Teager Energy 기반의 수정된 파워 스펙트럼 편차를 이용한 음성 검출

Voice Activity Detection Using Modified Power Spectral Deviation Based on Teager Energy

송지현*, 송영록, 심현민, 이상민
J. H. Song, Y. R. Song, H. M. Shim, S. M. Lee

요 약

본 논문에서는 잡음 상황에서 강인한 음성 특성을 나타내는 TE (teager energy) 기반의 특징벡터를 이용한 음성 검출 알고리즘을 제안하였다. 입력 신호에 TEO (teager energy operator)를 적용하고, 이를 이용하여 음성 검출 알고리즘에서 우수한 성능을 보여주는 파워 스펙트럼 편차를 구하였다. 또한, 제안된 음성 검출 알고리즘의 성능 향상을 위하여 통계적 모델 기반의 우도비를 TE 기반의 파워 스펙트럼 편차의 가중치 요소로 적용하였다. 제안된 알고리즘의 성능 검증을 위해서 전체 오차율, ROC (receiver operating characteristics), PESQ (perceptual evaluation of speech quality)와 같은 객관적 실험을 수행하였다. 실험 결과 5dB SNR 이하의 낮은 SNR을 갖는 비 정상 잡음 환경에서 제안한 음성 검출 알고리즘이 약 2.6%의 전체 오차율 감소와 약 0.053의 PESQ 점수 향상을 나타내었다.

ABSTRACT

In this paper, we propose a novel voice activity detection (VAD) algorithm using feature vectors based on TE (teager energy). Specifically, power spectral deviation (PSD), which is used as the feature for the VAD in the IS-127 noise suppression algorithm, is obtained after the input signal is transfomed by Teager energy operator. In addition, the TE-based likelihhod ratio are derived in each frame to modifiy the PSD for further VAD. The performance of our proposed VAD algorithm are evaluated by objective testing (total error rate, receiver operating characteristics, perceptual evaluation of speech quality) under various environments, and it is found that the proposed method yields better results than conventional VAD algorithms in the non-stationary noise environments under 5 dB SNR (total error rate = 2.6% decrease, PESQ score = 0.053 improvement).

Keyword : Voice activity detection, Teager energy, Power spectral deviation

1. 서론

음성 검출기는 입력된 신호로부터 음성을 검출하는 시스템으로 음성코덱, 잡음 제거와 같은 다양한 음성 시스템의 전처리로서 사용되는 알고리즘으로 그 중요성이 증가하고 있다 [1-3]. 특히, 잡음 제거 알고리즘의 경우 잡음 제거를 위해 사용되는 중요 특징벡터의 추출이 음성 검출 알고리즘에 의해서 이루어지기 때문에 음성 검출 성능이 중요한 요소임을 알 수 있다 [4-5]. 초기 음성 검출 알고리즘은

접 수 일 : 2014.02.13
심사완료일 : 2014.02.25
게재확정일 : 2014.02.27
* 송지현 : 인하대학교 전자공학과 박사과정
neverjin0109@naver.com (주저자)
송영록 : 인하대학교 전자공학과 연구교수
gateway32@inha.ac.kr (공동저자)
심현민 : 인하대학교 정보전자공동연구소 연구교수
hmshim@inha.ac.kr (공동저자)
이상민 : 인하대학교 전자공학과 교수
sanglee@inha.ac.kr (교신저자)
※ 본 연구는 서울시 산학연 협력사업(SS100022)과 중소기업 기술개발 사업의 미래선도과제 (S2044922) 지원에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

음성과 잡음에 대해 상이한 통계적 분포를 갖는 특징벡터와 문턱값과의 비교를 통해서 이루어 졌다. 음성 검출에 사용되는 중요 특징벡터로 파워 스펙트럼 편차(power spectral deviation), 스펙트럼 에너지(spectral energy), 영교차율(zero-crossing), 선형예측 계수(liner prediction coefficient), 주기성(periodic measure), 포만트 형태(formant shape)등이 있다 [3-5]. 이러한 특징 벡터들은 잡음 레벨이 비교적 적은 상황에서 우수한 음성 검출 성능을 보여주지만, 잡음이 심각한 상황이나 잡음 성분이 빠르게 변화하는 비 정상 잡음 환경에서는 잡음과 음성에 대해서 뚜렷이 구분되는 특징을 나타내지 못하기 때문에 분류 성능이 현저하게 저하되는 단점이 존재한다.

본 논문에서는 낮은 SNR을 갖는 잡음 환경에서도 우수한 음성 검출 성능을 제공하기 위해서 TE (teager energy) 기반의 스펙트럼 편차를 이용한 음성 검출 알고리즘을 제안한다. 구체적으로, TEO (teager operator)를 통해서 입력 신호에 포함되어있는 잡음 성분을 줄여주고, 향상된 음성 신호를 이용하여 파워 스펙트럼 편차를 추출한다. 또한, 추출한 특징벡터의 음성과 잡음에 대한 분류 특성을 향상시키기 위해서 통계적 모델 기반의 우도비(likelihood ratio)를 가중치 요소로 적용하였다.

본 논문의 구성은 다음과 같다. 2장에서 TEO의 잡음 제거 원리에 대해서 살펴보고 3장에서는 제안한 음성 검출 알고리즘에 사용되는 TE 기반의 특징벡터와 음성 검출 결정식에 대해서 살펴본다. 4장에서는 객관적 실험을 통해서 제안한 음성 검출 알고리즘의 성능을 검증하고, 5장에서 결론을 맺는다.

2. TEO (teager energy operator)

이산 신호에서 TEO는 다음과 같이 정의 된다 [6-8].

$$\Psi_d[x(n)] = x(n)^2 - x(n+1)x(n-1) \tag{1}$$

여기서 $x(n)$은 이산 신호를 나타내고, $\Psi_d[x(n)]$는 이산 시간에서의 TE를 나타낸다. 실제 환경에서는 입력 음성 신호에 배경잡음이 존재하기 때문에 마이크로폰의 입력 신호($y(n)$)는 다음과 같이 나타낼 수 있다.

$$y(n) = s(n) + d(n) \tag{2}$$

여기서 $s(n)$와 $d(n)$은 각각 깨끗한 음성 신호와 잡음 신호를 나타내고, 0의 평균 값을 갖고, 서로 상관 관계가 없다고 가정한다. 식 (2)를 식 (1)에 적용하면 입력 신호의 $\Psi_d[y(n)]$는 다음과 같이 나타 낼 수 있다.

$$\Psi_d[y(n)] = \Psi_d[s(n)] + \Psi_d[d(n)] + \widetilde{\Psi}_d[s(n),d(n)] \tag{3}$$

여기서 $\widetilde{\Psi}_d[s(n),d(n)]$는 $s(n)$과 $d(n)$의 cross-$\widetilde{\Psi}_d$를 나타내고, 다음과 같이 도출 된다.

$$\widetilde{\Psi}_d[s(n),d(n)] = s(n)d(n) - 0.5s(n-1)d(n+1) - 0.5s(n+1)d(n-1) \tag{4}$$

위에서 $s(n)$와 $d(n)$는 0 평균값을 갖고, 서로 독립이라고 가정하였기 때문에 cross-$\widetilde{\Psi}_d$의 기댓값은 0을 갖고, $\Psi_d[y(n)]$의 기댓값은 다음과 같이 구해진다 [8].

$$E\{\Psi_d[y(n)]\} = E\{\Psi_d[s(n)]\} + E\{\Psi_d[d(n)]\} \tag{5}$$

일반적으로 $E\{\Psi_d[s(n)]\}$는 $E\{\Psi_d[d(n)]\}$에 비해서 상대적으로 큰 값을 갖기 때문에 식(5)는 다음과 같이 나타낼 수 있다 .

$$E\{\Psi_d[y(n)]\} \approx E\{\Psi_d[s(n)]\} \tag{6}$$

식(6)에서 볼 수 있는 것처럼 TEO를 통해서 전처리된 입력 신호는 원신호에서 잡음 성분이 줄어든 형태로 나타나기 때문에 이를 이용해서 구해진 특징 벡터는 잡음 상황에서 보다 강인한 음성 특성을 제공할 수 있다.

3. TE-LR을 적용한 파워 스펙트럼 편차 기반의 음성 검출 알고리즘

이전 장에서 TEO를 통해서 잡음 상황에서 보다 강인한 음성 특성을 유지하는 특징벡터를 추출할 수 있음을 알 수 있었다. 이를 기반으로 본 논문에서는 강인한 음성 특성을 제공하는 TEO 기반의 특징벡터인 파워 스펙트럼 편차(TE-PSD, TE-power spectral deviation)와 통계적 모델의 음성 존재와 부재에 대한 우도비(TE-likelihood ratio, TE-LR)를 이용한 음성 검출 알고리즘을 제안하였고, 그림 1에 블록다이어그램을 나타내었다.

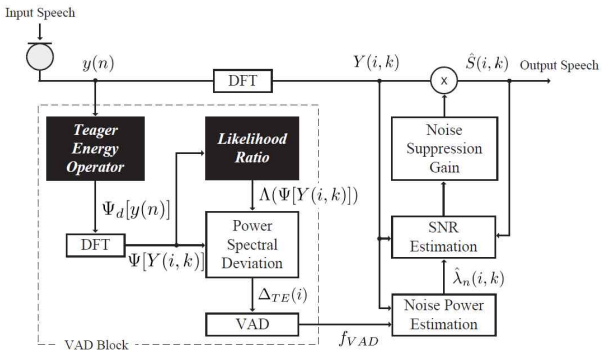

그림 1. 제안된 음성 검출 알고리즘의 블록다이어그램

3.1 TE 기반의 파워 스펙트럼 편차

TE-PSD($\Delta_{TE}(i)$)는 다음과 같이 구해진다 [2].

$$\Delta_{TE}(i) = \log_{10}\left(\frac{1}{M}\sum_{k=1}^{M}|Y_{TE}(i,k) - \overline{Y}_{TE}(i,k)|\right) \quad (7)$$

여기서 i는 프레임 인덱스를 나타내고, M은 주파수 밴드수를 나타낸다. $Y_{TE}(i,k)$는 오염된 신호의 파워 스펙트럼($|\Psi[Y(i,k)]|^2$)을 나타내고, $\overline{Y}_{TE}(i,k)$는 추정된 잡음 신호의 파워 스펙트럼을 나타내고 현재 프레임이 잡음이라고 판단되면 업데이트 된다.

$$\overline{Y_{TE}}(i,k) = \alpha \bullet \overline{Y_{TE}}(i-1,k) + (1-\alpha) \bullet Y_{TE}(i,k) \quad (8)$$

여기서 α는 0과 1 사이의 값을 갖는 스무딩 파라미터를 나타낸다.

3.2 TE 기반의 우도비

통계적 모델 기반의 음성 검출기에서 TE 기반의 우도비($\Lambda(\Psi[Y(i,k)])$)는 다음과 같이 구해진다 [9-10].

$$\Lambda(\Psi[Y(i,k)]) = \frac{p(\Psi[Y(i,k)]|H_1)}{p(\Psi[Y(i,k)]|H_0)} = \frac{1}{1+\zeta(i,k)} exp\left[\frac{\eta(i,k)\zeta(i,k)}{\zeta(i,k)}\right] \quad (9)$$

여기서 H_0와 H_1는 음성 부재와 음성 존재를 나타내고, $\eta(i,k)$ 와 $\zeta(i,k)$는 각각 *a posteriori* signal-to-noise ratio (SNR)와 *a priori* SNR을 나타낸다.

$$\eta(i,k) \equiv \frac{|\Psi[Y(i,k)|^2}{\sigma_d(i,k)} \quad (10)$$

$$\zeta(i,k) \equiv \frac{\sigma_s(i,k)}{\sigma_d(i,k)} \quad (11)$$

$\sigma_d(i,k)$와 $\sigma_s(i,k)$는 잡음 신호와 깨끗한 음성 신호의 추정된 전력을 나타낸다. 현재 프레임의 깨끗한 음성 신호의 전력은 직접적으로 구할 수 없기 때문에 Decision-Directed (DD) 방식을 통해서 구해진다 [5].

각 주파수에 대해서 서로 통계적 독립이라고 가정하면 현재 프레임에 대한 TE 기반의 우도비($\beta(i)$)는 다음과 같이 구해진다.

$$\beta(i) = \prod_{k=1}^{M}\Lambda(\Psi[Y(i,k)]) \quad (12)$$

3.3 제안된 음성 검출 알고리즘

TEO 기반의 특징벡터를 이용한 제안된 음성 검출 알고리즘의 최종 결정식은 다음과 같다.

$$f_{VAD} = \begin{cases} speech, & \text{if } \Delta'_{TE}(i) > T \\ non-speech, & otherwise \end{cases} \quad (13)$$

여기서 T는 제안된 음성 검출기의 문턱값을 나타내고, $\Delta'_{TE}(i)$는 TE-LR을 적용한 수정된 TE-PSD를 나타낸다.

$$\Delta'_{TE}(i) = \log_{10}\left(\frac{\beta(i)}{M}\sum_{k=1}^{M}|Y_{TE}(i,k) - \overline{Y_{TE}}(i,k)|\right) \quad (14)$$

그림 2는 다양한 특징 벡터에 대한 결과값을 나타내었다. 그림 2(c)를 통해서 TEO를 통해서 오염된 음성 신호의 잡음 성분이 줄어드는 것을 확인 할 수 있다. 또한. 그림 2(d)~그림 2(e)를 통해서 기존의 PSD와 LR 특징벡터가 잡음에 대해서 민감하게 변화 되는 것을 확인 할 수 있다. 반면에, 그림 2(f)에서와 같이 제안한 특징 벡터의 경우 기존의 특징벡터에 비해서 우수한 음성과 잡음 분류 특성을 보여주는 것을 볼 수 있다.

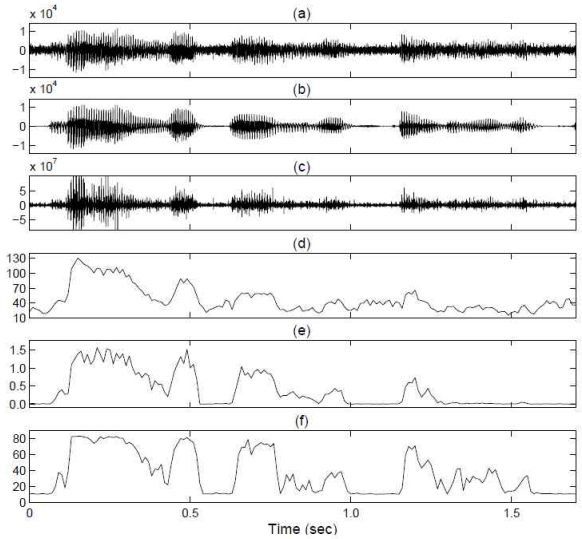

그림 2. (a) 오염된 음성 파일 (b) 깨끗한 음성 파일 (c) TE 적용된 음성 파일 (d) 파워 스펙트럼 결과 (e) 우도비 결과 (f) 제안된 TE 기반의 파워스펙트럼 결과

4. 실험

본 논문에서는 제안된 음성 검출기의 성능 평가를 위해서 다양한 잡음 환경에서 객관적인 실험을 수행하였다. 성능평가 지표는 전체 오차율(total error rate)와 ROC (receiver operating characteristics), PESQ (perceptual evaluation of speech quality)를 이용하였다.

실험에 사용된 음성 데이터베이스는 남성 4명, 여성 4명에 의해서 녹음된 456 s 길이의 NTT 데이터

베이스가 사용되었고, 8 kHz로 샘플링 되었다. 음성 데이터 전체의 유성음, 무성음, 무음의 비율은 44.8%, 13.4%, 41.8%로 구성되어 있다. 잡음 환경을 office, babble, car 잡음이 사용되었고, 각 잡음 환경에 대해서 0, 5, 10 , 15 dB SNR의 오염된 음성 파일을 생성하였다. 음성 검출 확률을 위해 10 ms 마다 매뉴얼 마킹된 파일을 만들었다.

제안된 TE 기반의 특징벡터를 이용한 음성 검출 알고리즘의 성능 확인을 위해서 CMAP-SG (conditional maximum *a posteriori* - spectral gradient) 기반의 통계적 음성 검출기와 비교하였다 [11]. 또한, 실생활에서의 사용 가능성에 대해서 확인하기 위해 ITU-T G.729B Appendix III 음성 코덱의 음성 검출기와 성능을 비교하였다 [12].

표 1은 제안한 음성 검출기와 기존 음성 검출기의 전체 오차율을 나타낸다. 전체 오차율은 비음성 프레임을 음성으로 잘못 검출한 확률 (FAR, false acceptance rate)과 음성 프레임을 비음성으로 잘못 검출한 확률 (FRR, false rejection rate)에 의해서 구해진다. 표 1을 통해서 제안한 음성검출기의 전체 오차율이 CMAP-SG에 비해서 비 정상 잡음 환경 (offce, babble)에서 더 우수한 성능을 나타내었다. car 잡음 환경에서는 제안한 알고리즘에 비해서 기존의 CMAP을 적용한 음성 검출기의 성능이 우수한 것을 볼 수 있다. 이는 car와 같은 정상 잡음 환경에서는 잡음 특성이 시간에 대해서 빠르게 변하지 않기 때문에 이전 프레임의 검출 결과를 현재 프레임의 음성 검출 결정 식에 반영하는 CMAP의 효과가 최대화되기 때문이다. 음성 코덱과 비교하였을 때 전체 잡음 환경에서 더 우수한 성능을 보여주었고, 이는 제안한 음성 검출 알고리즘의 성능이 실생활에서 충분히 적용 가능함을 나타낸다.

표 1. 제안한 음성 검출 알고리즘과 기존 음성 검출 알고리즘(CMAP-SG, G.729B App.III)의 전체 오차율 비교

Environments		Methods		
Noise	SNR (dB)	CMAP -SG	G.729B App.III	Proposed
Office	0	16.94	16.28	15.32
	5	15.78	14.87	12.07
	10	10.84	12.84	9.77
	15	8.16	12.11	8.02
Babble	0	22.76	21.55	19.52
	5	16.86	16.24	15.04
	10	10.09	13.11	8.3
	15	8.21	12.45	7.53
Car	0	6.02	23.94	6.82
	5	5.62	22.44	6.53
	10	5.51	21.97	6.11
	15	5.34	20.34	5.86

다음으로 ROC에 대해서 조사하였다. ROC는 FAR과 FRR를 통해서 음성 검출기의 전체 성능을 나타내고, FAR과 FRR의 trade-off 관계를 보여준다. ROC의 해석은 알고리즘의 ROC가 왼쪽 위에 위치할수록 더 우수한 성능을 나타낸다. 그림 3~5는 제안한 알고리즘과 비교 알고리즘의 ROC 커브를 나타낸다. 음성 코덱의 음성 검출은 다양한 특징벡터의 비교구문을 통해서 이루어지기 때문에 실제로 ROC를 그릴 수 없어서 동작포인트로 나타내었고, 제안한 알고리즘의 실생활 적용 가능성에 대해서 직관적으로 비교하기 위해서 나타내었다. 실험결과 제안한 음성 검출 알고리즘의 성능이 통계적 모델 기반의 음성 검출 알고리즘에 비해서 car 잡음 환경을 제외하고는 전체적으로 우수하게 나오는 것을 확인 할 수 있었다.

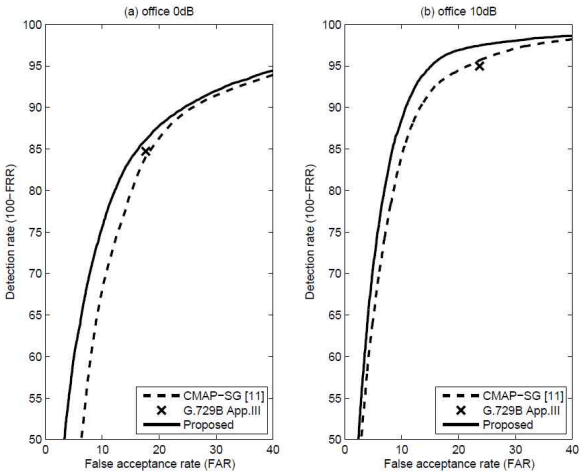

그림 3. Office 잡음에서의 ROC 결과

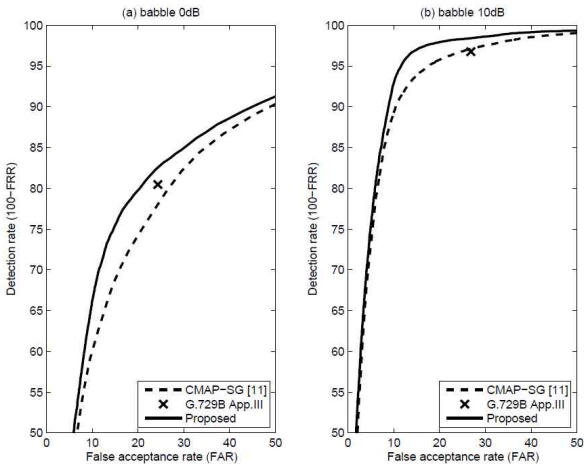

그림 4. Babble 잡음에서의 ROC 결과

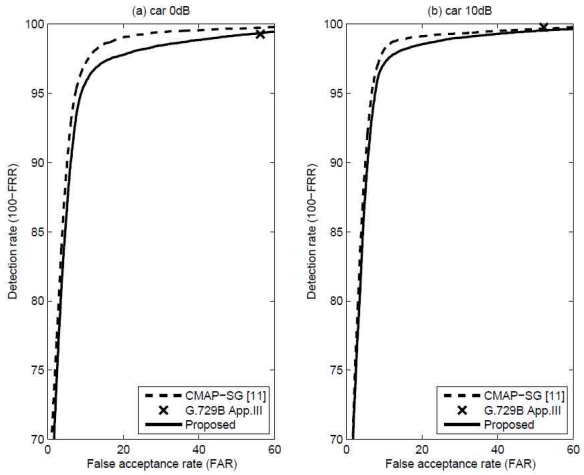

그림 5. Car 잡음에서의 ROC 결과

다음으로 PESQ 테스트를 수행하였다 [13]. PESQ 테스트를 위해서 제안한 알고리즘의 음성 검출 결과를 MMSE 기반의 잡음 제거 알고리즘에 적용하여 잡음이 제거된 음성 파일을 생성하였다. PESQ 실험을 위해서 사용된 음성 데이터베이스는 8 kHz로 샘플링된 8초 길이의 90개의 음성 파일이 사용되었다. 각 음성 파일은 두 개의 의미있는 문장으로 구성되어 있다. 표 2는 PESQ 테스트의 결과를 나타내고 있다. 표 2를 통해서 office, babble 잡음 환경에서 평균 0.03, 0.051의 PESQ 점수 향상을 확인할 수 있고, car 잡음 환경에서는 평균 0.002 정도 줄어든 결과를 볼 수 있었다.

표 2. 다양한 잡음 상황에서의 PESQ 결과 비교

Environments		PESQ	
Noise	SNR (dB)	CMAP-SG	Proposed
Office	0	1.848	1.891
	5	2.193	2.225
	10	2.556	2.583
	15	2.841	2.859
Babble	0	1.982	2.055
	5	2.362	2.425
	10	2.674	2.711
	15	2.965	2.997
Car	0	3.152	3.151
	5	3.443	3.439
	10	3.694	3.692
	15	3.944	3.941

5. 결론

본 논문에서는 잡음 상황에서 강인한 음성 검출 성능을 보여주는 TE 기반의 특징벡터를 이용한 음성 검출 알고리즘을 제안하였다. 구체적으로 TE를 이용하여 구한 파워 스펙트럼 편차와 문턱값과의 비교를 통해서 음성을 검출하였다. 이때, 보다 효과적인 파워 스펙트럼 편차를 구하기 위해서 가중치 요소로 통계적 모델 기반의 우비도 (TE-LR)을 적용하였다. 제안한 음성 검출 알고리즘과 기존 음성 검출 알고리즘 (통계적 모델 기반의 음성 검출기 (CMAP-SG), G.729B App.III)의 객관적 테스트 (전체 오차율, ROC, PESQ)로부터 제안된 음성 검출 알고리즘의 성능이 기존의 음성 검출 알고리즘에 비해서 비 정상 잡음 환경에서 보다 향상된 결과를 나타내었고, 정상 잡음 환경에서는 유사한 성능을 보여주었다.

참 고 문 헌

[1] L. Karray, C. Mokbel and J. Monne, “Solutions for robust speech/non-speech detection in wireless environment,” presented at the IVTTA, Sep. 1988.

[2] TIA/EIA/IS-127, Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems, 1996.

[3] R. J. McAualy and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, pp. 137-145, Apr. 1980.

[4] G. Evangelopoulos, P. Maragos, “Multiband modulation energy tracking for noisy speech detection,” IEEE Trans. ASLP, vol.14, no.6, pp. 2024-2038, 2006.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[6] F. Jabloun, A. E. Cetin and E. Erzin, “Teager energy based feature parameters for speech recognition in car noise,” IEEE Signal Processing Letters, vol. 6, pp. 259-261, 1999.

[7] S.-H. Chen, H.-T.Wu, Y. Chang, T. K. Truong, “Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator,” Pattern Recognition Letters, vol.28, no.11 pp. 1327-1332, 2007.

[8] Y. S. Park, and S. Lee, “Voice activity

detection using global speech absence probability based on teager energy for speech ehancement," IEICE Trans. Inform. System, vol. E95-D, no. 10, Oct. 2012.

[9] 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.

[10] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Processing Letters, vol. 7, no. 5, pp. 108-110, May 2000.

[11] S. K. Kim and J. H. Chang, "Voice activity detection based on conditional MAP criterion incorporating the spectral gradient," Signal Processing 92, pp. 1699-1705, 2012.

[12] ITU-T, Appendix III: G.729 Annex B enhancement in voice-over-IP applications - Option 2, 2005.

[13] ITU-T Recommendation P.862, "Perceptual evaluatioon of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

송 지 현

2007년 2월 인하대학교 전자공학과 학사
2009년 2월 인하대학교 전자공학부 석사
2009년 3월~현재 인하대학교 전자공학부 박사과정

관심분야 : 생체신호처리, 음성 향상, 음성 검출, 패턴 인식

송 영 록

2001년 2월 인천대학교 정보통신공학과 졸업 (학사)
2003년 8월 인천대학교 대학원 정보통신공학과 졸업 (석사)
2009년 2월 인천대학교 대학원 정보통신공학과 졸업 (박사)
2009년 7월 - 현재 인하대학교 정보전자공동연구소 연구교수

관심분야 : Ubiquitous Computing, Semantic Web, Bio-signal Processing

심 현 민

2001년 2월 인하대학교 전자공학과 졸업 (학사)
2003년 2월 인하대학교 대학원 전자공학과 졸업(석사)
2007년 2월 인하대학교 대학원 전자공학과 졸업(박사)
2007년 4월 - 2012년 8월 LIG넥스원 S/W연구센터 수석연구원
2012년 9월 - 현재 인하대학교 정보전자공동연구소 연구교수

관심분야 : implantable rehabilitation engineering, mobile robotics, embedded system design

이 상 민

1987년 2월 인하대학교 전자공학과 학사
1989년 2월 인하대학교 전자공학과 석사
2000년 2월 인하대학교 전자공학과 박사
2006년 6월~현재 인하대학교 전자공학과 부교수

관심분야 : Brain-Machine interface, Bio-Siganl Processing, Psyco-Acoustic