# 이중채널 잡음음성인식을 위한 공간정보를 이용한 통계모델 기반 음성구간 검출

신민화 박지훈 김홍국
광주과학기술원 정보통신공학부
{mhshin, jh_park, hongkook}@gist.ac.kr

# Statistical Model-Based Voice Activity Detection Using Spatial Cues for Dual-Channel Noisy Speech Recognition

Min Hwa Shin Ji Hun Park Hong Kook Kim
School of Information and Communications, Gwangju Institute of Science and Technology

## 요약

본 논문에서는 잡음환경에서의 이중채널 음성인식을 위한 통계모델 기반 음성구간 검출 방법을 제안한다. 제안된 방법에서는 다채널 입력 신호로부터 얻어진 공간정보를 이용하여 음성 존재 및 부재 확률모델을 구하고 이를 통해 음성구간 검출을 행한다. 이때, 공간정보는 두 채널간의 상호 시간 차이와 상호 크기 차이로, 음성 존재 및 부재 확률은 가우시안 커널 밀도 기반의 확률모델로 표현된다. 그리고 음성구간은 각 시간 프레임 별 음성 존재 확률 대비 음성 부재 확률의 비를 추정하여 검출된다. 제안된 음성구간 검출 방법의 평가를 위해 검출된 구간만을 입력으로 하는 음성인식 성능을 측정한다. 실험결과, 제안된 공간정보를 이용하는 통계모델 기반의 음성구간 검출 방법이 주파수 에너지를 이용하는 통계모델 기반의 음성구간 검출 방법과 주파수 스펙트럼 밀도 기반 음성구간 검출 방법에 비해 각각 15.6%, 15.4%의 상대적 오인식률 개선을 보였다.

## 1. 서론

음성인식 시스템 내에서의 음성구간 검출의 역할은 잡음음성으로부터 음성구간을 검출함으로써 음성인식의 디코딩과정에서 불필요한 연산량을 감소시키고, 잡음으로 인한 오인식을 방지하는데 있다. 기존의 음성구간 검출 방법에서는 음성신호의 단구간 에너지, 영 교차율, 피치 구간, 스펙트럴 차 등의 특징 파라메터를 사용한다[1][2]. 이 파라메터들은 높은 신호 대 잡음비 (signal-to-noise ratio: SNR) 조건하에서는 음성의 존재 패턴을 반영하는데 매우 효과적이지만, 낮은 SNR 환경에서는 음성구간과 잡음구간에서의 파라메터 간의 변별력이 떨어진다. 이러한 문제를 해결하기 위해 최근에는 주파수 영역에서의 통계적 모델을 이용한 접근 방식이 제안되었다[3][4]. 특히, [4]에서는 Ephraim와 Malah에 의해 제안된 주파수 빈별 에너지 기반의 확률모델[5]을 이용하여 음성 존재 및 부재 확률을 추정하며, 추정된 확률의 검증을 통해 음성구간을 검출함으로써 음성구간 검출 정확성을 향상시켰다. 또한 Davis는Welch-Barlett 방법을 사용하여 보다 낮은 편차를 갖는 스펙트럼을 추정하고, 잡음의 통계적 특성을 고려한 문턱값을 설정 및 적응하는 방식을 제안함으로써 환경 변화에 따른 음성구간 검출 성능 저하를 개선하였다[4]. 상기 기술된 두 가지 방식 모두 주파수 영역에서의 통계적 정보를 기반으로 높은 SNR 환경 뿐만 아니라 낮은 SNR 환경에서도 안정적인 음성구간 검출이 가능하게 되었으나, 이는 정적잡음 환경에 국한된 경향을 보이고 있으며, 동적잡음 환경에서는 우수한 음성구간 검출 성능이 보장되지 못하는 단점을 지닌다.

본 논문에서는 동적 잡음환경에서의 음성인식을 위한 공간정보를 이용하여 음성 존재 확률모델을 추정하고 이를 이용하여 음성구간을 검출하는 방법을 제안한다. 즉, 제안된 방법은 이중채널 입력음성으로부터 획득한 공간정보를 이용하여 음성의 존재 및 부재에 대한 확률모델을 생성하며, 이를 이용하여 각 시간 프레임에 대한 음성구간 결정 파라메터로서 음성 존재 확률 대비 음성 부재 확률 비를 추정함으로써 음성구간을 검출한다.

## 2. 공간정보를 이용하는 통계모델 기반 음성구간 검출

그림 1은 제안된 음성구간 검출 방법의 구성도를 나타낸다. 제안된 음성구간 검출 방법은 우선 16 kHz의 표본화율로 얻어진 좌·우 채널 입력신호에 대해 사람의 청각특성을 반영한 감마톤 필터뱅크에 적용한다. 이때, 필터뱅크의 수는 32개이며 이를 통해 음성신호는 청각 주파수 신호로 변환된다. 이러한 시간-주파수 분석으로부터 각 프레임 및 주파수 밴드별로 좌·우 청각 주파수 신호간의 상호 시간 차이 (interaural time difference: ITD)와 상호 크기 차이 (interaural level difference: ILD)를 추출한다[6]. 추출된 $(i,j)$번째 시간-주파수 영역별 ITD와 ILD는 다음 식과 같이 음성 존재 대비 음성 부재 확률비 $\lambda(i,j)$를 추정하는데 적용된다.

$$\lambda(i,j) = \frac{P(t_{i,j}, l_{i,j}|\pi_{j,s})}{P(t_{i,j}, l_{i,j}|\pi_{j,s}) + P(t_{i,j}, l_{i,j}|\pi_{j,n})} \quad (1)$$

여기서 $P(t_{i,j}, l_{i,j}|\pi_{j,s})$는 주어진 $j$번째 주파수 밴드에 대한 음성존재 확률모델 $\pi_{j,s}$에 대한 $(i,j)$번째 시간-주파수 영역에서 추출된 ITD 값 $t_{i,j}$와 ILD 값 $l_{i,j}$를 특징벡터의 확률을 나타내며, $P(t_{i,j}, l_{i,j}|\pi_{j,n})$는 주어진 $j$번째 주파수 밴드의 음성부재의 확률모델 $\pi_{j,n}$에 대한 동일한 ITD와 ILD 확률을 의미한다. 본 논문에서 사용된 음성존재 및 음성부재 확률모델, $\pi_{j,s}$와 $\pi_{j,n}$는 공간 파라메터, 즉 ITD와 ILD를 특징벡터로 하며, 가우시안 커널 밀도 추정 방식[6]을 적용하여 학습된 확률모델로써, 마이크로폰 배열의 정면을 목표 음성 방향으로, 정면 이외의 방향을 잡음 방향으로 가정하여 학습된다.

제안된 방법은 음성구간 결정을 위해 문턱값 기반 비교 방식을 이용한다. 음성구간 검출을 위한 특징 파라메터로 식 (1)을 통해 구한 $i$번째 프레임의 32개 확률비의 평균 값 $\Lambda(i)$를 이용하여 입력 신호의 초기 10 프레임에서 식 (2)에서와 같이 음성 문턱값 $T_s$과 잡음 문턱값 $T_n$을 계산한다.

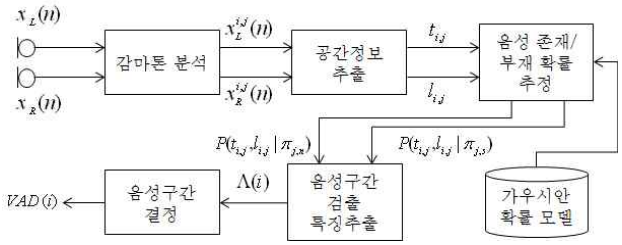


그림 1. 공간정보를 이용한 통계모델 기반 음성구간 검출 방법

$$T_c = \mu_n + \alpha_c \sigma_n \tag{2}$$

여기서 $c$는 $s$와 $n$으로 각각 표현될 수 있으며, $\mu_n$과 $\sigma_n$은 잡음 구간에서의 평균 확률 비들의 평균과 표준편차를 각각 나타내며, $\alpha_c$는 음성과 잡음에 대한 각각의 문턱값을 구하기 위한 결정 파라메터로서 $\alpha_s$는 5, $\alpha_n$은 1의 값으로 설정한다. 문턱값 $T_s$와 $T_n$은 다음 식에 적용되어 음성구간 검출을 위한 기준값으로 이용된다.

$$VAD(i) = \begin{cases} 1, & \Lambda(i) > T_s \\ 0, & \Lambda(i) < T_n \\ VAD(i-1), & otherwise \end{cases} \tag{3}$$

여기서 $VAD(i)$는 $i$번째 프레임에 대한 음성 존재 및 부재를 나타내는 1은 음성 존재를, 0은 음성 부재를 각각 의미한다.

초기 10 프레임 이후의 시간 프레임에서는 식 (3)의 규칙에 의해 음성의 존재 및 부재 여부를 결정한다. 특히, 음성 부재 프레임으로 결정되었을 경우 해당 프레임의 평균 확률 비 $\Lambda(i)$는 시간에 따라 변하는 잡음의 특성을 고려하기 위해 문턱값 $T_s$와 $T_n$의 적응에 활용된다.

## 3. 성능평가

이중채널 환경에서의 음성구간 검출 성능을 평가하기 위해 200개의 단어를 목표음성으로, 군중잡음과 경음악, 음성신호를 잡음원으로 사용하여 인위적으로 이중채널 잡음음성 데이터를 구축하였다[6]. 그림 2는 음성구간 검출 방법에 따른 오거부율(false rejection rate: FRR) 및 오보율(false alarm rate: FAR)과 음성인식을 통한 단어 오인식률(word error rate)을 보여준다. 여기서, FRR은 음성구간을 잡음구간으로 오인한 것을, FAR은 잡음구간을 음성구간으로 오인한 것을 의미한다. 그림 2(a)는 수동으로 검출한 음성구간을 참고로 하여 측정한 FRR과 FAR성능을 보여준다. 그림에서 보는 바와 같이, 제안된 방법의 FRR과 FAR이 기존의 주파수 에너지를 이용한 통계모델 기반 방법과 주파수 스펙트럼 밀도 기반 방법들에 비해 낮음을 알 수 있다.

그림 2(b)는 잡음환경에서의 음성인식을 통한 단어 오인식률을 나타낸다. 성능평가에 사용된 음성인식시스템은 특징벡터로 39차 MFCC를 이용하였으며, 음향 모델은 4개의 가우시안 혼합 밀도를 갖는 left-to-right hidden Markov model로 표현되었다. 또한 음성인식 시스템에는 2,250개의 단어가 유한 상태 네트워크의 문법으로 정의된 언어모델이 적용되었다. 그림 2(b)의 인식결과는 군중잡음, 음악잡음, 음성잡음에 대한 인식결과의 평균값이며, 특히 baseline 및 기존의 음성구간 검출 방법들은 이중채널 잡음음성 중 상대적으로 높은 SNR을 갖는 왼쪽 채널의 신호를 이용하여 성능을 비교하였다. 그림에서 보는 바와 같이, 제안된 방법의 단어 오인식률이 기존의 음성구간 검출 방법 보다 낮으며, 수동 음성구간 검출 방법의 성능과 가장 유사함을 알 수 있다. 결과적으로, 본 논문에서 제안한 공간정보를 이용한 통계모델 기반 음성구간 검출 방법이 주파수 빈별 에너지를 이용한 통계모델 기반의 방법과 주파수 스펙트럼 밀도 기반의 방법에 비해 평균적으로

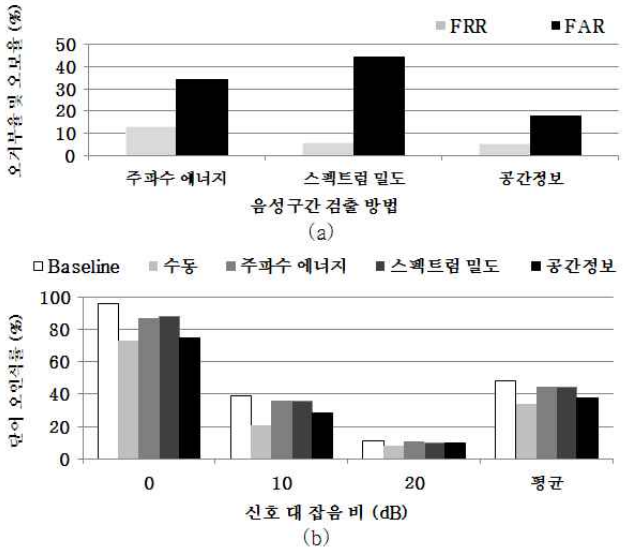


그림 2. 음성구간 검출 성능: (a) 오거부율(FRR) 및 오보율(FAR), (b) 단어 오인식률

각각 15.6.%와 15.4%의 오인식률의 개선을 보였다.

## 4. 결론

본 논문에서는 동적 잡음환경에서 음성인식 성능향상을 위한 통계모델 기반의 음성구간 검출 방법을 제안하였다. 제안된 방법은 이중채널 입력신호로부터 획득한 ITD와 ILD를 특징으로 하여 생성한 가우시안 확률모델을 기반으로 음성 존재 및 부재 확률을 추정하였다. 또한 음성구간 결정을 위한 특징 파라메터로 시간 프레임 별 음성 존재 대비 부재 확률 비를 추출하여 문턱값 비교를 통해 음성구간을 검출하였다. 제안된 음성구간 검출 방법은 음성구간 검출 오보율과 음성인식을 통한 단어 오인식률 측정을 통해 주파수 에너지를 이용한 통계모델 기반 방법 및 스펙트럼 밀도 기반 방법의 성능과 비교하였다. 비교 결과, 제안된 방법은 주파수 에너지를 이용한 통계모델 기반 방법과 주파수 스펙트럼 밀도 기반 방법에 비해 각각 15.6%와 15.4%만큼의 상대적 오인식률을 개선하였다.

## 감사의 글

본 논문은 2010년도 GIST "테라헤르츠 정보통신 융합기술 연구" 프로그램으로 수행된 연구임.

## 참고문헌

[1] 3GPP TS 26.194, *Adaptive multi-rate - Wideband (AMR-WB) speech codec: Voice activity detector(VAD)*, Dec. 2004.

[2] ETSI EN 301 708 v7.1.1, *Voice Activity Detector (VAD) for Adaptive Multi-rate (AMR) Speech Traffic Channels*, Dec. 1999.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[4] A. Davis, S. Nordholm, and R. Tognery, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412-424, Mar. 2006.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[6] J. H. Park, J. S. Yoon, and H. K. Kim, "HMM-based mask estimation for a speech recognition front-end using computational auditory scene analysis," *IEICE Trans. on Information and System*, vol. E91-D, no. 9, pp. 2360-2364, Sept. 2008.