인 음성 검출을 하기 위한 여러 이론들이 제안되었다. Wilpon과 Rabiner는 HMM 모델 기반의 접근법을 제안하였다[4]. 숫자음 단어의 음성 검출 실험을 통해 그 성능을 입증한 이 논문에서는 미리 학습된 단어 HMM 모델과 잡음 HMM 모델을 이용하여 두 모델의 유사도를 비교하여 음성 검출을 한다. Gazor는 직교 변환한 도메인에서의 음성 데이터는 가우시안 모델보다 라플라시안 모델에 유사하다고 가정하여[5], 음성 가설과 묵음 가설의 likelihood ratio를 구해 음성 검출을 하였다[6]. [7]에서는 음성과 잡음이 주파수 대역에서 다르게 분포된다는 것을 이용 한다. 일부 주파수 대역을 가리는 band-partitioning을 한 후 엔트로피 값을 계산하여 음성 검출을 하도록 제안하였다. [8]에서는 non-stationary 잡음이 화자의 발성 중간에 발생할 경우 문턱값 조정이 필요하므로, geometrically adaptive energy threshold (GAET) 방법이 제안되었다. 음성의 검출을 음성의 활동 영역과 비활동 영역의 경계선을 찾

* 교신저자
본 연구는 한국과학재단 특정기초연구 (R01-2006- 000-11162-0) 지원으로 수행되었음.

아내는 것으로 이해하고, 영상의 경계선을 찾는데 쓰였던 Canny의 edge detector가 음성 검출에 응용이 되기도 하였다[9]. 가우시안 잡음으로부터 영향을 받지 않는 음성의 검출을 위해 higher order statistics (HOS) 방법이 제안되었고[10], 여기에 로그 kurtosis를 취하고 잡음 환경의 정도를 반영하여 잡음 환경의 변화에 덜 민감하도록 만들어 성능을 향상 시킨 방법이 제안되었다[11].

음성 검출 알고리즘은 에너지 기반에 영교차율을 고려한 전통적 기법과 SNR이 낮은 상황에서 여러 다른 잡음 조건에 좋은 특성을 보이도록 하는 기법이 제안되었다. 본 논문에서는 지금까지 발전해 온 음성 검출 이론들을 재조명해봄으로써 음성 검출의 연구 방향에 대한 이해를 돕고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 특성 검출별 음성 검출 알고리즘의 장단점을 비교 분석하고, 3장에서는 모델별 음성 검출 알고리즘의 장단점을 비교 분석한 후 마지막으로 4장에서 결론을 맺는다.

II. 특징 검출별 음성 검출

이 장에서는 특징 검출별로 그 동안 알려져 있는 음성 검출 알고리즘들을 비교 분석 해본다.

2.1 에너지 및 영교차율 기반 알고리즘

가장 대표적으로 알려져 있는 방법은 에너지 기반에 영교차율을 고려한 음성 검출 기법이다. 이 방법은 비교적 수학적 계산이 간단하며, 계산양이 많지 않고, 음성의 기본적인 특징인 에너지와 주파수 성질을 잘 표현하는 장점이 있어 많은 음성 인식 관련 응용 분야에 적용이 되고 있다. Rabiner와 Sambur는 음성의 짧은 구간의 에너지 특성과 영교차율을 기반으로 음성 검출 알고리즘을 제안하고 30dB 이상의 실험 환경에서 그 성능을 검증해 보였다[1][2]. 입력 신호의 에너지 $E(n)$과 영교차율 $Z(n)$은 각각 식(1), (2)로 구한다.

$$E(n)=\sum_{i=0}^{M-1}|x(i+n)| \tag{1}$$

$$Z(n)=\sum_{i=0}^{M-1}[|sgn[x(i)]-sgn[x(i-1)]|] \tag{2}$$

여기서 $x(i)$는 입력 신호이며, M은 coding frame size이다.

에너지와 영교차율 기반의 알고리즘을 이용한 음성 검출 방법은 잡음 정도가 높은 환경에서 에너지가 작고 영교차율의 특성이 잡음과 거의 유사한 파열음이나 마찰음과 같은 무성음은 찾아내기 어렵다.

2.2 Higher Order Statistics (HOS)

음성 신호의 higher order statistics를 계산하여 특별히 가우시안 잡음에서부터 음성 검출을 하는 방법이 제안되었다[10]. 음성을 삼각함수 모델로 표현 할 수 있다고 가정하고 짧은 순간의 LPC residual의 3차와 4차 cumulant 즉 skewness 식(3)과 kurtosis 식(4)를 구한다. 음성의 경우 에너지 E_s와 harmonics M으로 구성된다고 볼 수 있다. 에너지의 영향을 없애기 위해 에너지 값을 기준으로 정규화하여 비율을 구하면 식(5)를 얻고 이를 기준으로 음성과 가우시안 잡음을 판단해 낸다.

Skewness

$$\gamma_3^{(s)}=\frac{2}{2\sqrt{3}}(E_s)^{-\frac{3}{2}}\left[\frac{M-1}{M}\right], \tag{3}$$

Kurtosis

$$\gamma_4^{(s)}=E_s^2\left[\frac{4}{3}M-4+\frac{7}{6M}\right], \tag{4}$$

여기서 E_s는 LPC residual의 에너지이며, M은 harmonic의 개수이다.

Skewness-to-Kurtosis Ratio (SKR)

SKR 값은 Skewness와 Kurtosis의 에너지 값인 $(E_s)^{1.5}$와 $(E_s)^2$로 정규화한후 그 비율을 구한다.

$$SKR=\frac{(\gamma_3)^2}{(\gamma_4)^{1.5}}=\frac{9(M-1)^2}{8M\left[\frac{4M}{3}-4+\frac{7}{6M}\right]^{1.5}} \tag{5}$$

Log-Kurtosis (LK)

Li는 [11]에서 normalize된 로그 kurtosis 또는 로그 skewness 하나만을 사용하도록 제안한다. 즉, SKR을 쓰는 것 대신 입력 신호의 에너지 영향을 적절히 줄여주는 normalized kurtosis (skewness) 함수에 로그를 취하여 향상된 식 (6)과 같이 제안한다.

$$10\log_{10}\left(\gamma_4^{(x)}\right)=20\log_{10}E_s+10\log_{10}\left[\frac{4}{3}M-4+\frac{7}{6M}\right] \tag{6}$$

로그 kurtosis (또는 skewness)는 하나만 계산하므로 계산양이 줄고 입력 신호의 에너지가 kurtosis 함수 에 의해 normalize되므로 덜 민감한 효과를 얻는다.

하지만, HOS 알고리즘은 단독시스템으로 사용할 수가 없어 기존의 에너지 기반 알고리즘 등과 혼합하여 사용하여야 하며, 또한 non-stationary 잡음에 좋지 않은 성능을 보일 수밖에 없다는 단점이 있다.

2.3 Spectral Entropy

음성과 잡음의 주파수 대역에서의 데이터 분포의 형태가 다르다는 점을 이용하여 그 엔트로피를 계산하여 음성을 잡음으로부터 구분해 내는 방법이 제안 되었다[7]. 이 논문에서는 음성과 잡음의 엔트로피가 다르게 나타나도록 하기 위해, 주파수 대역에서 일정한 간격을 마스킹 하는 band-partitioning한 후 엔트로피를 구하였다. 그러나, 이 방법은 아직 cross-talk, 또는 음악 소리와 같은 넓은 대역에 에너지가 분포하는 경우 음성 검출을 하기 어렵다는 단점이 있다.

III. 모델 기준으로 본 음성 검출

이 장에서는 음성 검출의 판단 기준으로 사용되는 모델별로 알고리즘들을 비교 분석한다.

3.1 에너지와 영교차율의 고정된 문턱값

[1]에서는 음성 비활동 구간 동안에 SNR에 따라 미리 정한 문턱값을 기준으로 음성 검출의 판단을 내렸다. 또한, SNR이 다소 낮은 환경에서는 에너지만으로 음성 검출에 부족하므로, 모음의 앞뒤에 있을 수 있는 무성음의 영교차율을 고려하도록 하였다.

3.2 GAET

Non-stationary 잡음과 같이 세기가 일정치 않은 잡음이 음성 활동 구간에도 발생할 수 있다. 그러므로 음성 활동 구간에도 잡음의 크기에 따라 문턱값을 적응시켜야 할 필요가 있다.

[8]에서는 GAET 방법이 제안되어 입력의 크기에 따라 음성 판단과 잡음 판단의 문턱값을 적응시켰다.

3.3 Hidden Markov Model을 이용하는 방법

Wilpon과 Rabiner는 hidden Markov model (HMM)에 기반을 둔 음성 검출 방법을 제안하였다[4]. 제안된 알고리즘은 학습을 통하여 단어 모델과 잡음 모델을 참조 모델로 준비하고, 입력 데이터가 들어오면 각각의 모델과의 유사도를 비교하여 단어의 끝점을 찾아내도록 구현하였고, 숫자음의 음성 검출 실험에서 좋은 성능을 보였다. HMM을 기반으로 한 음성 검출 개념도는 그림 1과 같다. 음성 신호 $s(n)$이 입력으로 들어오면 linear prediction coefficient (LPC) 분석과 cepstral 변환을 통해 특징 벡터를 추출하고, 이 특징 벡터를 이용하여 미리 학습된 단어 HMM과 잡음 HMM 각각의 유사도를 측정하여 더 높은 유사도를 갖는 모델이 선택되게 함으로써 음성 검출을 한다. 이때 음성과 잡음의 두 개의 모델을 사용할 수도 있다.

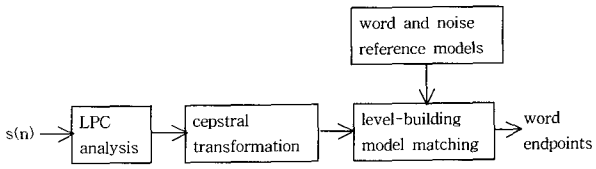

그림 1. HMM 기반 음성 검출 알고리즘

HMM을 이용한 음성 검출 알고리즘은 입력 신호를 파라미터화하여 사용하므로, 검출하고자 하는 신호와 주변 잡음에 대한 더 많은 정보를 이용하는 장점이 있다. 하지만, 이와 같은 통계적 모델을 이용한 음성 검출 방법은 학습 데이터와 실험 데이터의 환경이 크게 다른 경우 음성 검출 오류가 많이 발생하게 된다.

실제 시스템에 적용 할 경우 항상 학습 데이터와 실험 데이터가 유사성을 유지하기 어렵고, 또한 모든 종류의 주변 잡음을 정확히 모델링하고 있기 어려워 실제 시스템에 적용하기 어려운 문제가 있다. 또한, 계산량이 많아 실시간 처리 시스템과 같은 응용 분야에 적용하기 어려운 단점이 있다.

3.4 Laplacian-Gaussian 모델을 이용하는 기법

Gazor는 입력된 음성과 잡음이 discrete cosine trasform (DCT) 또는 Karhunen-Loève trasform (KLT)과 같은 직교 변환을 통해 decorrelated한 도메인으로 치환 했을 때, 음성 데이터는 라플라시안 확률 분포 함수로 잡음은 가우시안 확률 분포 함수로 표현 할 때 그 유사도가 높다고 증명하였다[10].

잡음의 확률분포

DCT 도메인으로 변환된 잡음 콤포넌트 $\{z_i(m)\}_{i-1}^{K}$가 가우시안이라고 하면, $z_i(m)$의 확률 분포 함수는 식(7)과 같다.

$$f_{z_i}(z_i(m))=\frac{1}{\sqrt{2\pi\sigma_i^2(m)}}e^{-\frac{z_i^2(m)}{2\pi\sigma_i^2(m)}},\ \forall i=1,2,\cdots K \tag{7}$$

여기서 $\sigma_i^2(m)$은 i번째 콤포넌트의 노이즈 분산이다.

음성의 확률분포

DCT 도메인으로 변환된 음성 콤포넌트 $\{s_i(m)\}_{i-1}^{K}$가

라플라시안이라고 하면, 음성의 확률 분포 함수는 식(8)과 같다.

$$f_{s_i}(s_i(m))=\frac{1}{2a_i(m)}e^{-\frac{|s_i(m)|}{2a_i(m)}},\ \forall i=1,2,\cdots K \quad (8)$$

$a_i(m)$은 깨끗한 음성의 i번째 eigenvalue의 라플라시안 factor이다.

Likelyhood Ratio Test (LRT)

음성구간과 묵음구간을 식(9)와 같이 가설을 세운다.

$$\begin{cases}\text{묵음구간, } H_0\text{: } V(m)=Z(m)\\ \text{음성구간, } H_1\text{: } V(m)=S(m)+Z(m)\end{cases} \quad (9)$$

여기서, $V(m)$, $S(m)$,과 $Z(m)$은 각각 잡음이 섞인 음성 $v_i(m)$, 음성 $s_i(m)$, 과 잡음 $z_i(m)$의 K-dimension 벡터이다.

입력이 주어지면 식(10)을 이용해 우도의 비율을 구해 그 비율에 따라 음성 구간인지 묵음 구간인지 판단한다.

$$L(m)=\frac{f_{V|H_1}(V(m))}{f_{V|H_0}(V(m))}, \quad (10)$$

[6]에서는 음성의 검출 여부에 따라 음성과 잡음 확률 분포 함수의 파라미터를 각각 적응 시키고 적응된 파라미터를 기준으로, 상대적으로 적은 파라메타를 사용한다는 장점이 있다. 또한, 음성의 경우 가우시안보다 라플라시안 모델을 사용하므로 유사도를 높였다고 볼 수 있다. 하지만, 음성의 끝 부분에 무성음이 존재 하여 SNR이 급격하게 작아지는 경우는 잡음으로 오판하게 된다. 그래서, HMM기반의 hangover기법을 사용하지만 상대적으로 false alarm의 비율이 높아지게 된다.

Ⅳ. 결론

본 논문에서 최근까지 알려져 있는 알고리즘을 사용하는 특징과 모델별로 나누어 비교 분석하였다. 지금까지 알려져 있는 특징 검출 방법들은 음성 검출을 위한 일반적인 판단 기준을 제공하고 있지 못 하고 있다. 따라서 새로운 특징 검출 기법뿐만 아니라 기존의 모델 기법들을 연동한 향상된 음성 검출 방법이 필요하다.

참고문헌

[1] L. R Rabiner and M. R. Sambur, "An algorithm for determining the endpoints for isolated utterances", *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297-315, Feb. 1975.

[2] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, 1993.

[3] ITU-T, A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70 Annex B, Nov. 1996.

[4] J. G. Wilpon and L. R. Rabiner, "Application of hidden Markov models to automatic speech endpoint detection", *Computer Speech and Language*, no. 2, pp. 321-341, Nov. 1987.

[5] S. Gazor and W. Zhang, "Speech Probability Distribution", *IEEE Signal Processing Letters*, vol. 10, no. 7, July 2003.

[6] S. Gazor and W. Zhang "A soft voice activity detector based on a Laplacian-Gaussian model", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498-505, Sept. 2003.

[7] B. F. Wu and K. C. Wang, "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no 5, pp. 762-775, Sept. 2005.

[8] S. G. Tanyer and H. Özer "Voice Activity Detection in Nonstationary Noise", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478-482, July 2000.

[9] M. Petrou and J. Kittler, "Optimal Edge Detectors for Ramp Edges", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 483-491, May. 1991.

[10] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217-231, Mar. 2001.

[11] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An Improved voice activity detection using higher order statistics", *IEEE Transactions on Speech and Audio Processing*, vol 13, no. 5, pp. 965-974, Sept. 2005.