강인한 핵심어 인식을 위해 유용한 주파수 대역을 이용한 음성 검출기

지 미 경, 김 회 린
한국정보통신대학교 공학부

Accurate Speech Detection based on Sub-band Selection for Robust Keyword Recognition

Mikyong Ji, Hoirin Kim
School of Engineering, Information and Communications University
E-mail : {lindaji,hrkim}@icu.ac.kr

Abstract

The speech detection is one of the important problems in real-time speech recognition. The accurate detection of speech boundaries is crucial to the performance of speech recognizer. In this paper, we propose a speech detector based on Mel-band selection through training. In order to show the excellence of the proposed algorithm, we compare it with a conventional one, so called, EPD-VAA (EndPoint Detector based on Voice Activity Detection). The proposed speech detector is trained in order to better extract keyword speech than other speech. EPD-VAA usually works well in high SNR but it doesn't work well any more in low SNR. But the proposed algorithm pre-selects useful bands through keyword training and decides the speech boundary according to the energy level of the sub-bands that is previously selected. The experimental result shows that the proposed algorithm outperforms the EPD-VAA.

I. 서론

음성인식에서 가장 중요한 문제 중 하나가 신호 대 잡음 비가 낮은 환경에서 음성 구간을 정확하게 검출해 내는 것이다. 즉 음성 구간을 정확히 검출해 내는 것이 음성인식에 큰 영향을 주기 때문이다. 본 논문에서는 신뢰성 있는 실시간 음성 검출 방법을 제안한다. 음성구간을 정확히 검출할 수 있다면, 인식기가 음성의 시작 전과 음성의 끝점 뒤의 묵음구간을 처리하는데 사용하였던 자원을 사용하여 인식 결과를 좀 더 정확하거나 혹은 응답시간을 더욱 빠르게 하는데 이용할 수 있다. 초기에 제안된 알고리즘의 대부분은 에너지와 영교차율에 기반하여 음성과 비 음성 부분을 구별하여 음성의 시작점과 끝점을 검출하는 방법을 사용하였다[1],[2]. 이러한 방법들은 구조가 간단하고 구현이 용이하다는 장점을 가지고 있다. 그러나 에너지와 영교차율에 기반한 방법들은 잡음의 영향에 상대적으로 민감하기 때문에 잡음 환경에 적용될 경우 그다지 좋은 성능을 보이지 못하였다. 에너지-영교차율 기반의 음성검출 방법을 개선하기 위한 노력으로 선형예측 오차 에너지, 피치, 지속시간 등과 같은 파라미터를 추가적으로 이용하는 방법이 제안되었다[3],[4],[5]. 파라미터 특성상 기존의 간단한 방법에 비해 많은 계산량을 필요로 하지만 잡음에 좀 더 강인하다. 최근에는 이와 다르게 시간과 주파수 영역에서 파라미터를 추출하고 이를 혼용하여 보다 정확한 음성 검출을 하고자하는 연구가 활발히 진행되어 왔다[6][7][8]. 본 논문에서는 핵심어 인식에서 가장 중요한 핵심어를 보다 정확히 검출하는 음성 검출 시스템을 제안한다. 사전 훈련을 통해 미리 유용한 주파수 대역을 선발하여 이를 기반으로 정확한 음성 검출을 한다.

II. 에너지-영교차율 기반의 음성 검출

이 알고리즘은 에너지와 영교차율을 파라미터로 이용하여 여러 규칙을 적용하여 음성구간을 검출하는 방법이다[6]. 음성의 시작과 끝을 검출하기 위해 음성을 마찰음류(fricative-like)와 모음류(vowel-like)인 두 가지로 분류하고 각각에 적합한 임계치를 설정한다. 구해진 임계치를 입력 음성에서 구한 값과 비교하여 음성을 검출한다. 이 알고리즘은 실시간 음성 검출이 가능하고 동작과정 또한 매우 빠르며 계산량도 적기 때문에 실제로 많이 쓰이나 신호 대 잡음비가 낮은 잡음 환경에서는 그다지 좋은 성능을 보이지 못한다.

III. 적응 시간-주파수 파라미터에 기반한 음성 검출

이 알고리즘은 시간-주파수 파라미터에 기반한 음성 검출 방법을 개선한 방법이다[7]. 주파수 대역에서 구한 에너지를 주파수 영역 파라미터로 정하고 주파수 대역 전체의 rms (root mean square) 에너지의 합으로 이루어진 것을 시간 영역의 파라미터로 하여 두 파라미터를 smoothing과 normalization 과정을 거친 다음 적당한 비율로 가산하고 smoothing하여 추출한 것이 ATF (Adaptive Time-Freuency) 파라미터이다. 여기서의 주파수 영역의 파라미터는 0~4000 Hz의 주파수 대를 20개의 멜 필터뱅크를 적용한 대역별 에너지를 가산한 것이다. 이 방법은 유용한 주파수 대역을 선정하기 위해 매 발성마다 주파수 영역의 파라미터를 사용하여 주파수 대역을 선정한 후 음성 검출을 하기 때문에 잡음의 영향에서 더 낳은 성능을 보이는 것으로 보고되고 있지만 실시간 음성 검출이 어렵다는 점과 이 과정에서의 과도한 계산량이 단점으로 지적된다.

IV. 유용한 주파수 대역을 이용한 음성 검출

제안된 음성 검출 알고리즘은 ATF 파라미터를 변형한 파라미터 즉 주파수 영역의 파라미터를 사용하고 일반적으로 많이 사용되는 EPD-VAA의 음성구간 검출 규칙을 적용한 음성 검출 방법이다. 특히 핵심어 인식에서 가장 중요한 핵심어를 보다 잘 검출하기 위해 사전 훈련을 통해 유용한 주파수 대역을 미리 선발하여 주파수 대역의 임계치를 비교하여 매 프레임을 음성과 비음성으로 구분하고 이를 토대로 음성의 시작점과 끝점을 결정한다. 즉 비 핵심어에 해당하는 구간보다는 핵심어에 해당하는 음성 구간을 더욱더 정확히 검출하고자 한다. 이 때 매 프레임을 음성 또는 비 음성으로 구분할 때, 유용한 주파수 대역마다 임계치와 비교하여 임계치보다 큰 값의 에너지를 가지는 대역이 일정 개수 이상이면 음성으로 분류한다.

이 장에서는 제안된 알고리즘에서 유용한 주파수 대역 선정 기법과 음성구간을 검출하는 방법에 대해 중점을 둔다.

4.1 핵심어에 대한 유용한 주파수 대역 선정기법

제안된 음성 검출 알고리즘에서의 유용한 주파수 대역 선정 기법은 그림 1과 같다.

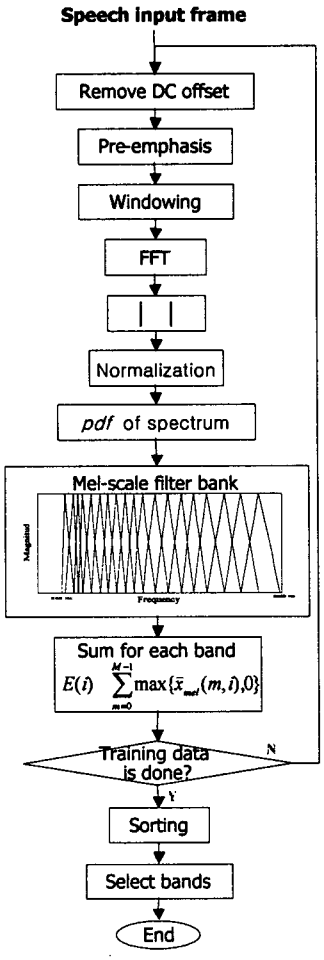


그림 1. 유용한 주파수 대역 선발을 위한 절차

입력 프레임의 DC bias를 제거하고 음성의 고주파 부분을 강조하기 위해 프리엠퍼시스를 적용하여 Hamming 윈도를 취한다. 그 다음 FFT 과정을 통해 스펙트럼을 구하고 음성의 앞부분에 묵음 구간으로 간

주되는 부분을 이용하여 normalization을 한다. 본 논문에서 사용되는 DB는 동시에 여러 종류의 마이크(HeadSet, NoteBook, Stand 마이크)를 사용하여 동시에 녹음된 음성이기 때문에 같은 핵심어 사이의 차이는 오직 에너지라 가정한다. 따라서 스펙트럼에서의 상대적인 값이 중요하기 때문에 전체 주파수 에너지에 의해 normalization 한다(Pdf of spectrum). 핵심어 DB를 이용하여 같은 과정을 반복하여 주파수 대역별 에너지를 모두 가산하여 순서를 정하고 큰 에너지의 주파수 대역을 선택한다.

4.2 음성 구간 결정 알고리즘

앞서 언급한 바와 같이 제안된 음성 검출 방법은 변형된 ATF 파라미터에 EPD-VAA의 음성 검출 규칙을 적용한 것이다. 훈련을 통해 유용한 주파수 대역을 미리 선발하고 실시간 음성 검출 시 음성의 첫 묵음 구간을 이용하여 미리 선발된 유용한 주파수 대역에 대한 임계치를 계산한다. 매 입력 프레임마다 유용한 주파수 대역의 대역별 에너지를 구하고 이를 임계치와 비교하여 프레임을 음성과 비음성으로 구별하고 음성이 일정기간 동안 계속되면 음성의 시작으로 간주한다. 이와 비슷하게 일정기간 동안 비음성이 지속되면 음성의 끝을 선언한다.

V. 실험 결과

5.1 DB

실험에 사용된 DB는 다양한 마이크(HeadSet, Notebook, Stand 마이크)를 사용하여 동시에 녹음되었다. 또 마이크와 입 사이의 거리(Stand 마이크 20cm, Stand 마이크 50cm, Stand 마이크 100cm)에 따라 총 5 종류의 DB를 사용하였다. 음성 검출의 테스트를 위해 사용된 DB의 구조는 그림 1과 같다. 이 중 핵심어에 해당하는 부분만을 잘라 유용한 주파수 대역 선정 시 사용하였다.

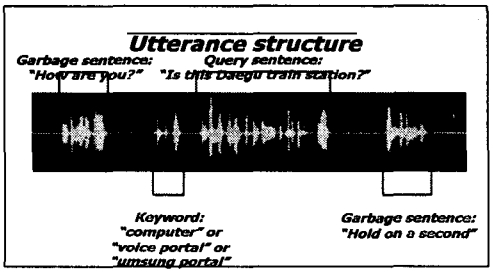
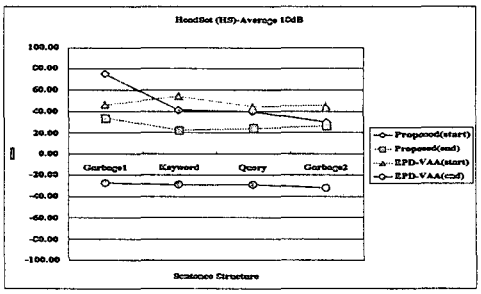


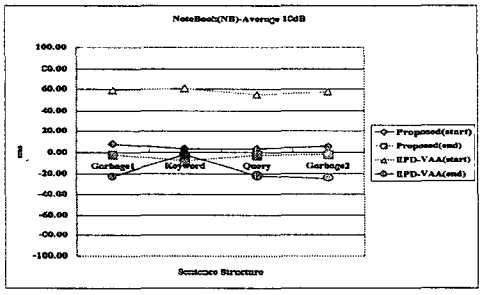
그림 2. 음성 검출에 사용된 DB의 구조

5.2 평가

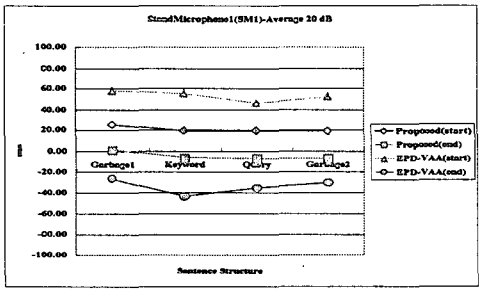
실험 결과를 통해 모든 DB 종류에서 제안된 알고리즘이 핵심어에 대해 보다 정확한 음성구간을 검출을 하고 있음을 알 수 있다. 비단 핵심어 뿐 아니라 비핵심어 음성에 대해 대체적으로 정확한 음성구간을 검출함을 보여준다. 특히 신호 대 잡음비가 낮은 경우 제안된 알고리즘이 EPD-VAA에 비해 월등한 성능을 보인다.



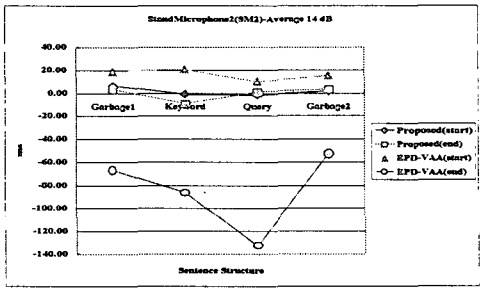
(a)



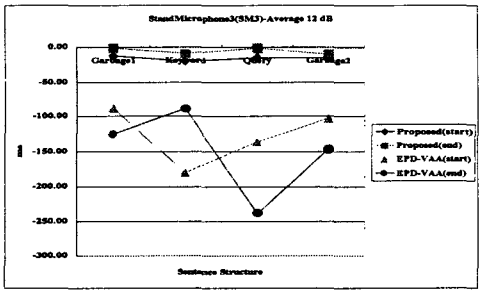
(b)



(c)

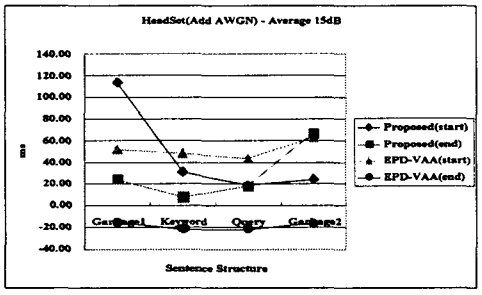


(d)

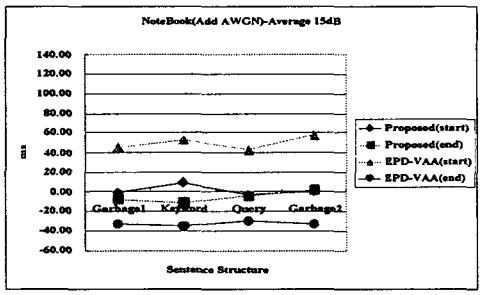


(e)

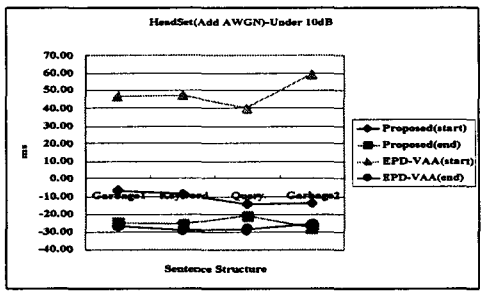
그림 3. 음성의 레이블 정보와 검출된 음성구간 사이의 평균거리



(a)



(b)



(c)

그림 4. 음성의 레이블 정보와 검출된 음성구간 사이의 평균거리(백색 가우시안 잡음을 더한 경우)

V. 결론

본 논문에서는 핵심어 인식을 위해 유용한 주파수 대역을 미리 선발하고 선발된 주파수 대역을 이용하여 핵심어 검출을 보다 정확히 하는 음성 검출기를 제안하고 이를 가장 일반적인 알고리즘(EPD-VAA)과 비교하였다. EPD-VAA 음성 검출 방법은 상대적으로 계산량이 적어 빠르고 구조가 간단하기 때문에 구현하기가 용이하기 때문에 실제로 많이 이용된다. 그러나 이 방법은 상대적으로 잡음의 영향에 민감하기 때문에 잡음 환경에서 좋은 성능을 유지하기 어렵다. 따라서 훈련을 통해 유용한 주파수 밴드를 미리 선정하여 핵심어를 보다 정확히 검출할 뿐 아니라 주파수 대역 선정을 통해 잡음에 영향을 덜 받는 음성 검출 방법을 제안하였고 실험 결과를 통해 더 나은 성능을 보였다.

참고문헌

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, pp. 130-134, Prentice-Hall, 1978.

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

[3] L. F. Ramel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, " An improved endpoint detector for isolated word recognition," IEEE ASSP Magazine, vol. 29, pp 777-785, Aug. 1981.

[4] C. Tsao and R. M. Gray, "An endpoint detector for LPC speech using residual error look-ahead for vector quantization applications," Proc. ICASSP-84,pp. 18b.7.1-4, 1984

[5] M. Hamada, Y. Takizawa, and T. Norimatsu, "A noise robust speech recognition system," Proc. ICSLP-90, pp. 893-896, 1990

[6] J. C. Junqa, B. Mark, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise, " IEEE Trans. Speech and Audio Processing, vol. 2, no. 3, pp. 406-412, July 1994

[7] G. D. Wu and C. T. Lin, "Word boundary detection with Mel-scale frequency bank in noisy environment," IEEE Trans. Speech and Audio Processing, vol. 8, no. 5, pp. 541-554, Sept. 2000.

[8] C. T. Lin, J-Ylin, and G-D Wu, "A roubust word boundary detection algorithm for variable noise-level environment in cars," IEEE Trans. Intelligent Transportation Systems, vol. 3, no. 1, pp. 89-101, March 2002.