말소리와 음성과학 제2권 제3호 (2010)
pp. 141~148

이중채널 잡음음성인식을 위한 공간정보를 이용한 통계모델 기반 음성구간 검출

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

신 민 화[1] · 박 지 훈[2] · 김 홍 국[3] · 이 연 우[4] · 이 성 로[5]
Shin, Min Hwa · Park, Ji Hun · Kim, Hong Kook · Lee, Yeonwoo · Lee, Seong Ro

ABSTRACT

In this paper, voice activity detection (VAD) for dual-channel noisy speech recognition is proposed in which spatial cues are employed. In the proposed method, a probability model for speech presence/absence is constructed using spatial cues obtained from dual-channel input signal, and a speech activity interval is detected through this probability model. In particular, spatial cues are composed of interaural time differences and interaural level differences of dual-channel speech signals, and the probability model for speech presence/absence is based on a Gaussian kernel density. In order to evaluate the performance of the proposed VAD method, speech recognition is performed for speech segments that only include speech intervals detected by the proposed VAD method. The performance of the proposed method is compared with those of several methods such as an SNR-based method, a direction of arrival (DOA) based method, and a phase vector based method. It is shown from the speech recognition experiments that the proposed method outperforms conventional methods by providing relative word error rates reductions of 11.68%, 41.92%, and 10.15% compared with SNR-based, DOA-based, and phase vector based method, respectively.

Keywords: voice activity detection (VAD), dual-channel speech, speech recognition, spatial cues

1. 서론

음성구간 검출(Voice Activity Detection, VAD)은 음성신호처리 시스템에서 입력 신호로부터 음성구간과 비음성구간을 검출하는 방법으로서, 다양한 음성 기반 알고리즘의 전처리 기법으로 적용되어 시스템 성능향상을 위한 중요한 역할을 수행한다. 예를 들면, 음성향상 기법에 적용되어 비음성구간으로부터 잡음의 통계 특성을 추정하고 활용함으로써 음성의 품질을 향상시킬 수 있으며 [1], 음성부호화에서는 비음성구간에 대해 낮은 비트율을 할당하는 불연속 전송을 통해 통신 채널의 효율적인 사용을 가능하게 한다 [2]. 또한, 음성인식 시스템 내에서는 잡음음성으로부터 음성구간을 검출함으로써 음성인식의 디코딩 과정에서의 불필요한 연산량을 감소시키고, 음성구간 전·후의 잡음으로 인한 오인식을 방지하는데 기여한다 [3].

일반적으로 음성구간 검출은 입력신호로부터 음성존재 및 부재 판별에 대한 특징 파라미터를 추출하는 단계와 이를 바탕으로 음성신호의 존재 여부를 판단하는 단계로 구성되며, 최근까지 음성구간 검출 성능 향상을 위해 다양한 특징 파라미터를 기반으로 한 음성구간 검출 기법들이 제안되었다. 이중에서도

1) 전자부품연구원 minhwas@gmail.com
2) 광주과학기술원 jh_park@gist.ac.kr
3) 광주과학기술원 hongkook@gist.ac.kr, 교신저자
4) 목포대학교 ylee@mokpo.ac.kr
5) 목포대학교 srlee@mokpo.ac.kr
이 논문은 2007년 정부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구이며 (KRF-2007-314-D00245) 또한 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2010-C1090-1021-0007).

접수일자: 2010년 8월 1일
수정일자: 2010년 9월 16일
게재결정: 2010년 9월 28일

음성구간 검출을 위한 가장 오래되고 대표적인 특징 파라미터로는 음성신호의 단구간 에너지(short-term energy)와 영교차율(Zero Crossing Rate, ZCR)을 들 수 있다 [4]. 단구간 에너지와 영교차율 기반의 음성구간 검출 방법은 적은 계산량으로 효율적인 음성구간 검출이 가능한 장점을 지닌 반면, 낮은 신호 대 잡음비(Signal-to-Noise Ratio, SNR) 환경에서는 음성구간과 잡음구간에서의 파라미터 간 변별력이 떨어지며 이로 인한 음성구간 검출 성능이 저하되는 단점을 보인다. 이를 극복하기 위해 대역 제한된 신호의 파워 또는 대역 통과 필터뱅크로부터의 출력신호들의 에너지의 합 [5], [6], 스펙트럼의 모양 [7], 잡음 스펙트럼의 특성 [8], 음성신호의 주기성 [9] 등을 특징 파라미터로 이용한 방법들이 제안되었다. 특히, "참고문헌[10]"에서는 Welch-Barlett 방법을 사용하여 보다 낮은 편차를 갖는 스펙트럼을 추정하고, 이를 통해 SNR을 측정하여 음성구간 검출 파라미터로 이용하고, 비음성구간에서의 통계적 특성을 고려한 문턱값을 설정 및 적응하는 방식을 제안함으로써 환경 변화에 따른 음성구간 검출 성능을 개선하였다. 그러나 이와 같은 성능은 정적잡음 환경에 국한된 경향을 보이고 있으며, 비정적 잡음환경에서는 우수한 음성구간 검출 성능이 보장되지 못하는 단점을 지닌다.

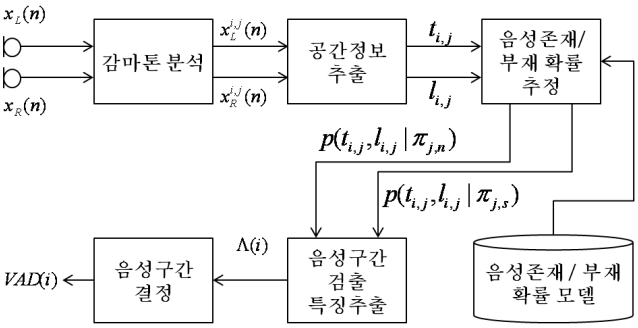

그림 1. 제안된 공간정보를 이용한 통계모델 기반 음성구간 검출 방법의 블록도

Figure 1. Block diagram of the proposed statistical model-based voice activity detection employing spatial cues

이와 관련하여 최근에는 정적잡음 환경뿐만 아니라 비정적 잡음 환경에서의 음성구간 검출 성능 향상을 위하여 다채널 입력신호를 이용한 음성구간 검출 방법들이 제안되었다. 예를 들어 이중 채널 입력으로부터 특정 시간 주파수 영역에서 신호도착방향(Direction of Arrival, DOA)의 엔트로피를 추정하고 이를 통하여 음성구간을 검출하는 방법이 제안되었다 [11]. 그러나 이 방법은 목표 음성신호로부터 추정된 신호도착방향은 균일하게 분포하며, 잡음신호로부터 추정된 신호도착방향은 균일하지 않게 분포하지 않는 다는 가정 하에 제안되었기 때문에 목표 음성신호와 마찬가지로 방향성이 뚜렷한 잡음신호가 존재하는 환경에서는 음성구간 검출 성능이 저하된다. 또 다른 예로써, 마이크로폰 배열에서의 위상 벡터(phase vector)를 이용하여 음성구간을 검출하는 방법이 제안되었다 [12]. 하지만 이 경우 또한 방향성을 지닌 잡음이 존재하는 환경에서는 안정적인 음성구간 검출이 가능하지만, 높은 성능을 위해서는 많은 수의 마이크로폰을 필요로 하는 단점이 있다.

본 논문에서는 기존의 음성구간 검출 방법의 단점을 극복하기 위해 공간정보기반의 음성구간 검출 방법을 제안한다. 즉, 제안된 방법은 이중채널 마이크로폰 신호로부터 음성구간과 비음성구간에서의 상호 시간 차이(Interaural Time Difference, ITD)와 상호 크기 차이(Interaural Level Difference, ILD)를 추출하여 각각 ITD, ILD 분포를 확률모델로 생성하여 이를 기반으로 음성구간을 검출한다.

본 논문의 구성은 다음과 같다. 서론에 이어, 2장에서는 공간정보를 이용한 음성구간 검출 방법을 제안하고, 3장에서는 기존의 방식과의 비교실험 결과를 보여준다. 마지막으로 4장에서는 본 논문의 결론을 맺는다.

2. 공간정보를 이용하는 통계모델 기반 음성구간 검출

<그림 1>은 본 논문에서 제안하는 공간정보를 이용한 음성구간 검출 방법의 구성도를 보여준다. 제안된 음성구간 검출 방법은 이중채널 음성을 입력으로 하며, 좌·우 입력채널간의 ITD와 ILD를 기반으로 음성존재 및 부재에 대한 확률모델을 생성하고, 이를 바탕으로 음성구간을 검출한다. 본 장에서는 공간정보를 이용한 통계모델 기반 음성구간 검출 방법의 각 단계에 대해 자세히 설명하도록 한다.

2.1 감마톤 분석

제안된 음성구간 검출 방법은 우선 16 kHz의 표본화율의 이중채널 잡음신호를 각각 사람의 청각특성을 반영한 감마톤 필터뱅크에 적용하여 32개의 청각 주파수 신호로 변환시킨다 [13]. 본 논문에서 사용된 감마톤 필터뱅크는 50 Hz에서 8 kHz 사이의 주파수를 균일 장방형 대역(equivalent rectangular bandwidth) 단위로 나눈 32개의 감마톤 필터로 구성된다 [14]. 이 때, 입력신호는 10 ms씩 중첩되는 20 ms 크기의 프레임으로 나누어 감마톤 필터뱅크에 적용되며, 이를 통해 각 시간-주파수 영역별 좌·우 청각 주파수 신호 $x_L^{i,j}$과 $x_R^{i,j}$을 얻는다. 이때, i는 시간 프레임 인덱스를, j는 주파수 밴드 인덱스를 각각 나타낸다.

2.2 공간정보 추출

음성의 존재 및 부재 확률 추정을 위한 파라미터로 각 시간 프레임 및 주파수 밴드별로 공간정보, 즉 ITD와 ILD를 구한다. 우선 ITD를 추출하기 위해서 감마톤 분석을 통해 얻어진 (i,j) 번째 시간-주파수 영역의 좌·우 청각 주파수 신호 $x_L^{i,j}$와 $x_R^{i,j}$

사이의 정규상호상관계수(normalized cross-correlation) $CC^{i,j}(\tau)$를 다음과 같이 구한다.

$$CC^{i,j}(\tau)=\frac{\sum_{n=0}^{N-1}x_L^{i,j}(n)x_R^{i,j}(n-\tau)}{\sqrt{\sum_{n=0}^{N-1}(x_L^{i,j}(n))^2}\sqrt{\sum_{n=0}^{N-1}(x_R^{i,j}(n))^2}} \quad (1)$$

여기서 τ는 정규상호상관계수의 시간 지연값으로 -8에서 8의 값을 가지며, 이는 16 kHz 표본화율에서 -0.5 ms부터 0.5 ms에 해당한다. 또한 N은 한 프레임 당 샘플 수를 나타내고, 동일한 표본화율 환경에서 20 ms에 해당하는 320 샘플의 값으로 설정된다. 다음으로, 식 (1)에서 구한 (i,j)번째 시간-주파수 영역의 $CC^{i,j}(\tau)$가 최대값을 가질 때의 지연값을 다음 식과 같이 구함으로써 해당영역의 ITD를 구한다.

$$t_{i,j}=\arg\max_{\tau}|CC^{i,j}(\tau)| \quad (2)$$

ILD의 경우 (i,j)번째 시간-주파수 영역의 좌·우 청각 주파수 신호 $x_L^{i,j}$과 $x_R^{i,j}$간의 에너지 차이를 이용하여 해당영역의 ILD 값, $l_{i,j}$을 다음 식과 같이 계산한다.

$$l_{i,j}=10\log_{10}\left[\frac{\sum_{n=0}^{N-1}(x_L^{i,j}(n))^2}{\sum_{n=0}^{N-1}(x_R^{i,j}(n))^2}\right] \quad (3)$$

2.3 음성구간 검출을 위한 파라미터

2.2절에서의 식 (2)와 식(3)을 통해 좌·우 청각 주파수 신호로부터 추출된 ITD, $t_{i,j}$와 ILD, $l_{i,j}$는 음성존재 및 부재에 대한 확률모델의 특징벡터로 입력되어 (i,j)번째 시간-주파수 영역의 음성존재 및 부재 확률을 추정하는데 이용된다. 특히 음성존재 및 부재에 대한 확률 모델은 각 주파수 대역별로 ITD와 ILD를 입력으로 가우시안 커널 밀도 추정(Gaussian kernel density estimation) [15]을 통해 학습된다. 이 때 마이크로폰 배열의 정면을 목표 음성 방향으로 가정하여 학습 데이터로부터 획득한 ITD와 ILD를 통해 음성존재에 대한 확률 모델을 생성한다. 마찬가지로, 정면 이외의 방향에 대한 학습 데이터로부터 획득된 ITD와 ILD을 통해 음성부재에 대한 확률모델, 즉 잡음모델을 학습한다.

음성구간 검출을 위한 특징 파라미터로는 다음과 같이 각 (i,j)번째 시간-주파수 영역별로 추정된 음성존재 대비 부재 확률의 비로 계산하고, 이 값을 전체 주파수 밴드에 대해 모두 합한 $\Lambda(i)$를 이용한다.

$$R_{i,j}=\frac{P(t_{i,j},l_{i,j}|\pi_{j,s})}{P(t_{i,j},l_{i,j}|\pi_{j,n})} \quad (4)$$

$$\Lambda(i)=\sum_{j=0}^{J-1}R_{i,j} \quad (5)$$

여기서 $P(t_{i,j},l_{i,j}|\pi_{j,s})$는 주어진 j번째 주파수 밴드에 대한 음성존재 확률모델 $\pi_{j,s}$에 대한 (i,j)번째 시간-주파수 영역에서 얻어진 ITD값 $t_{i,j}$와 ILD값 $l_{i,j}$로 구성된 벡터의 확률을 나타낸다. 또한 $P(t_{i,j},l_{i,j}|\pi_{j,n})$는 j번째 주파수 밴드의 음성부재의 확률모델 $\pi_{j,n}$에 대한 동일한 입력 벡터의 확률을 의미한다. 그리고 J는 감마톤 필터뱅크의 전체 주파수 밴드의 수로 본 논문에서는 32의 값을 가진다.

2.4 음성구간 결정 규칙

본 논문에서 제안된 음성구간 검출 방법은 음성구간 결정을 위해 문턱값 기반 비교 방식을 이용한다. 음성구간 검출을 위한 문턱값은 초기 I 개 프레임 이내의 입력 신호에 음성이 존재하지 않는다는 가정 하에, 이 구간에서의 특징 파라미터 $\Lambda(i)$의 분포를 통해 구한다. 즉, $\Lambda(i)$값에 대한 평균 μ_n과 표준편차 σ_n을 이용하여 다음과 같이 음성에 대한 문턱값 T_s와 잡음에 대한 문턱값 T_n을 계산한다.

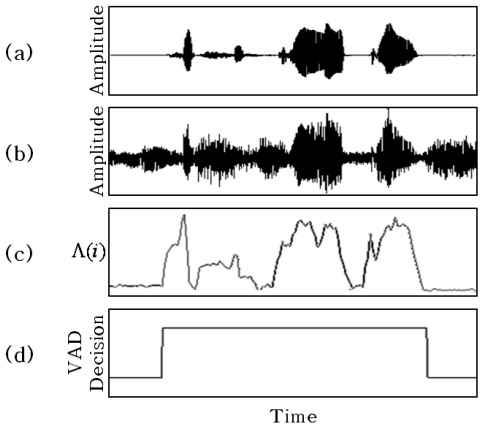

그림 2. 경음악 잡음이 포함된 잡음음성에 대한 음성구간 검출 결과; (a) 목표음성, (b) 잡음음성, (c) 음성구간 검출을 위한 특징 파라미터, (d) 음성구간 검출 결과

Figure 2. VAD results for noisy speech mixed with classic music; (a) target speech, (b) noisy speech, (c) VAD feature, and (d) VAD result

$$T_c=\mu_n+\alpha_c\sigma_n \quad (6)$$

여기서 c는 s와 n으로 각각 표현될 수 있으며, α_c는 음성과 잡음에 대한 각각의 문턱값을 구하기 위한 결정 파라미터로서, α_s는 5, α_n은 1의 값으로 설정한다. 또한 μ_n과 σ_n은 초기 I 개 프레임에 대해 각각 계산되며, 본 논문에서는 I=10으로 하여 10개의 프레임

에 대한 평균과 표준편차를 다음과 같이 각각 구한다.

$$\mu_n = \frac{1}{I}\sum_{i=1}^{I} \Lambda(i) \tag{7}$$

$$\sigma_n = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(\mu_n - \Lambda(i))^2} \tag{8}$$

다음으로는 식 (5)를 통해 구한 특징 파라미터 $\Lambda(i)$와 식 (6)을 통해 구해진 음성과 잡음에 대한 문턱값들, T_s와 T_n,을 이용하여 다음과 같은 규칙을 통해 음성구간을 결정한다.

$$VAD(i) = \begin{cases} 1, & \Lambda(i) > T_s \\ 0, & \Lambda(i) < T_n \\ VAD(i-1), & otherwise \end{cases} \tag{9}$$

여기서 $VAD(i)$는 i번째 프레임에 대한 음성존재 및 부재를 나타내며 1은 음성존재를, 0은 음성부재를 각각 의미한다. 초기화로 $VAD(0) = 0$ 으로 설정한다. 만약 i번째 프레임이 음성부재 프레임으로 결정되었을 경우, 파라미터 $\Lambda(i)$는 시간에 따라 변하는 잡음의 특성을 고려하기 위하여 문턱값 T_s와 T_n의 적응에 활용된다. 즉, 음성부재 프레임의 $\Lambda(i)$가 반영된 새로운 μ_n과 σ_n은 식 (10)-(12)와 같이 적응되며, 최종적으로 식 (6)을 통해 새로운 문턱값 T_s와 T_n을 추정한다.

$$\mu_n = \gamma \cdot \mu_n + (1-\gamma) \cdot \Lambda(i) \tag{10}$$

$$\mu_{n-squraed} = \gamma \cdot \mu_{n-squraed} + (1-\gamma) \cdot \Lambda(i)^2 \tag{11}$$

$$\sigma_n = \sqrt{\mu_{n-squared} - \mu_n^2} \tag{12}$$

여기서 $\mu_{n-squared}$는 비음성구간에서의 특징 파라미터 $\Lambda(i)$의 제곱의 평균값을 나타낸다. 또한 γ는 문턱값 적응 정도를 조정하기 위한 평활 계수(smoothing coefficient)로, 본 논문에서는 실험적인 방법을 통해 0.95로 설정하였다.

<그림 2>는 경음악 잡음이 혼재된 잡음음성에 대한 제안된 공간정보를 이용하는 통계모델 기반 음성구간 검출방법의 음성구간 검출 결과를 보여준다. <그림 2(a)>는 목표음성만을 나타내며, <그림 2(b)>는 경음악 잡음이 0 dB SNR로 혼합된 잡음음성을 보여준다. 또한, <그림 2(c)>는 <그림 2(b)>의 잡음음성을 입력으로 하여 추출한 특징 파라미터를, <그림 2(d)>는 식 (9)의 규칙을 통해 구한 최종 음성검출 결과를 각각 보여준다. <그림 2(a)>와 <그림 2(d)>의 비교를 통해 알 수 있듯이 제안된 음성구간 검출 방법은 실제 음성구간과 유사하게 음성구간을 검출함을 확인할 수 있다.

3. 성능 평가

본 장에서는 제안된 공간정보를 이용하는 통계모델 기반 음성구간 검출방법의 성능을 단채널 신호를 통해 추정한 SNR 기반의 음성구간 검출 방법 [10], 이중채널 환경에서의 동작하는 DOA의 엔트로피를 이용한 음성구간 검출 방법 [11] 및 이중채널 입력신호의 위상벡터 기반 음성구간 검출 방법 [12]의 성능과 각각 비교한다.

3.1 이중채널 잡음음성 데이터베이스

제안된 음성검출 방법의 성능 평가를 위해 ETRI 한국어 헤드셋 인식용 단어 데이터베이스 [16]를 사용하여 이중채널 잡음음성 데이터를 인위적으로 구축하였다. 200개의 단어음성에 대해서는 0°에 위치하게 하는 머리 전달 임펄스 응답(Head-Related Impulse Response, HRIR) [17]을 적용하고, 군중잡음(babble noise), 공장잡음(factory noise), 경음악(classic music), 그리고 목표음성이 아닌 다른 방향으로부터의 음성(interference speech)을 잡음신호로 사용하여, 20°, 30°, 40° 및 50°에 위치하도록 하는 HRIR를 적용하여 방향이 전환된 잡음신호를 더해 이중채널용 테스트 잡음음성 데이터베이스를 제작하였다. 이때 잡음은 0, 10 및 20 dB의 SNR을 갖도록 가공하였다.

음성존재 및 부재에 대한 확률 모델의 학습에는 성능 평가용 음성 데이터베이스에 사용되지 않은 96개의 단어음성을 목표음성을 사용하여 0°에 해당하는 HRIR을 적용하였다. 또한, 성능 평가용 데이터베이스와 동일한 네 종류의 96개 잡음신호들을 0, 10 및 20 dB의 SNR을 갖도록 가공한 후, 20°, 40°, 60° 및 80°에 해당하는 머리 전달함수를 적용하여 최종적으로 96개의 잡음음성 데이터를 제작하였다.

3.2 음성구간 검출 성능

우선 제안된 음성구간 검출 방법의 성능을 SNR 기반 방법, DOA 엔트로피 기반 방법, 위상벡터 기반 방법들의 성능과 오보율(False Alarm Rate, FAR) 및 오거부율(False Rejection Rate, FRR)을 측정하여 비교하였다. FAR과 FRR은 200개의 성능평가용 단어음성에 대해 수동으로 검출한 음성구간을 기준으로 각각 다음과 같이 측정된다.

$$FAR = \frac{\text{비음성구간이 음성구간으로 검출된 프레임 수}}{\text{전체비음성구간 프레임 수}} \tag{13}$$

$$FRR = \frac{\text{음성구간이 비음성구간으로 검출된 프레임 수}}{\text{전체음성구간 프레임 수}} \tag{14}$$

<그림 3>과 <그림 4>는 10 dB와 0 dB SNR의 잡음환경에 대한 각 음성구간 검출 방법들의 문턱값들을 조정하면서 FAR과 FRR을 측정하여 구한 동작특성(Receiver Operating Characteristic, ROC) 곡선을 각각 보여준다. 이 때, 위상벡터 기반의 음성구간

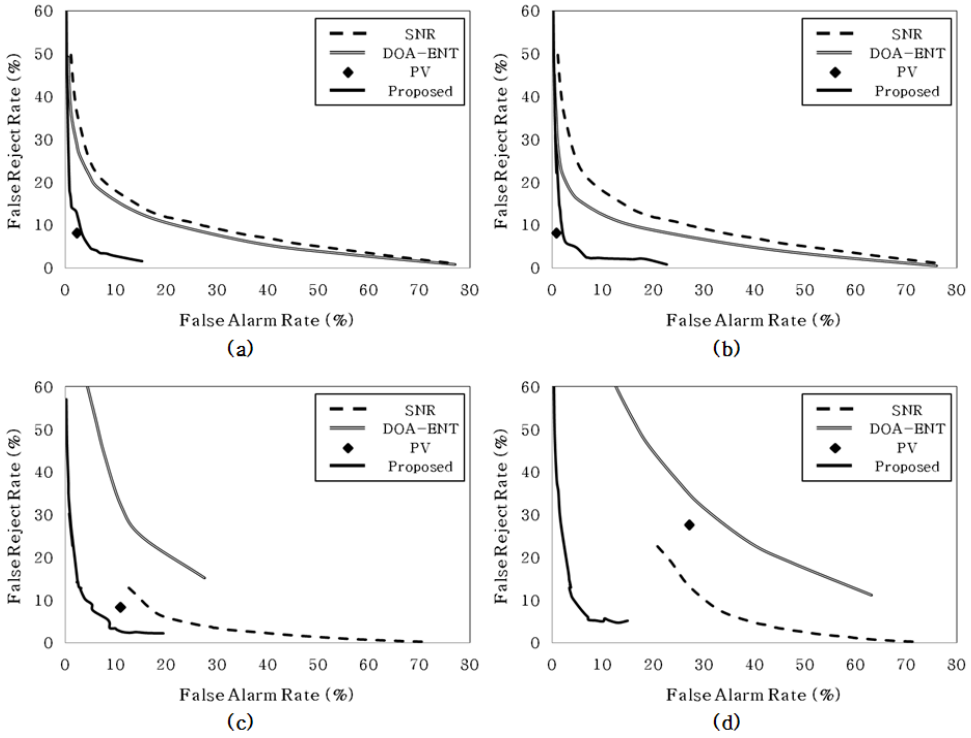

그림 3. 다양한 잡음 환경에서의 ROC 곡선 (10 dB SNR); (a) babble, (b) factory, (c) classic music, (d) interference speech
Figure 3. ROC curves for various noise environments (10 dB SNR); (a) babble, (b) factory, (c) classic music, and (d) interference speech

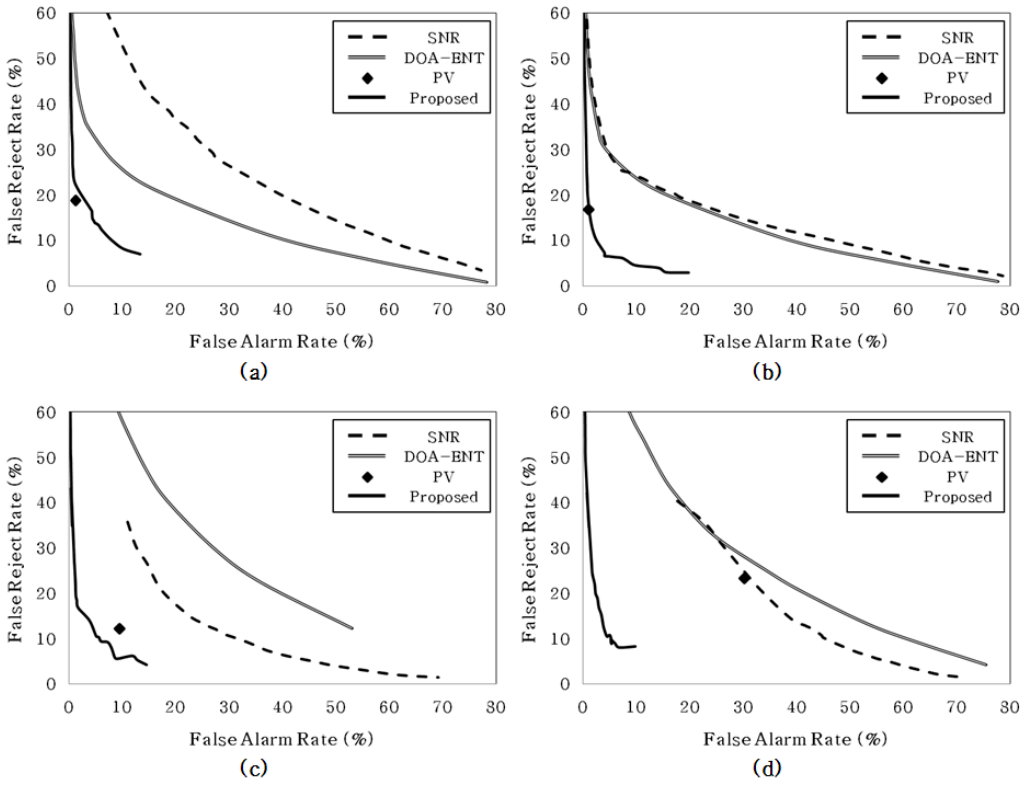

그림 4. 다양한 잡음 환경에서의 ROC 곡선 (0 dB SNR); (a) babble, (b) factory, (c) classic music, (d) interference speech
Figure 4. ROC curves for various noise environments (0 dB SNR); (a) babble, (b) factory, (c) classic music, and (d) interference speech

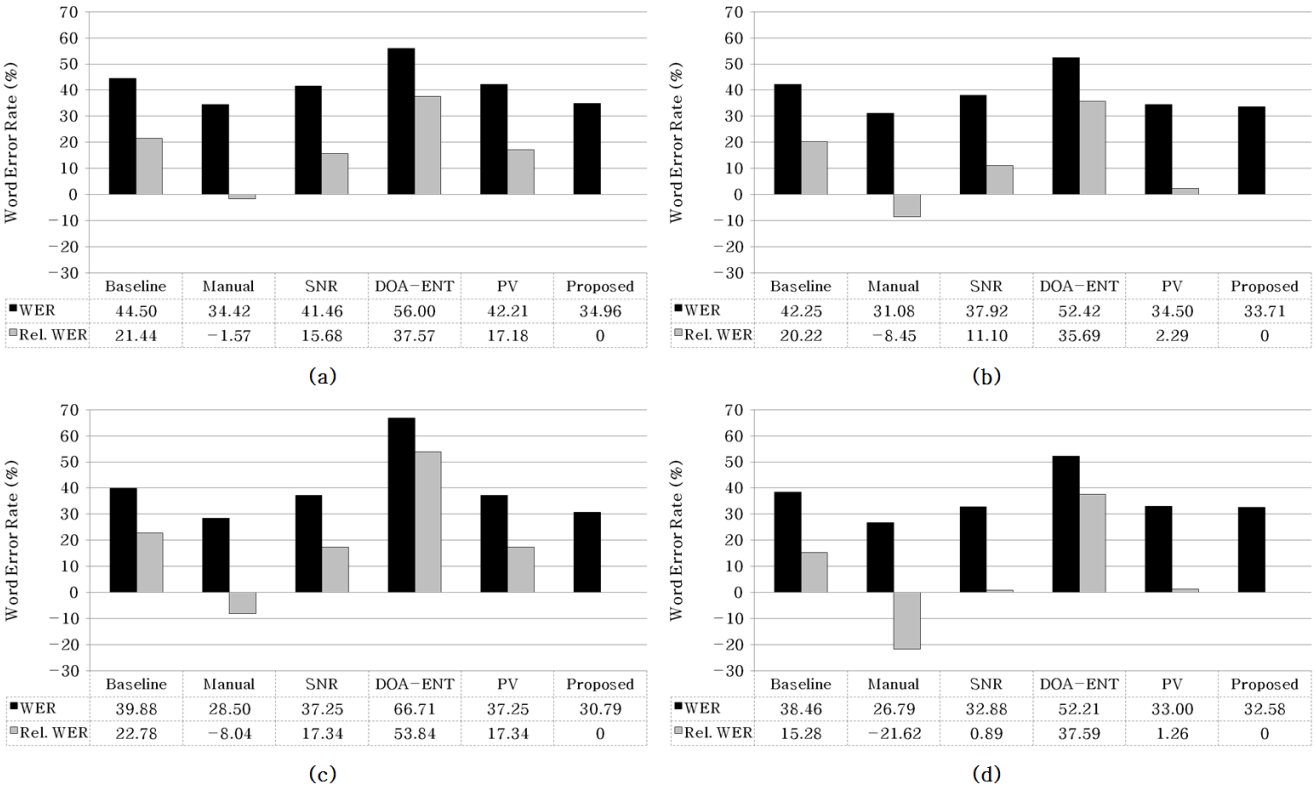

그림 5 . 잡음원의 방향에 따른 단어 오인식률; (a) 20°, (b) 30°, (c) 40°, (d) 50°
Figure 5. Word error rate according to different noise directions; (a) 20°, (b) 30°, (c) 40°, and (d) 50°

검출 방법의 경우, 고정된 문턱값을 사용함으로 인해 ROC 곡선이 아닌 점으로 표현되었다. 그림에서 보는 바와 같이, 여러 가지의 잡음환경과 SNR 환경에서 제안된 공간정보 기반의 음성구간 검출 방법이 SNR 기반의 음성구간 검출 방법 및 DOA 엔트로피 기반의 음성구간 검출 방법에 비해 향상된 FAR과 FRR를 보이는 것을 알 수 있었다. 특히, 위상 벡터를 이용한 방법과 비교할 경우, 군중잡음과 공장잡음, 경음악잡음 환경에서는 유사한 성능을 가지나 음성잡음에서 향상된 성능을 보임을 알 수 있었다.

3.2 음성인식 실험

VAD가 음성인식에 미치는 영향을 평가하기 위하여 본 논문에서는 VAD에 의해 검출된 구간만을 입력으로 하는 음성인식 실험을 수행하였다. 이를 위해 ETRI 한국어 헤드셋 인식용 단어 데이터베이스 [16] 중 성능평가에 사용되지 않는 18,240개의 단어음성을 학습용 데이터로 이용하여 음성인식 시스템을 구축하였다.

음성인식 시스템에는 39차 특징 벡터를 사용하였으며, 이를 위하여 12차 멜-켑스트럼 계수와 로그 에너지를 추출하였고, 1차, 2차 미분계수를 구하여 39차 특징벡터를 구성하였다. 음향모델은 트라이폰(triphone) 단위의 은닉 마코프 모델(Hidden Markov Model, HMM)을 기반으로 하며, 각 트라이폰은 3개의 상태(state)를 갖는 left-to-right로 표현되었다. 이 때 각 상태는 4개의 가우시안 혼합 밀도를 가지며, 결정트리(decision tree)를 통해 트라이폰들의 상태를 결합하여 총 2,296개의 상태를 갖는 음향모델을 구성하였다. 또한 사용된 어휘 수는 2,250개의 단어이며, 문법으로는 유한 상태 네트워크(Finite State Network, FSN)가 적용되었다.

<그림 5>는 baseline 음성인식 시스템과 수동 음성구간 검출 방법(Manual), SNR 기반 방법(SNR), DOA 엔트로피 기반 방법(DOA-ENT), 위상벡터 기반 방법(PV), 그리고 제안된 공간정보 기반 방법(Proposed)들에 의한 음성인식 시스템들의 단어 오인식률 (WER)과 제안된 공간정보 기반 방법 대비 상대적 오인식률 차이 (Rel. WER)를 보여준다. 그림에서 각 차트의 단어 오인식률은 잡음원의 방향별 평균 단어 오인식률로 각각 0, 10, 20 dB SNR 환경에서 군중잡음, 공장잡음, 경음악 잡음, 음성잡음 신호에 대한 평균값을 나타낸다. 그림에서 보는 바와 같이 baseline의 경우 음성구간 검출 없이 인식과정을 수행하므로 가장 높은 단어 오인식률을 보이고 있으며, 반대로 수동 음성구간 검출 방법의 경우 음성구간 검출 오류가 존재하지 않으므로 가장 낮은 단어 오인식률을 보임을 알 수 있다. 또한, DOA 엔트로피 기반의 음성구간 검출 방법의 경우 다른 음성구간 검출 방법에 비해 높은 단어 오인식률을 보이며 특히 baseline과 비교해서 성능개선이 거의 없음을 보였다. 이는 DOA 엔트로피

기반의 방법이 목표음성으로부터 추정된 DOA는 균일하게 분포하며, 잡음으로부터 추정된 DOA는 균일하지 않게 분포한다는 가정 하에 제안된 방법이나, 본 논문에서 사용된 성능평가용 데이터의 경우 잡음신호 역시 방향성이 뚜렷하기 나타나기 때문이다. 제안된 공간정보를 이용하는 통계모델 기반 음성구간 검출방법의 경우 SNR 기반의 방법, DOA 엔트로피 기반의 방법, 위상벡터 기반의 방법들에 비해 낮은 단어 오인식률을 보이며, 수동 음성구간 검출 방법과 가장 유사함을 알 수 있다. 이는 제안된 방법의 음성구간 검출 성능이 가장 우수함을 보여주는 것이다. 결과적으로는 제안된 공간정보를 이용하는 통계모델 기반 음성구간 검출 방법이 SNR 기반 음성구간 검출 방법, DOA 엔트로피 기반 음성구간 검출 방법 및 위상벡터 기반 음성구간 검출 방법에 비해 평균적으로 각각 11.68%, 41.92%, 10.15%의 오인식률의 개선을 보였다.

4. 결론

본 논문에서는 동적잡음 환경에서의 음성인식 성능향상을 위한 공간정보 기반의 음성구간 검출 방법을 제안하였다. 제안된 방법은 이중채널 입력신호로부터 획득한 ITD와 ILD를 이용하여 생성한 가우시안 확률모델을 기반으로 음성존재 및 부재 확률을 추정하였다. 또한 음성구간 결정을 위한 특징 파라미터로 시간-주파수 영역별 음성존재 대비 부재 확률 비를 추출하여 문턱값 비교를 통해 음성구간을 검출하였다. 제안된 음성구간 검출 방법은 동작특성 곡선과 검출된 구간만을 입력으로 하는 음성인식을 통한 단어 오인식률 측정을 통해 기존의 SNR 기반 음성구간 검출 방법, DOA 엔트로피 기반 음성구간 검출 방법 및 위상벡터 기반 음성구간 검출 방법의 성능과 비교하였다. 비교 결과, 제안된 공간정보를 이용하는 통계모델 기반 음성구간 검출 방법의 성능이 가장 우수하였다. 특히, 제안된 방법이 SNR 기반 방법, DOA 엔트로피 기반 방법 및 위상벡터 기반 방법에 비해 평균적으로 각각 11.68%, 41.92%, 10.15%의 상대적 오인식률을 개선하였다.

참고문헌

[1] Le Bouquin-Jeannès, R. & Faucon, G. (1995). “Study of voice activity detector and its influence on a noise reduction system,” *Speech Communication*, Vol. 16, No. 3, pp. 245-254, Apr.

[2] ETSI TS 101 707 V7.5.0 (2000). *Digital Cellular Telecommunications System (Phase 2+); Discontinuous Transmission (DTX) for Adaptive Multi-Rate (AMR) Speech Traffic Channels.*

[3] Junqua, J.-C., Mak, B. & Reaves, B. (1994). “A robust algorithm for word boundary detection in the presence of noise,” *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 3, pp. 406-412, July.

[4] Rabiner, L. R. & Sambur, M. R. (1975). “An algorithm for determining the endpoints of isolated utterances,” *Bell System Technical Journal*, Vol. 54, No. 2, pp. 297-315, Feb.

[5] Mak, B., Junqua, J.-C. & Reaves, B. (1992). “A robust speech/non-speech detection algorithm using time and frequency-based features,” *Proc. of ICASSP*, Vol. 1, pp. 269-272, Mar.

[6] ETSI Standard (2007). *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms*. ETSI ES 202 050 V1.1.5.

[7] Ramírez, J., Segura, J. C., Benítez, C., de la Torre, A. & Rubio, A. (2004). “Efficient voice activity detection algorithms using long-term speech information,” *Speech Communication*, Vol. 42, Nos. 3-4, pp. 271-287, Apr.

[8] de la Torre, A., Ramírez, J., Benítez, C., Segura, J. C., García, L. & Rubio, A. J. (2006). “Noise robust model-based voice activity detection,” *Proc. of Interspeech*, pp. 1954-1957, Sept.

[9] Tuker, R. (1992). “Voice activity detection using a periodicity measure,” *IEE Proceedings-I, Communications, Speech, and Vision*, Vol. 139, No. 4, pp. 377-380, Aug.

[10] Davis, A., Nordholm, S. & Togner y, R. (2006). “Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 2, pp. 412-424, Mar.

[11] Rubio, J. E., Ishizuka, K., Sawada, H., Araki, S., Nakatani, T. & Fujimoto, M. (2007). “Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates,” *Proc. ICASSP*, Vol. 4, pp. 385-388, Apr.

[12] Kim, G. & Cho, N. I. (2007). “Voice activity detection using phase vector in microphone array,” *Electronic Letters*, Vol. 43 No. 14, pp. 783-784, July.

[13] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (2007). *An Efficient Auditory Filterbank Based on the Gammatone Functions*, APU Report 2341, MRC, Applied Psychology Unit, Cambridge U.K.

[14] Glasberg B. R. & Moore, B. C. (1990). “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, Vol. 47, Nos. 1-2, pp. 103-138, Aug.

[15] Parzen, E. (1962). “On estimation of a probability density function and mode,” *The Annals of Mathematical Statistics*, Vol. 33, No. 3. pp. 1065-1076. Aug.

[16] Kim, S., Oh, S., Jung, H.-Y., Jeong, H.-B. & Kim, J.-S. (2002). "Common speech database collection," *Proc. of the Acoustical Society of Korea,* Vol. 21, No. 1, pp. 21-24, July
(김상훈, 오승신, 정호영, 전형배, 김정세 (2002). "공통음성 DB 구축," 한국음향학회 하계학술대회논문집, 제21권, 제1호, pp. 21-24, 7월.)

[17] Gardner, W. G. & Martin, K. D. (1995). "HRTF measurements of a KEMAR," *Journal of the Acoustical Society of America*, Vol. 97, No. 6, pp. 3907-3908, June.

• **신민화 (Shin, Min Hwa)**
전자부품연구원 멀티미디어IP 연구센터
경기도 성남시 분당구 야탑동 68번지
Tel: 031-789-7380
Email: minhwas@gmail.com
관심분야: 음성인식
현재 전자부품연구원 멀티미디어IP 연구센터 위촉연구원 재직중

• **박지훈 (Park, Ji Hun)**
광주과학기술원 정보통신공학부
광주광역시 북구 오룡동 1번지
Tel: 062-715-3121
Email: jh_park@gist.ac.kr
관심분야: 음성인식
현재 광주과학기술원 정보통신공학부 박사과정 재학중

• **김홍국 (Kim, Hong Kook),** 교신저자
광주과학기술원 정보통신공학부
광주광역시 북구 오룡동 1번지
Tel: 062-715-2228
Email: hongkook@gist.ac.kr
관심분야: 음성 및 오디오 처리, 음성인식
현재 광주과학기술원 정보통신공학부 교수

• **이연우 (Lee, Yeonwoo)**
목포대학교 공과대학 정보공학부
전남 무안군 청계면 도림리 61번지
Tel: 061-450-2745
Email: ylee@mokpo.ac.kr
관심분야: 이동통신, 해양텔레매틱스
현재 목포대학교 공과대학 정보공학부 정보통신공학과 교수

• **이성로 (Lee, Seong Ro)**
목포대학교 공과대학 정보공학부
전남 무안군 청계면 도림리 61번지
Tel: 061-450-2436
Email: srlee@mokpo.ac.kr
관심분야: 이동통신, 해양텔레매틱스
현재 목포대학교 공과대학 정보공학부 정보전자공학과 교수