2013년 10월 전자공학회 논문지 제 50 권 제 10 호
Journal of The Institute of Electronics Engineers of Korea Vol. 50, NO. 10, October 2013 http://dx.doi.org/10.5573/ieek.2013.50.10.181

논문 2013-50-10-20

CASA 시스템의 청각장면과 PAR를 이용한 음성 영역 검출에 관한 연구

(A Study on Voice Activity Detection Using Auditory Scene and Periodic to Aperiodic Component Ratio in CASA System)

김 정 호*, 고 형 화*, 강 철 호*

(Jung-Ho Kim©, Hyung-Hwa Ko, and Chul-Ho Kang)

요 약

인간의 청각은 청각 장면 분석을 통해 배경 잡음이나 여러 사람들이 동시에 말하는 상황에서도 특정 목적을 가지는 음성 신호를 청취할 수 있는 능력을 가지고 있다. 인간의 청각 능력 시스템을 잘 반영한 CASA 시스템을 이용해 음성을 분리를 할 수 있다. 그러나 CASA 세그먼트에서 음성의 위치를 잘못 결정 했을 때 CASA 시스템의 성능은 감소된다. 본 논문에서는 CASA 시스템에서 잘못된 음성 영역 위치로 인해 발생되는 성능 감소를 개선하기 위하여 청각 장면, 그리고 주기 성분과 비주기 성분의 비율(PAR)을 결합한 음성 영역 검출 알고리즘을 제안한다. 음성 영역 검출의 성능을 평가하기 위하여 백색 잡음과 자동차 잡음 환경에서 신호 대 잡음비의 변화에 따라 실험을 수행하였다. 본 논문에서는 신호 대 잡음비 15~0dB에서 기존의 알고리즘(Pitch 와 Guoning Hu)과 제안한 알고리즘을 비교한 결과, 음성 영역 검출의 정확도가 백색잡음과 자동차 잡음에서 신호 대 잡음비 15dB 에서 최대 4%, 0dB에서 최대 34% 씩 각각 향상되었다.

Abstract

When there are background noises or some people speaking at the same time, a human's auditory sense has the ability to listen the target speech signal with a specific purpose through Auditory Scene Analysis. The CASA system with human's auditory faculty system is able to segregate the speech. However, the performance of CASA system is reduced when the CASA system fails to determine the correct position of the speech. In order to correct the error in locating the speech on the CASA system, voice activity detection algorithm is proposed in this paper, which is a combined auditory scene analysis with PAR(Periodic to Aperiodic component Ratio). The experiments have been conducted to evaluate the performance of voice activity detection in environments of white noise and car noise with the change of SNR 15~0dB. In this paper, by comparing the existing algorithms (Pitch and Guoning Hu) with the proposed algorithm, the accuracy of the voice activity detection performance has been improved as the following: improvement of maximum 4% at SNR 15dB and maximum 34% at SNR 0dB for white noise and car noise, respectively.

Keywords : Auditory Scene Analysis, CASA, VAD

Ⅰ. 서 론

인간의 청각 시스템은 청각 장면 분석(ASA: Auditory Scene Analysis)[1]을 통해 배경 잡음이나 여러 사람들이 동시에 말하는 상황에서도 특정 목적을 가

* 정회원, 광운대학교 전자통신공학과
(Deptartment of Electronics and Communication Engineering, Kwangwoon University)
© Corresponding Author(E-mail: anpaul@kw.ac.kr)
※ 이 논문은 2011년도 광운대학교 연구년에 의하여 연구되었음.
접수일자: 2013년7월9일, 수정완료일: 2013년10월3일

지는 음성 신호를 청취할 수 있는데 이러한 현상을 칵테일 파티 효과(Cocktail party effect)라고 한다[2]. 다양한 잡음이 혼합된 음성 신호에서 칵테일 파티 효과처럼 음성을 분리하면 음성 인식 시스템 성능을 향상 시킬 수 있다. 그러나 성능 향상을 위하여 잡음과 혼합된 음성 신호에서 음성 성분을 분리하는 작업이 필요하다. 여기서 음성 분리(Speech segregation)를 수행하려면 음성 신호의 본질적 속성을 분석해야 한다. 음성 심리학(Psychoacoustic)과 생물학적(Biology) 발견에 기반을 둔 CASA(Computational Auditory Scene Analysis) 시스템은 다양한 잡음에 대한 사전적 정보 및 학습 없이 혼합된 음성 신호에서 높은 음성 분리 성능을 보여주고 있다[3]. 이러한 CASA 시스템의 동작은 2단계로써 CASA 세그먼트(Auditory Segmentation)와 그룹화(Grouping)로 구성되어 있다. CASA 세그먼트는 청각 장면을 시간-주파수(T-F) 단위로 변화해 음성(예로 음소, 음절, 단어)영역의 위치를 구별하지만 잡음 환경에서 시간-주파수에서 잘못된 음성 영역 위치 결정으로 CASA 시스템의 음성 분리 성능은 감소된다. 그러나 혼합된 음성 신호에서 음성 영역의 위치를 정확하게 알 수 있다면 CASA 시스템의 성능을 향상시킬 수 있다.

본 논문에서는 CASA 시스템의 음성 분리 성능 개선을 위해 청각 장면, 그리고 주기 성분과 비주기 성분의 비율(PAR: Periodic to Aperiodic component Ratio)[4~5]을 결합한 음성 영역 검출(VAD: Voice Activity Detection)을 제안한다. 제안한 알고리즘의 성능을 평가하기 위하여 백색 잡음과 자동차 잡음 환경에서 신호 대 잡음비(SNR: Signal to Noise Ratio)의 변화에 따라 기존 알고리즘과 제안한 알고리즘의 성능을 비교하였다.

논문의 구성은 다음과 같다. Ⅱ장에서 연구 배경에 대해 알아보고 Ⅲ장에서는 제안한 음성 영역 검출 알고리즘에 대해 소개한다. Ⅳ장은 실험 결과에 대해 기술하며, Ⅴ장에서 결론을 내린다.

Ⅱ. 연구 배경

2.1 청각 기관의 인지 특성

인간의 청각 구조는 크게 외이(Outer ear), 중이(Middle ear), 내이(Inner ear) 세가지로 구분한다. 외이에서 들어오는 미약한 신호가 중이의 고막(Tympanic membrance)을 진동한다. 진동 신호는 3개의 뼈(Malleus, Incus, Stapes)를 통해 신호를 증폭한다. 증폭 신호는 내이의 달팽이관(Cochlea)에 있는 유모 세포(Hair cell)로 전달된 후 청각 신경(Auditory nerve)을 통해 뇌로 전달된다.

2.2 청각 신경의 신호 분석

외이와 중이를 통해 증폭된 신호는 달팽이관 안에 있는 세포막의 유모 세포에서 가청 주파수 영역별로 인식한다. 달팽이관에서 특정 구간별로 주파수 응답(frequency response) 특성을 표현한다.

사람의 청각 기관(외이, 중이, 달팽이관)을 모델링하기 위해 CASA 시스템은 감마톤 필터(Gammatone filter) 채널을 가진 ERB 필터뱅크(Equivalent Rectangular Bandwidth filterbank)를 사용한다. ERB 필터뱅크는 Glasberg와 Moore 알고리즘을 사용한다[6].

$$g(t,f_c)= t^{n-1}e^{-2\pi Bt}\cos(2\pi f_c t+\phi) \quad (1)$$

식(1)에서 B는 필터 대역폭, n은 필터 차수이고 n=4를 사용하였다. f_c는 중심 주파수를 나타낸다.

2.3 CASA 시스템

CASA 시스템은 음향심리학과 생물학적 발견을 기반으로 인간 청각 시스템의 기능을 모방하고, 음향 신호의 이해를 장면 분석으로 해석한다. CASA 시스템의 처리 과정은 청각 신경 분석(Auditory peripheral Analysis), 특징 추출(Feature extraction), 청각 세그멘트(Auditory segmentation), 그룹화(Grouping)이다[3]. CASA 시스템의 목적은 이상적인 2진 시간-주파수 마스크(Ideal binary T-F mask) 계산을 통해 다양한 잡음 환경에서 음성 분리를 한다.

2.4 음성 영역의 검출

음성 영역 검출은 음성 부호화, 음성 분리, 음성 인식 등 음성통신 어플리케이션에서 음성과 잡음 영역을 찾아낸다. 그러나 음성 신호는 다양한 잡음으로부터 왜곡되기 때문에 혼합된 음성 신호에서 음성 영역을 찾는 것은 어렵다. 잘못된 음성 영역은 음성의 품질 결정과 음성 어플리케이션의 성능 감소와 같은 심각한 문제를 발생시킨다. 따라서 음성 영역 검출을 위하여 신뢰할 수 있는 알고리즘을 사용해야 한다.

Ⅲ. 제안한 음성 영역 검출 알고리즘

3.1 청각 장면과 PAR을 결합한 음성 영역 검출 알고리즘

CASA 시스템은 잡음에 대한 사전 정보 없이 음성을 분리한다. CASA 시스템은 2단계로 구성되며, 첫번째 CASA세그먼트는 청각 장면을 시간-주파수(T-F) 단위로 변환하고 두번째 그룹화는 CASA 세그먼트를 stream으로 구성한다.

여기서 시간-주파수로 표현되는 청각 장면 정보로부터 Pitch[3] 또는 CASA 세그먼트를 이용해 음성 영역을 검출한다. 그러나 잡음 환경에서 음성 영역을 결정하는데 많은 어려움이 있다. 2007년 Guoning Hu[7]는 채널간의 상관성(correlation)을 이용하여 음성 영역을 검출하는 알고리즘을 제안하였지만 잡음 환경에서 정확하게 음성 영역을 검출하지 못하였다. 음성 영역 검출을 위한 다양한 알고리즘이 있지만 잡음 환경에서 성능을 보장하지 못하기 때문에 잡음에 강인한 음성 영역 검출 알고리즘을 사용해야 한다.

본 논문에서는 CASA 시스템의 성능을 향상 시키기 위해 청각 장면 정보와 주기 성분과 비주기 성분의 비율(PAR)을 결합한 음성 검출 알고리즘을 제안한다. 주기 성분과 비주기 성분 비율을 기반으로 하는 음성검출 알고리즘의 특징은 신호 대 잡음 비의 변화에 더 강인하고, 통계 기반에 의해 음성의 존재 유무를 결정한다.

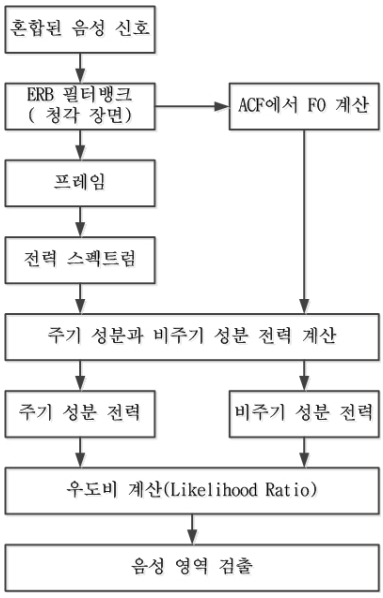

그림 1. 제안한 알고리즘의 구성도
Fig. 1. Block Diagram of Proposed VAD algorithm.

$$s(t,c)=s(t)*g(t,f_c) \tag{2}$$

식(2)에서 $s(t)$는 혼합된 음성 신호이고, $g(t,f_c)$는 ERB 필터뱅크이고, f_c는 ERB 필터 채널의 중심 주파수이고, c는 필터 채널이다. $s(t,c)$는 $s(t)$와 $g(t,f_c)$의 시스템 응답이다. 혼합된 음성 신호가 ERB 필터뱅크를 통과하면 청각 정보를 얻는다. 청각 정보는 말 의미가 아닌 말 소리에 정보가 있다.

청각 분석(Auditory Analysis)에서 사람의 청각 모델인 ERB 필터뱅크를 이용해 각 채널 별로 주파수 분석을 할 수 있다. 시간-주파수(T-F)로 표현하는데 이것을 코클로그램(Cochleagram) 이라 한다. 코클로그램은 128 채널로 구성되고, 전체 대역은 50~5000Hz의 대역을 가진다.

음성 신호($s(t)$)는 주기 성분과 비주기 성분을 가지는 준 주기 신호이다. 따라서 혼합된 음성의 청각 장면 신호로부터 주기 성분($s_p(t)$)과 비주기 성분($s_a(t)$)으로 분해 할 수 있다.

$$s(t)=s_p(t)+s_a(t) \tag{3}$$

식(3)에서 혼합된 음성 신호($s(t)$)는 프레임 단위로 처리한다.

$$|S(n,m)|^2=|S_p(n,m)|^2+|S_a(n,m)|^2 \tag{4}$$

식(4)는 식(3)의 STFT이다. n은 프레임 개수이고, m은 주파수 index이다. ERB 필터뱅크 처리된 $s(t,c)$는 식(4)를 바탕으로 식(5)와 같이 표현된다.

$$|s(n,c)|^2=|s_p(n,c)|^2+|s_a(n,c)|^2 \tag{5}$$

식(5)는 식(4)과 같은 식(3)의 STFT이다. c는 필터 채널(또는 필터 채널 index)이다.

$$\rho(n)=\frac{1}{C}\sum_{c=0}^{C-1}|s(n,c)|^2 \tag{6}$$

식(6)에서 $\rho(n)$는 short-time power이고, C는 ERB 필터뱅크의 전체 채널 수이다.

$$\rho(n)=\rho_p(n)+\rho_a(n) \tag{7}$$

식(3),(4),(5),(7)에서 주기 성분과 비주기 성분 구분은 기본주파수(F0)를 이용한다. 기본주파수는 주기 성분으

로 ERB 필터뱅크의 ACF에서 얻는다. 기본주파수 영역은 50~500Hz이다. 프레임 n에서 고조파(harmonic) 개수 $v(n)=\frac{\max(F0)}{f_0(n)}$ 이다.

$$\rho(n)=\hat{\rho}_p(n)+\hat{\rho}_a(n) \tag{8}$$

식(8)은 식(7)[6]으로부터 주기 성분과 비주기 성분을 근사화 한다.

$$\hat{\rho}_p(n)=\eta\frac{\sum_{k=1}^{v(n)}|s(n,[kf_0(n)])|^2-v(n)\rho(n)}{1-\eta v(n)} \tag{9}$$

식(9)는 근사화된 주기 성분 전력($\hat{\rho}_p(n)$)이다. 여기서 $s(n,[kf_0(n)])$는 k차 고조파 $kf_0(n)$에 근접한 주파수 값을 가지는 채널 index 으로부터 프레임 n에 대한 주파수 값을 얻는다. η는 임의 상수이고, $\eta=0.01$이다.

$$\hat{\rho}_a(n)=\rho(n)-\hat{\rho}_p(n)) \tag{10}$$

식(10)는 근사화된 비주기 성분 전력($\hat{\rho}_a(n)$)이다.

$$\hat{f}_0(n)=\arg\max_{f_0(n)}\left(\sum_{k=1}^{v(n)}|s(n,[kf_0(n)])|^2-v(n)\rho(n)\right) \tag{11}$$

식(11)에서는 REPS[8]으로부터 업데이트된 기본주파수(F0)를 계산한다.

3.2 주기성분 전력 및 비주기성분 전력의 우도비 (Likelihood Ratio) 계산

근사화된 주기 성분 전력($\hat{\rho}_p(n)$)으로부터 음성 영역을 결정할 수 있다. 그러나 잘못된 영역 결정은 CASA 시스템 성능에 영향을 주기 때문에 통계적 모델을 이용해 음성 영역 검출의 정확성을 높이고 있다[9].

변수 $H_n=0$인 경우에 혼합된 음성에서 음성이 없는 비주기 성분을 $\rho(n)=\hat{\rho}_a(n)$이라고 가정하면 비주기 성분의 에러는 식(12)과 같다.

$$\epsilon_a(n)=\rho(n)-\hat{\rho}_a(n)=\hat{\rho}_p(n) \tag{12}$$

$$p(\rho(n)|H_n=0)=\frac{1}{\sqrt{2}\,\alpha\hat{\rho}_a(n)}exp\left(-\frac{1}{2\alpha^2}\left(\frac{\hat{\rho}_p(n)}{\hat{\rho}_a(n)}\right)^2\right)) \tag{13}$$

식(13)는 비음성 영역에 대해 혼합된 음성의 우도비이다. 평균은 0이고, 표준편차는 $\alpha\hat{\rho}_a(n)$를 가진다. α는 상수이다. 반대로 식(14)에서는 변수 $H_n=1$일 때 혼합된 음성에서 음성 영역이 존재하는 주기 성분의 에러를 측정한다.

$$\epsilon_p(n)=\rho(n)-\hat{\rho}_p(n)=\hat{\rho}_a(n) \tag{14}$$

$$p(\rho(n)|H_n=1)=\frac{1}{\sqrt{2}\,\beta\hat{\rho}_p(n)}exp\left(-\frac{1}{2\beta^2}\left(\frac{\hat{\rho}_a(n)}{\hat{\rho}_p(n)}\right)^2\right) \tag{15}$$

식 (15)는 음성 영역에 대한 혼합된 음성의 우도비이다. 평균은 0이고, 표준편차는 $\beta\hat{\rho}_p(n)$를 가진다. β는 상수이다.

$$\left(L(n)=\frac{p(\rho(n)|H_n=1)}{p(\rho(n)|H_n=0)}\right)\underset{H_0}{\overset{H_1}{>\atop<}}\left(\tau=\frac{p(H_n=0)}{p(H_n=1)}\right) \tag{16}$$

식(16)에서는 프레임 n에 대해 우도비($L(n)$)와 임계값(τ)을 비교하여 음성의 존재 유무를 결정한다.

Ⅳ. 실험 결과

음성 영역을 검출하기 위해 실험에 사용된 데이터는 ETRI의 PBW 음성 데이터 베이스 445개의 단어 음성, OHIO 대학 PNL의 100 non-speech sounds와 NOIZEUS(Noisy speech corpus)를 사용하여 “꽃밭”외의 깨끗한 음성과 각각의 잡음을 15~0dB로 혼합하여 성능 평가하였다. ERB필터뱅크는 128채널로 구성되고, 프레임은 20ms으로 하여 10ms 씩 이동하면서 계산하

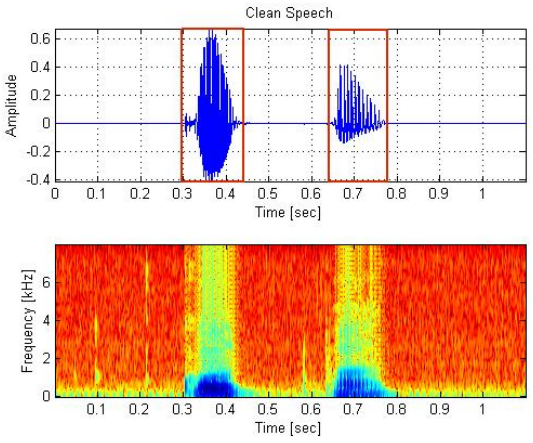

그림 2. 깨끗한 음성 남성(“꽃밭”)
Fig. 2. Clean speech man(‘kkochbat’).

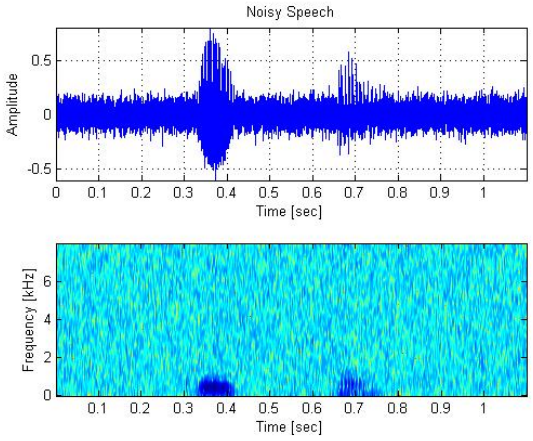

그림 3. 백색잡음이 혼합된 음성(SNR 0dB)
Fig. 3. Noisy speech created by additive white noise at (SNR 0dB).

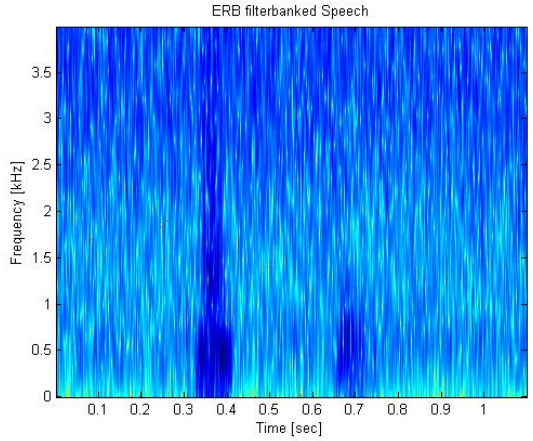

그림 4. ERB 필터뱅크를 통과한 음성
Fig. 4. Speech after ERB filter bank.

였다.

그림 2는 깨끗한 음성의 시간영역과 스펙트로그램(spectrogram)이다. 시간영역에서 음성 영역은 직접 레이블(label) 하였다.

그림 3은 잡음에 의해 음성 신호가 왜곡되었다.

그림 4는 혼합된 음성이 ERB 필터뱅크를 통과한 청각 장면 정보의 코클로그램이다.

신호 대 잡음비가 15~0dB인 경우, 제안한 알고리즘(그림 7)을 일반적으로 많이 사용하는 Pitch(그림 5)와 Guoning Hu 알고리즘(그림 6)과 음성 영역 검출을 위한 성능 평가를 하였다.

Guoning Hu 알고리즘은 채널간의 상관성으로부터 음성 영역 분류 후 임계값을 이용해 음성 영역을 결정한다. 그러나 잡음이 혼합된 음성인 경우에 그림 6과

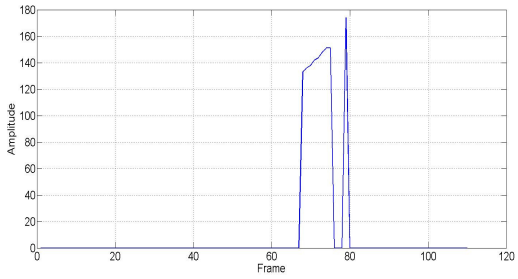

그림 5. 피치 알고리즘으로 음성 영역 검출
Fig. 5. VAD using Pitch algorithm.

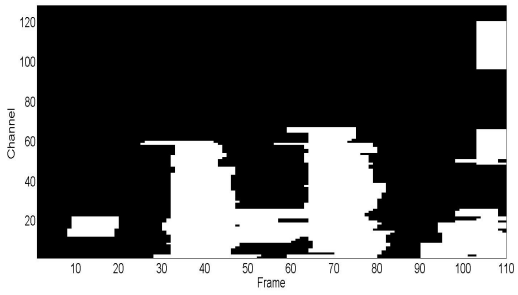

그림 6. Guoning Hu 알고리즘으로 음성 영역 분류
Fig. 6. Speech classification of Guoning Hu algorithm.

같이 음성 영역을 분류하는 과정에서 잡음이 포함된다. 잡음의 영향으로 잘못된 음성 영역을 결정한다.

음성 영역을 직접 레이블한 그림 2를 그림 5, 그림6의 기존 알고리즘과 비교하면 기존 알고리즘들이 음성 영역을 잘못 검출하고 있음을 알 수 있다. 그러나 그림 7의 제안한 알고리즘은 그림 2의 깨끗한 음성 영역과 비교할 때, 비슷한 위치에서 음성을 검출하고 있다. 본 논문의 성능평가를 위한 그림 8, 9에서 발성 정확율(Corr: utterance correct rate)는 식 (17)에 의해 계산한다. 식 (17)은 잡음이 혼합된 음성에서 음성 영역 검출의 정확성을 의미한다. 그리고 음성 인식(Speech Recognition)을 이용해 발성 정확율(Corr)이 정확하게 음성을 검출 했는지는 워드 정확율(word accuracy)로 평가한다[10].

$$Corr(\%) = \frac{\text{정확하게 검출된 음성 영역 개수}}{\text{이상적인 전체 음성 영역 개수}} \times 100\% \quad (17)$$

그림 8~9에서는 음성 영역의 검출 성능을 평가하기 위하여 백색 잡음과 자동차 잡음 환경에서 신호 대 잡음비의 변화(SNR 15~0dB)에 따라 실험을 수행하였다. 기존의 알고리즘들과 제안한 알고리즘을 비교한 결과

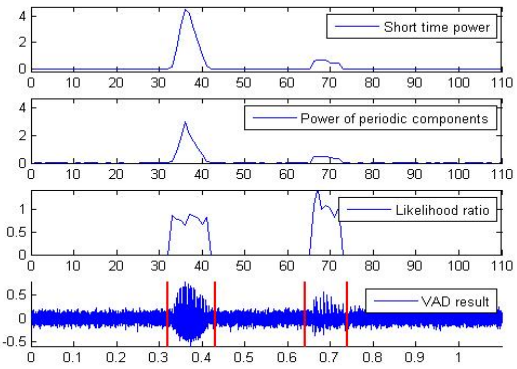

그림 7. 제안한 알고리즘으로 음성 영역 검출
Fig. 7. VAD of proposed algorithm.

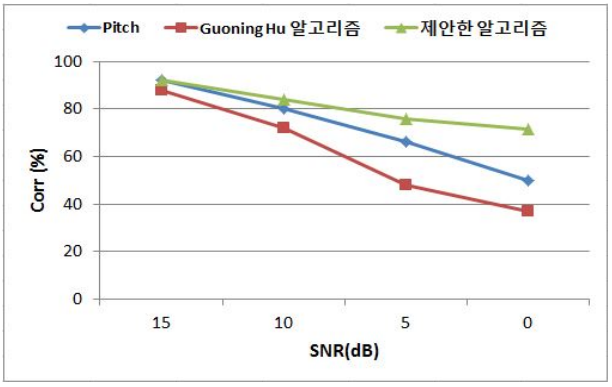

그림 8. 백색잡음 환경에서 음성 영역 검출 성능 평가
Fig. 8. Detection performance in white noise environment.

그림 9. 자동차 환경에서 음성 영역 검출 성능 평가
Fig. 9. Detection performance in car noise environment.

백색잡음과 자동차 잡음에서 각각 15dB 에서 최대 4%, 0dB에서 최대 34% 씩 음성 영역의 검출성능이 향상되었다.

V. 결 론

다양한 잡음 환경에서 음성의 고유 특성 및 음성 정보의 손상을 줄이고 음성 분리 성능을 높이기 위해서는 정확한 음성 영역의 검출이 필요하다. 본 논문에서는 CASA 시스템의 청각 장면 정보와 주기 성분과 비주기 성분의 비율(PAR)을 이용한 음성 검출 알고리즘을 결합한 알고리즘을 제안하였다. 성능평가를 위한 실험 결과 기존의 알고리즘(Pitch, Guoning Hu)과 비교할 때, 본 연구에서 제안한 알고리즘이 신호 대 잡음비의 다양한 변화에 더욱 강인함을 확인하였다.

REFERENCES

[1] A. S. Bregman, "*Auditory Scene Analysis: The Perceptual Organization of Sound*," Cambridge, MIT Press, 1990.

[2] 정상봉, 구자일, 홍준표, "웨이블렛 변환과 독립 성분 분석을 이용한 음성 블라인드 소스 분리에 대한 연구", 전자공학회논문지, 제40권 IE편, 제2호, 15-22쪽, 2003년 6월

[3] DeLiang Wang and G. J. Brown, "*Computational auditory scene analysis," Comput. Speech Lang.*, vol. 8, pp. 297-336, 1994.

[4] M. Fujimoto and K. Ishizuka, T. Nakatani and N. Miyazaki, "Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio," *Proc. Interspeech* 7, pp 230-233, 2007.

[5] Naotoshi Seo, "Individual voice activity detection using periodic to aperiodic component ration based activity detection(PARADE) and Gaussian mixture speaker models," (http://note.sonots.com/SciSoftware/IVAD.html)

[6] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research* 47, pp. 103-138, 1990.

[7] Guoning Hu and Deliang Wang, "Auditory Segmentation Based on Onset and Offset Analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp 396-405, 2007.

[8] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Am.* 116, pp. 3690-3700, 2004.

[9] Jongseo Sohn, Nam Soo Kim, "A statistical

model-based voice activity detection," *IEEE Signal Process*, vol. 6, pp. 1-3, 1999
[10] M. Fujimoto and K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," *ICASSP*, pp. 4441-4444, 2008.

저 자 소 개

김 정 호(정회원)
2000년 대진대학교 통신공학과 학사 졸업
2002년 광운대학교 전자통신 공학과 석사 졸업
2006년～현재 광운대학교 전자공학과 박사 수료
<주관심분야 : 음성신호처리, 음성분리, 잡음처리>

강 철 호(정회원)
1975년 한양대학교 전자공학과 학사 졸업.
1979년 서울대학교 전자공학과 석사 졸업.
1988년 서울대학교 전자공학과 박사 졸업.
1977년～1981년 국방과학연구소
1983년～현재 광운대학교 전자통신공학과 교수
<주관심분야 : 음성신호처리, 통신신호처리>

고 형 화(정회원)
1979년 서울대학교 전자공학과 학사 졸업
1982년 서울대학교 전자공학과 석사 졸업
1989년 서울대학교 전자공학과 박사 졸업
1985년～현재 광운대학교 전자통신공학과 교수
<주관심분야 : 영상통신, 임베디드 시스템, JBIG2>