Journal of Korea Multimedia Society Vol. 17, No. 9, September 2014(pp. 1064-1069)
http://dx.doi.org/10.9717/kmms.2014.17.9.1064

우도비를 이용한 적응 밴드 분할 기반의 음성 검출기

김상균[†], 심현민[††], 이상민[†††]

Voice Activity Detection based on Adaptive Band-Partitioning using the Likelihood Ratio

Sang-Kyun Kim[†], Hyeon-Min Shim[††], Sangmin Lee[†††]

ABSTRACT

In this paper, we propose a novel approach to improve the performance of a voice activity detection(VAD) which is based on the adaptive band-partitioning with the likelihood ratio(LR). The previous method based on the adaptive band-partitioning use the weights that are derived from the variance of the spectral. In our VAD algorithm, the weights are derived from LR, and then the weights are incorporated with the entropy. The proposed algorithm discriminates the voice activity by comparing the weighted entropy with the adaptive threshold. Experimental results show that the proposed algorithm yields better results compared to the conventional VAD algorithms. Especially, the proposed algorithm shows superior improvement in non-stationary noise environments.

Key words: Voice Activity Detection, Adaptive Band-Partitioning, Likelihood Ratio

1. 서 론

음성 검출기(voice activity detection, VAD)는 잡음 환경에서 음성이 존재하는 구간을 검출하는 알고리즘으로서 음성 인식, 음성 향상 그리고 음성 부호화 등의 다양한 음성 신호처리 분야에서 사용된다. 예를 들면, 음성 부호화기는 사용 가능한 주파수 대역에 제한되어 있기 때문에 제한된 주파수 대역에서 최대한 많은 정보를 전송하기 위해서는 입력신호의 정보량에 따라 전송률을 가변적으로 부여해야만 한다[1,2]. 이러한 가변 전송률 음성 부호화기(variable bit rate speech codec)에서는 입력 신호에 음성이 있으면 전송률을 높이고 잡음만 존재 하면 전송률을 낮게 선택해야하기 때문에 입력 신호에 음성이 존재하는지 아닌지를 결정하는 음성 검출기는 매우 중요한 기술이다. 또한 음성 향상이나 음성 인식을 할 때, 음성 검출기의 정확도는 전체 시스템의 성능에 큰 영향을 미친다[3,4]. 이러한 이유로 음성 검출기의 성능을 개선시키기 위해 현재까지 다양한 알고리즘들이 연구되고 있다.

초기에 음성 검출을 위한 특징 벡터들은 선형예측 부호화 파라미터, 영교차율, 에너지 레벨, 포만트 모양, 주기성, 캡스트럴 계수 등이 있다[5]. 이러한 특징 벡터들은 현재까지도 이용되고 있으며 에너지 레벨 차이, 영교차율, 스펙트럼 차이들은 국제 표준 음성 부호화기인 ITU-T G.729 Annex. B [6]에 채택되어 사용되고 있다.

이후 통계적 모델 기반의 음성 검출 방법이 제안

※ Corresponding Author : Sangmin Lee, Address: (402-751) Hi-tech 704, InHa Univ., 100 Inharo, Nam-Gu InCheon, Korea, TEL : +82-32-860-7420, FAX : +82-32-868-3654 , E-mail : sanglee@inha.ac.kr
Receipt date : Jul. 1, 2014, Revision date : Jul. 23, 2014
Approval date : Aug. 1, 2014

[†] Division of Electronic Engineering, Inha Univ. (E-mail : greenwhity@nate.com)
[††] Division of Electronic Engineering, Inha Univ. (E-mail : hmshim@inha.ac.kr)
[†††] Division of Electronic Engineering, Inha Univ.
※ This research was supported by Seoul R&BD Program (SS100022).

되었고 발표 당시 성능이 매우 우수한 것으로 알려져 있다[7]. 이 음성 검출기는 Ephraim과 Malah의 연구에서 제안한 MMSE(minimum mean square error) 기반의 음성 향상 기법 [8]에 사용된 음성의 존재와 부재에 대한 통계적 모델을 가우시안 분포로 가정하여 우도비 테스트(likelihood ratio test, LRT)에 적용한 것이다. 또한 직접 구할 수 없는 음성 파라미터인 사전 신호 대 잡음비(*a priori* signal-to-noise ratio, *a priori* SNR)를 decision-directed(DD) 기법을 이용하여 추정을 한다.

Wu의 연구에서 제안한 방법은 가중치가 적용된 엔트로피(entropy)를 이용하여 적응적으로 주파수 밴드를 선택하는 것이다[9]. 엔트로피만 사용했을 때의 문제점은 주파수 밴드별 에너지의 편차가 같아도 주파수 영역에서 신호의 위치가 다를 수 있다는 것이다. 이 문제를 해결하기 위해 인접한 두 주파수 밴드와의 분산을 구하고 이를 가중치로 사용하여 엔트로피에 부과한다. 가중치를 적용한 엔트로피와 적응적 주파수 밴드 선택은 음성 구간을 검출하는 과정에서 잡음의 영향을 줄여주는 효과를 얻을 수 있다. 또 다른 연구에서는 낮은 SNR 환경에서 음성 검출 성능을 높이고자 엔트로피를 이용하여 음성 검출을 한 후 퍼지 소속도 천이 c-means 클러스터링 방법을 제안하였으며 우수한 성능을 보였다[10]. 하지만 백색잡음 환경에서만 성능비교가 이루어져 추가적인 실험 결과가 필요하다.

기존의 적응적 주파수 밴드 선택 기반의 음성 검출 방법은 가중치를 이웃한 주파수 밴드와의 분산을 이용하였다[9]. 유성음의 경우 음성이 존재하는 주파수 대역은 인접한 주파수 밴드와 비교하였을 때 상대적으로 높은 에너지를 가진다. 하지만 분산의 경우 주파수 밴드가 이웃한 주파수 밴드보다 상대적으로 에너지가 낮은 경우에도 높은 값이 도출된다. 따라서 분산을 가중치로 사용하여 음성 구간을 검출할 경우 주파수 위치에 대한 정보가 왜곡되어 성능 저하의 원인이 된다. 따라서 본 논문에서는 음성 검출기의 성능을 향상시키기 위해 주파수 밴드별 우도비를 이용하여 가중치를 얻은 후 이를 이용하여 음성 구간을 검출 하는 음성 검출기 알고리즘을 제안한다. 제안된 음성 검출 방법은 다양한 잡음 환경에서 기존의 음성 검출 알고리즘들과 비교하였으며 향상된 성능을 보였다.

본 논문의 2장에서는 제안한 알고리즘을 설명하기 전에 기존의 가중치를 적용한 적응 밴드 분할 기법에 대해 소개하고 3장에서는 우도비를 적용한 새로운 적응 밴드 분할 알고리즘에 대해 논한다. 4장에서는 기존의 음성 검출 방법과 성능 비교를 실험 결과를 통해 보여주며, 마지막으로 5장에서 결론을 맺어 본 논문을 마친다.

2. 기존의 가중치를 적용한 적응 밴드 분할

적응 밴드 분할을 하기 전에 먼저 엔트로피에 적용할 가중치를 구해야 하며 그 과정은 다음과 같다. 먼저 시간 영역에서 입력된 신호를 $y(t)$라 놓고 여기서 t는 샘플링 인덱스(sampling index)를 나타낸다. 주어진 입력 신호 $y(t)$를 이산 퓨리에 변환(discrete Fourier transform, DFT)하여 주파수 영역으로 변환하면 $Y(k,l)$을 얻으며 여기서 k는 주파수 밴드를 나타내고 l은 프레임 인덱스를 나타낸다.

각 주파수 밴드에서의 확률은 다음과 같이 얻는다.

$$P_b(k,l) = \frac{Y(k,l)}{\sum_{k=1}^{m} Y(k,l)} \tag{1}$$

여기서 아래첨자 b는 밴드를 나타내고 m은 전체 주파수 밴드의 개수이다. 주파수 밴드별 기여도를 결정해 주는 가중치 $W(k,l)$은 밴드별 확률로부터 유도되며 다음과 같다[9].

$$W(k,l) = var\left[P_{offset}(k-1,l), P_{offset}(k,l), P_{offset}(k+1,l)\right] \tag{2}$$

여기서 $var[\bullet]$은 분산 값을 의미하고 $P_{offset}(k,l)$은 각 주파수 밴드별 에너지의 일반화를 나타내며 다음과 같다.

$$P_{offset}(k,l) = \frac{\min\{P_b(l)\}}{P_b(k,l)} \tag{3}$$

여기서 $\min\{\bullet\}$은 값들 중 최소값을 출력하는 연산자이다. 이렇게 구한 가중치 $W(k,l)$은 양 옆의 이웃 밴드들과 에너지 차이가 많으면 크고 차이가 없으면 작게 나오게 된다.

잡음의 영향을 줄이기 위해 모든 밴드를 사용하는 것이 아니라 적응적으로 변하는 유용한 밴드의 수 $UB(l)$을 이용하여 사용할 밴드의 수를 결정하며 아래와 같이 얻는다[9].

$$UB(l)=\begin{cases}30 & A(l)<5\\ \left[(4-30)\times\dfrac{A(l)}{25-5}+36.5\right] & 5<A(l)<25\\ 4 & otherwise\end{cases} \tag{4}$$

여기서 유용한 밴드수를 결정하는 파라미터 $A(l)$은 아래와 같이 주어진다[9].

$$A(l)=-\log\left[\frac{\min\{Y_b(k,l)\}}{\sum_{k=1}^{m}Y_b(k,l)}\right]. \tag{5}$$

최종적으로 가중치와 적응 밴드를 적용한 엔트로피는 아래와 같다.

$$T_b=\sum_{k=1}^{UB(l)}W(k,l)P_b(k,l)\log(1/P_b(k,l)) \tag{6}$$

위에서 구한 엔트로피를 문턱값과 비교하여 음성 구간을 결정한다.

3. 제안된 우도비 기반의 적응 밴드 분할

통계적 모델 기반의 우도비를 구하기 위해서는 l번째 프레임에서 k번째 신호가 잡음만 존할 경우와 음성과 함께 존재할 경우를 각각 가설 $H_0(k,l)$, $H_1(k,l)$으로 표현하며 아래와 같이 나타낸다.

$$H_0(k,l):\ Y(k,l)=N(k,l) \tag{7}$$

$$H_1(k,l):\ Y(k,l)=X(k,l)+N(k,l) \tag{8}$$

여기서 $Y(k,l)$은 2장에서와 마찬가지로 입력 신호이고 $X(k,l)$과 $N(k,l)$은 음성 신호와 잡음 신호의 이산 퓨리에 변환 계수이다.

음성과 잡음 신호의 스펙트럼 분포가 복소 가우시안 분포를 따른다고 가정하면, 가설 $H_0(k,l)$, $H_1(k,l)$을 조건부 확률로 적용한 확률밀도함수는 아래와 같다[7].

$$p(Y(k,l)|H_0)=\frac{1}{\pi\lambda_n(k,l)}\exp\left\{-\frac{|Y(k,l)|^2}{\lambda_n(k,l)}\right\} \tag{9}$$

$$p(Y(k,l)|H_1)=\frac{1}{\pi\left[\lambda_n(k,l)+\lambda_x(k,l)\right]}\exp\left\{-\frac{|Y(k,l)|^2}{\lambda_n(k,l)+\lambda_x(k,l)}\right\} \tag{10}$$

여기서 $\lambda_x(k,l)$와 $\lambda_n(k,l)$는 각 프레임에서 주파수 밴드별 음성과 잡음의 분산이며, 이때 k번째 주파수 밴드에 대한 우도비는 아래와 같이 구한다.

$$\Lambda(k,l)\equiv\frac{p(Y(k,l)|H_1)}{p(Y(k,l)|H_0)}=\frac{1}{1+\xi(k,l)}\exp\left\{\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}\right\} \tag{11}$$

여기서 $\xi(k,l)$은 사전 신호대 잡음비(*a priori* SNR: *a priori* signal-to-noise ratio)이고 $\gamma(k,l)$은 사후 신호대 잡음비(*a posteriori* SNR)이며 다음과 같이 얻을 수 있다[8].

$$\xi(k,l)=\frac{\lambda_x(k,l)}{\lambda_n(k,l)} \tag{12}$$

$$\gamma(k,l)=\frac{|Y(k,l)|^2}{\lambda_n(k,l)} \tag{13}$$

여기서 사후 신호대 잡음비 $\gamma(k,l)$은 음성 부재 구간에서 갱신되는 신호로부터 얻은 잡음 분산 $\lambda_n(k,l)$을 이용하여 추정하며, 사전 신호대 잡음비 $\xi(k,l)$은 DD(decision-directed) 기법을 이용하여 다음과 같이 추정한다[8].

$$\hat{\xi}(k,l)=\alpha\frac{|\hat{X}(k,l-1)|^2}{\lambda_n(k,l-1)}+(1-\alpha)Q[\gamma(k,l)-1] \tag{14}$$

여기서 $|\hat{X}(k,l-1)|$ 은 이전 프레임의 k번째 주파수 밴드에서 추정된 음성 신호의 스펙트럼 성분의 크기이며, MMSE를 기반으로 구한다[7]. 또한 α는 가중치 파라미터이며 일반적으로 [0.95, 0.99] 범위에서 값을 결정한다. $Q[\bullet]$ 연산자는 다음과 같이 정의된다.

$$Q[x]=\begin{cases}x & \text{if}\quad x\geq 0\\ 0 & otherwise.\end{cases} \tag{15}$$

우도비를 이용한 새로운 가중치는 아래와 같이 구한다.

$$W_\Lambda(k,l)=\frac{\Lambda(k,l)}{\sum_{k=1}^{m}\Lambda(k,l)} \tag{16}$$

여기서 새로운 가중치 $W_\Lambda(k,l)$은 각 프레임에서 우도비가 높은 주파수 밴드에 큰 값을 부여한다.

최종적으로 새로운 가중치를 적용한 적응 밴드의 엔트로피 $T_\Lambda(l)$을 문턱값 $\eta(l)$과 비교하여 음성 활동 구간을 검출하게 되며 다음과 같이 표현된다.

$$T_\Lambda(l)=\sum_{k=1}^{UB(l)}W_\Lambda(k,l)P_b(k,l)\log(1/P_b(k,l))\underset{H_0}{\overset{H_1}{\gtrless}}\eta(l) \tag{17}$$

여기서 문턱값은 아래와 같이 구한다.

$$\eta(l) = [\eta(l-1) + \alpha_T(l) \cdot T_\Lambda(l)]/2 \quad (18)$$

여기서 $\alpha_T(l)$은 음성 부재 확률에 의존하는 가중치 값으로 현재 프레임이 음성일 확률이 클수록 0에 가까운 값을 갖는다.

4. 실험 결과 및 고찰

본 논문에서 제안한 새로운 음성 검출 방법의 성능을 평가하기 위해 P_e(probability of total error), P_m(probability of miss) 그리고 P_{fa}(probability of false alarm)를 측정하였다. 통계적 모델 기반의 음성 검출법, 기존의 분산 기반의 밴드분할 알고리즘 그리고 실제 사용 가능성을 확인하기 위해서 G.729B 음성코덱[6]과 음성 검출 성능을 비교하였다. 실험에 사용된 데이터는 성능 평가 비교를 위해 사용된 음성 데이터의 길이를 고려하여 각각 4명의 젊은 남성, 여성화자가 영어 문장을 각각 57초씩 말하였으며, 이 데이터들을 모두 합하여 총 456초의 음성을 8kHz로 샘플링 하였다. 또한 평가를 위해 깨끗한 음성 데이터에 음성과 비음성 부분을 10 ms마다 수동으로 표시하였다. 음성 데이터의 음성 구간은 총 57.1%로 유성음 44.0%, 무성음 13.1%로 구성되었으며 잡음환경을 만들기 위해 White, Babble, Office, Street 잡음을 5, 10 그리고 15 dB SNR로 각각 456초의 깨끗한 음성 데이터에 더하여 사용하였다. Fig. 1의 제일 위에는 Babble 5 dB SNR 입력 신호의 파형을 보여주며, 두 번째는 실제 음성 구간을 쉽게 알 수 있도록 깨끗한 음성파일에서 얻은 매뉴얼을 보여준다. 그다음 세 번째는 기존의 분산을 가중치에 적용한 엔트로피이며 마지막 네 번째는 제안된 우도비를 가중치로 사용한 엔트로피를 보여준다. 14 s 이후의 구간을 보면 기존의 방법은 제안한 방법에 비해 잡음만 존재하는 구간에서도 엔트로피 값이 높게 나오는 것을 확인할 수 있으며 이는 음성 검출기의 성능을 저하시키는 원인이 된다.

Table 1은 위에서 설명한 456초의 잡음 섞인 데이터를 사용하여 기존의 음성 검출 알고리즘과 제안된 음성 검출 알고리즘의 P_e, P_m, P_{fa}를 나타낸 것이다. Table 1을 보면 기존의 분산을 가중치로 사용한 밴드 분할 알고리즘보다 제안한 방법의 P_e가 White 잡음

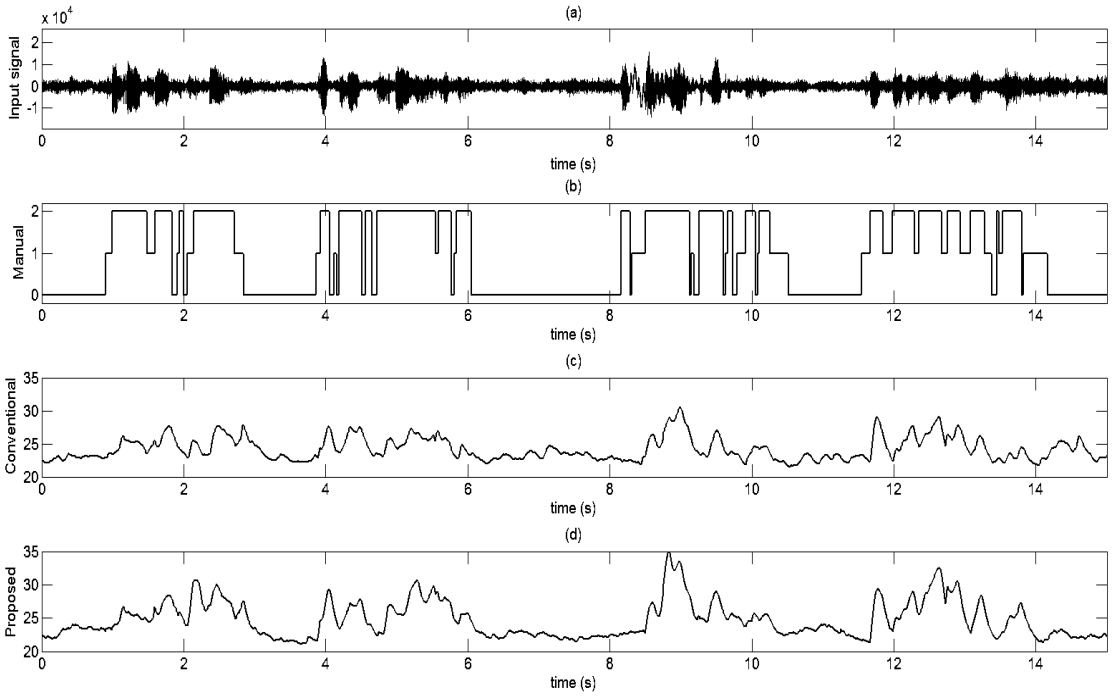


Fig. 1. (a) Waveform of the test file (Babble noise, SNR=5 dB) (b) Manual VAD (silence=0, unvoiced=1, voiced=2) (c) Variance-based entropy of conventional method (d) (c) LR-based entropy of proposed method

Table 1. Comparison of voice activity detection probability of error (P_e), probability of miss (P_m) and false alarm probability (P_{fa}) among the method of the statistical model-based, the conventional method and the proposed technique.

		Statistical Model-based			Conventional Weight-based			G.729 Annex B			Proposed		
Noise	SNR (dB)	P_e	P_m	P_{fa}	P_e	P_m	P_{fa}	P_e	P_m	P_{fa}	P_e	P_m	P_{fa}
White	5	12.8	15.7	9.1	11.2	9.7	13.2	25.2	42.8	0.7	9.9	6.5	14.4
	10	10.3	12.5	7.1	9.4	8.9	10.1	17.4	29.1	0.9	7.4	5.5	10.0
	15	9.1	10.5	6.5	8.6	8.1	9.4	12.9	20.0	3.1	6.8	5.2	9.1
Babble	5	25.7	34.0	15.4	22.2	14.0	33.5	27.8	29.9	24.7	19.6	12.2	30.0
	10	22.3	21.0	23.9	18.1	11.3	27.4	22.6	22.2	23.2	16.6	9.9	25.9
	15	17.9	16.9	19.2	16.7	10.9	24.7	18.6	13.9	25.2	14.3	11.4	18.4
Office	5	18.1	16.7	19.9	17.7	10.6	27.6	26.5	28.7	23.4	16.3	9.1	26.3
	10	15.6	13.4	17.8	14.3	9.5	21.0	22.7	22.3	23.4	12.7	6.9	20.8
	15	13.8	11.4	16.0	13.1	7.7	20.7	19.3	17.3	22.0	11.1	5.7	18.5
Street	5	16.4	12.1	22.3	19.8	14.9	26.7	23.1	23.7	22.1	14.1	10.8	18.7
	10	14.3	4.9	27.3	15.4	12.2	19.7	19.6	19.1	20.1	10.9	7.2	15.9
	15	12.5	2.6	26.2	14.0	10.4	19.0	15.8	14.1	18.0	9.7	6.6	14.1

[1.3, 2], Babble 잡음 [1.5, 2.6], Office 잡음 [1.4, 2], Street 잡음 [4.3, 5.7] 만큼 각각 향상되었다. 정상 잡음인 White 잡음보다 비정상 잡음인 나머지 잡음에서 성능 향상이 상대적으로 높은 것을 확일 할 수 있는데 이는 기존의 분산을 가중치로 사용했을 때보다 우도비를 가중치로 사용한 것이 비정상 잡음에 강인하다는 것을 보여준다.

5. 결 론

본 논문에서는 음성 검출기의 성능을 향상시키기 위해 우도비를 이용하여 효율적인 주파수 밴드를 적응적으로 결정하는 밴드 분할 방법을 제안하였다. 기존의 방법은 가중치를 구할 때 분산을 이용하였는데 이는 비정상 잡음에서 성능 저하의 원인으로 작용하였다. 이러한 단점을 보완하기 위해 제안된 알고리즘에서는 우도비를 사용하여 가중치를 도출하여 음성 검출 결정식에 부과하였다.

제안된 알고리즘의 성능 평가를 위해 P_e, P_m 그리고 P_{fa} 값을 Table 1에서 기존의 음성 검출 알고리즘들과 비교하였고 Fig. 1에서는 기존 방법의 성능 원인을 보기위해 분산과 우도비 기반의 가중치를 적용한 엔트로피를 보여주었다. Table 1과 Fig. 1에서 본 것처럼 제안된 음성 검출 알고리즘의 성능이 우수하다는 것을 알 수 있다.

REFERENCE

[1] Y. Gao, E. Shlomot, A. Benyassine, J. Thyssen, Huan-yu Su, and C. Murgia, "The SMV Algorithm Selected by TIA and 3GPP2 for CDMA Applications," *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 2, pp. 709-712, 2001.

[2] 3GPP2 Spec., *Source-controlled Variable-rate Multimedia Wideband Speech Codec (VMR-WB), Service Option 62 and 63 for Spread Spectrum Systems*, 3GPP2-C.S0052-A, v.1.0, 2005.

[3] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Improved Voice Activity Detection based on a Smoothed Statistical Likelihood Ratio," *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 2, pp. 7-11, 2001.

[4] J.H. Song and S.M. Lee, "Voice Activity Detection based on Generalized Normal-Laplace Distribution Incorporating Conditional MAP," *IEICE Transactions on Information and Systems*, Vol. E96-D, No. 12, pp. 2888-2891, 2013.

[5] Y.S. Park and S. Lee, "Voice Activity Detection using Global Speech Absence Probability based on Teager Energy for Speech Enhancement," *IEICE Transactions on Information and Systems,* Vol. E95-D, No. 10, pp. 2568-2571, 2012.

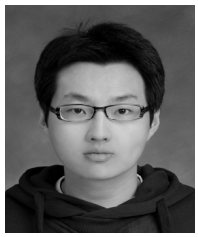
[6] ITU-T Rec. G.729, *Annex B, A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70*, 1996.

[7] 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, 1999.

[8] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-square Error Short-time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No. 6, pp. 1190-1121, 1984.

[9] B.F. Wu and K.C. Wang, "Robust Endpoint Detection Algorithm based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 13, No. 5, pp. 762-775, 2005.

[10] G.H. Lee, Y.J. Lee, J.H. Cho and M.N. Kim, "Voice Activity Detection Algorithm using Fuzzy Membership Shifted C-means Clustering in Low SNR Environment," *Journal of Korea Multimedia Society,* Vol. 17, No. 3, pp. 312-323, 2014.



김 상 균

2008년 2월 인하대학교 전자공학과 학사
2010년 10월 인하대학교 전자공학부 석사
2013년 3월~현재 인하대학교 전자공학부 박사과정

관심분야 : Speech Signal Processing, Acoustic 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 인하대학교 전자공학과 박사

2006년 6월~현재 인하대학교 전자공학과 부교수
관심분야 : Brain-Machine interface, Bio-Siganl Processing, Psyco-Acoustic