한국음향학회지 제26권 제8호 pp. 397~402 (2007)

우도비 특징 벡터를 이용한 SVM 기반의 음성 검출기

Voice Activity Detection Based on SVM Classifier Using Likelihood Ratio Feature Vector

조 규 행*, 장 준 혁*, 강 상 기**
(Q-Haing Jo*, Joon-Hyuk Chang*, Sangki Kang**)
*인하대학교 전자전기공학부, **삼성전자 정보통신총괄 통신연구소
(접수일자: 2007년 7월 5일; 수정일자: 2007년 8월 22일; 채택일자: 2007년 9월 20일)

본 논문에서는 기존의 통계적 모델 기반의 음성 검출기의 성능 향상을 위해 이진 분류에 우수한 support vector machine (SVM)을 도입한다. 기존의 통계적 모델 기반 음성 검출기의 경우 음성의 존재와 부재에 대한 가설로부터 각각의 통계적 모델을 세워 입력 데이타에 의해 결정된 각 주파수 채널별 우도비 (likelihood ratio)를 단순히 기하 평균을 취하여 문턱값과 비교, 음성 검출 여부를 판단한다. 제안된 음성 검출기는 기존의 기하 평균을 이용한 결정식을 대신하여 분류 오류 확률이 최소화되도록 각 주파수 채널별 우도비를 SVM의 특징 벡터로 적용한다. 제안된 SVM 기반의 통계적 모델 음성 검출기는 기존의 LRT를 이용한 음성 검출기 및 SVM 기반의 음성 검출기들과 비교하여 다양한 잡음 환경에서 우수한 성능을 나타낸다.

핵심용어: 음성 검출기, Support vector machine, 통계적 모델, 우도비

투고분야: 음성처리 분야 (2.4)

In this paper, we apply a support vector machine (SVM) that incorporates an optimized nonlinear decision rule over different sets of feature vectors to improve the performance of statistical model-based voice activity detection (VAD). Conventional method performs VAD through setting up statistical models for each case of speech absence and presence assumption and comparing the geometric mean of the likelihood ratio (LR) for the individual frequency band extracted from input signal with the given threshold. We propose a novel VAD technique based on SVM by treating the LRs computed in each frequency bin as the elements of feature vector to minimize classification error probability instead of the conventional decision rule using geometric mean. As a result of experiments, the performance of SVM-based VAD using the proposed feature has shown better results compared with those of reported VADs in various noise environments.

Key words: Voice activity detection, Support vector machine, Statistical model, Likelihood ratio

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

음성과 비음성 구간을 검출하는 음성 검출기 (voice activity detector, VAD)는 다중 접속 기술에서 한정된 주파수 대역을 효율적으로 사용하기 위한 가변 전송률 부호화기의 실현을 위해 필수적인 부분을 차지하고 있다 [1]. 이와 관련하여 다양한 형태의 알고리즘이 제안되었으며, 에너지 차이, 영교차율, 스펙트럼 차이 등의 특징을 이용한 다양한 알고리즘들이 연구되었다 [2]. 특히 Ephraim과 Malah의 연구에서 시작된 minimum mean square error (MMSE) 기반의 음성 향상 기법에 사용된 음성의 존재와 부재에 대한 통계적 모델을 우도비 테스트 (likelihood ratio test, LRT) 기반의 음성 검출기에 적용한 것이 매우 우수한 성능을 가진 것으로 알려져 있다 [3-10].

한편, 최근의 음성 검출기의 성능 개선을 위한 새로운 시도로서 기존의 음성 파라미터를 이진 분류에 뛰어난 성능을 보이는 support vector machine (SVM)에 적용하여 우수한 성능의 음성 검출기를 제시하였다 [11-12].

본 논문에서는 음성의 통계적 모델 기반의 음성 검출기의 성능 향상을 위해 기존의 방법에서 제시된 각 주파수 채널별 우도비의 단순한 기하 평균과 문턱값을 비교하는

책임저자: 장 준 혁 (changjh@inha.ac.kr)
420-751 인천시 남구 용현동 253 인하대학교 전자전기공학부
(전화: 032-860-7423; 팩스: 032-868-3654)

방법 대신, 주파수별 우도비를 SVM의 특징 벡터로 사용하여 최적화된 이진분류기법에 의존하는 새로운 방식을 제안하였으며, 다양한 잡음환경에서 기존의 통계적 모델 기반의 음성검출기와 성능을 비교하였다.

II. 통계 모델 기반의 음성 검출기의 이해

시간축 상에서 원래의 음성신호 $x(t)$에 잡음신호 $d(t)$가 인가된 입력신호 $y(t)$을 DFT (discrete Fourier transform)를 통해 주파수 축으로 변환되어 아래와 같이 표현된다.

$$Y_k(n) = X_k(n) + D_k(n) \tag{1}$$

여기서 $Y_k(n)$은 n번째 프레임에서의 k번째 주파수 성분이다. 주어진 가설 H_0, H_1이 각각 음성의 부재와 존재를 표현한다고 하면 각 주파수 채널별로 다음과 같이 기술된다.

$$H_0: \text{speech absent} \;:\; Y_k(n) = D_k(n) \tag{2}$$

$$H_1: \text{speech present} \;:\; Y_k(n) = X_k(n) + D_k(n). \tag{3}$$

음성과 잡음신호의 스펙트럼이 복소 가우시안 분포를 따른다는 가정으로부터 가설 H_0와 H_1을 조건으로 한 확률밀도함수는 아래와 같이 주어진다 [5].

$$p(Y_k|H_0) = \frac{1}{\pi\lambda_{d,k}} \exp\left\{-\frac{|Y_k|^2}{\lambda_{d,k}}\right\} \tag{4}$$

$$p(Y_k|H_1) = \frac{1}{\pi[\lambda_{d,k}+\lambda_{x,k}]} \exp\left\{-\frac{|Y_k|^2}{\lambda_{d,k}+\lambda_{x,k}}\right\} \tag{5}$$

여기서 $\lambda_{x,k}$와 $\lambda_{d,k}$는 각각 채널별 음성과 잡음의 분산이며, 이 때 k번째 주파수 밴드에 대한 우도비는 아래와 같이 구한다.

$$\Lambda_k \equiv \frac{p(Y_k|H_1)}{p(Y_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{6}$$

여기서 $\xi_k = \lambda_{x,k}/\lambda_{d,k}$와 $\gamma_k = |Y_k|^2/\lambda_{d,k}$는 각각 a priori signal-to-noise ratio (SNR)와 a posteriori SNR이다 [5]. 음성 부재 구간에서 갱신되는 잡음 신호로부터 구한 잡음 분산 $\lambda_{d,k}$를 이용하여 a posteriori SNR γ_k를 추정하며, 또한 a priori SNR ξ_k는 decision-directed (DD) 방식을 이용하여 아래와 같이 추정한다 [3].

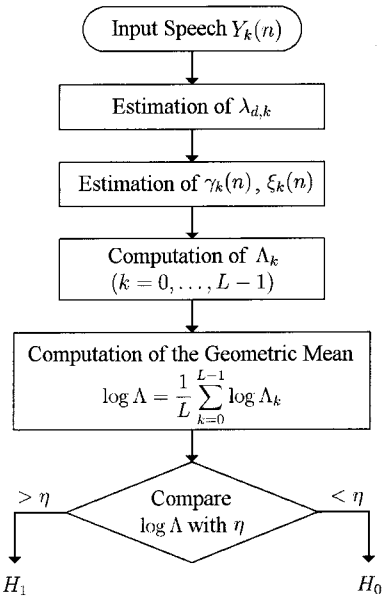

그림 1. 통계적 모델에 기반하여 우도비의 기하 평균을 이용한 음성 검출기의 결정 순서도

Fig. 1. The decision flowchart of VAD using geometric mean of likelihood ratio based on a statistical model.

$$\hat{\xi}_k(n) = \alpha \frac{|\hat{X}_k(n-1)|^2}{\lambda_{d,k}(n-1)} + (1-\alpha) P[\gamma_k(n) - 1] \tag{7}$$

여기서 $|\hat{X}_k(n-1)|$은 이전 프레임에서 추정된 음성 신호의 k번째 스펙트럼 성분의 크기에 대한 추정치이며, MMSE에 기반하여 구한다 [5]. 또한 α는 가중치 값이며, 연산자 $P[\cdot]$은 아래와 같이 정의된다.

$$P[x] = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

기존의 통계적 모델 기반의 음성 검출기에 대한 결정식은 각각의 주파수 채널에서 구해진 우도비를 기하 평균하여 아래와 같이 음성 검출 여부를 판단한다 [4-10].

$$\log \Lambda = \frac{1}{L}\sum_{k=0}^{L-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

여기서 L은 전체 주파수 대역의 개수이며, η는 음성 검출 문턱값이다. 그림 1은 기존의 음성의 통계적 모델을 바탕으로 한 음성 검출기의 결정 순서도를 보여주고 있다.

III. SVM 기반의 제안된 음성 검출 기법

Vapnik이 제안한 SVM은 통계적 학습 이론에 기반을 둔 패턴 분류기로써 분류 오류 확률을 최소화하는 구조적 위험 최소화 (structural risk minimization) 방법에 기초하고 있다 [13]. 선형적으로 분류 가능한 데이터에 대한 이진 분류에 있어 두 개의 클래스를 분류할 수 있는 무수히 많은 초평면 (hyperplane) 중 클래스의 가장 가까운 점들과 마진이 최대가 되는 최적 초평면을 구함으로써 높은 일반화 성능을 기대할 수 있다.

본 논문에서는 식 (6)에서 사용된 각 주파수 채널별 우도비 $\Lambda_k\,(k=0,\cdots,L-1)$를 식 (9)와 같이 기하평균을 구하여 문턱값과 비교하는 기존의 방법 대신 그림 2와 같이 우도비를 특징으로 사용하는 SVM에 기반한 음성 검출기를 제안한다. 그림 2의 support vector를 구성하기 위하여 훈련용 음성 데이터로부터 특징 벡터 Λ를 추출한다. 학습 데이터가 $(\Lambda_1, z_1), \cdots, (\Lambda_l, z_l)$, $\Lambda \in R^n, z \in \{1,-1\}$과 같이 주어졌을 때 초평면에 대한 방정식은 $(\mathbf{w} \cdot \Lambda) + b = 0$이다. 여기서 w는 가중치 벡터, b는 바이어스를 나타낸다. 이때 마진을 최대화하기 위해서는 아래의 조건을 만족해야 한다.

$$\text{Minimize}: \Phi(\mathbf{w}) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) \tag{10}$$

$$\text{Subject to}: \{(\mathbf{w} \cdot \Lambda_i) + b\} z_i \geq 1 \text{ for } i = 1,2,\ldots,l. \tag{11}$$

여기서 식 (10)의 최소화 문제를 해결하기 위해 식(10)과 식 (11)을 결합하여 아래와 같이 Lagrange multiplier α_i를 포함한 Lagrangian 함수를 구한다.

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^{l} \alpha_i [\{(\mathbf{w} \cdot \mathbf{x}_i) + b\} z_i - 1], \quad \alpha_i \geq 0, \ \forall i. \tag{12}$$

여기서 KKT (Karush-Kuhn-Tucker) 조건을 식(12)에 적용하여 식 (13)과 같이 α에 대한 수식으로 전개할 수 있으며, 식 (14)을 만족하면서 $Q(\alpha)$를 최대화하는 최적화 문제가 된다. 이 때 학습 데이터에 대하여 QP (quadratic programming)를 이용하여 α_i를 구한다.

$$\text{Maximize}: Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j z_i z_j (\Lambda_i \cdot \Lambda_j) \tag{13}$$

$$\text{Subject to}: \sum_{i=1}^{l} \alpha_i z_i = 0, \quad \alpha_i \geq 0, \ \forall i. \tag{14}$$

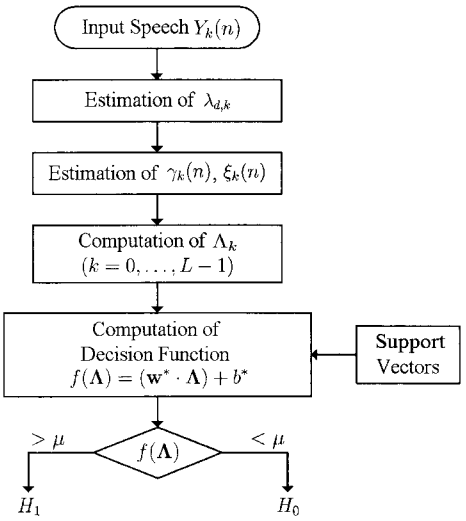

그림 2. 제안된 SVM 을 이용한 음성 검출기의 결정 순서도
Fig. 2. The decision flowchart of the proposed SVM-based VAD.

$Q(\alpha)$를 최대화하는 α_i^*와 KKT 조건에서 유도된 식으로부터 최적 가중치 벡터 $\mathbf{w}^*$와 바이어스 b^*를 구한다 [13].

본 논문에서 support vector를 구성하기 위한 훈련 데이터로 사용된 음성 데이터는 각각 4명의 남성, 여성화자가 말한 음성을 8 kHz로 샘플링 하였으며, 총 226초 길이의 깨끗한 음성 데이터를 음성과 비음성 부분으로 10 ms마다 수동으로 표시하여 클래스 z를 구성하였고, 또한 잡음 환경을 고려하여 vehicular, babble, street, white 잡음을 각각 5와 25 dB 사이의 여러 SNR에 대해 부과하였다.

그림 2과 같이 입력 신호가 주어졌을 때 실시간으로 특징 벡터 Λ를 추출하여 아래와 같은 판별함수를 바탕으로 음성과 비음성으로 분류한다.

$$f(\Lambda) = (\mathbf{w}^* \cdot \Lambda) + b^* = \sum_{i=1}^{l} \alpha_i^* z_i (\Lambda_i^* \cdot \Lambda) + b^* \underset{H_0}{\overset{H_1}{\gtrless}} \mu \tag{15}$$

여기서 Λ_i^*는 우도비의 support vector를 나타낸다.

일반적으로 입력 데이터의 경우 명확하게 선형 분류가 되지 않는 경우가 대부분이며, 선형 분류가 불가능한 경우 아래와 같은 커널 함수 K를 이용하여 고차원 공간으로 사상시켜 고차원 특징 공간에서의 선형 분류를 적용한다 [14].

$$K(\Lambda_i^*, \Lambda) = \Phi(\Lambda_i^*) \cdot \Phi(\Lambda). \tag{16}$$

식 (16)에서 사상함수 Φ가 존재할 수 있는 커널 함수 K가 주어진 경우 판별 함수는 아래와 같이 최종적으로 구한다.

$$f(\Lambda)=\sum_{i=1}^{l}\alpha_i^* z_i K(\Lambda_i^*,\Lambda)+b^* \underset{H_0}{\overset{H_1}{\gtrless}} \mu. \tag{17}$$

실제로 linear 커널을 이용한 선형 학습과 radial basis function (RBF) 커널을 사용한 비선형 SVM 학습을 실시한다 [11].

IV. 실험 결과 비교 및 분석

본 논문에서 제안된 SVM을 이용한 음성 검출기의 성능을 평가하기 위해 기존의 우도비 테스트를 이용한 통계적 모델 기반의 음성 검출기의 성능과 receiver operating characteristics (ROC) 곡선을 이용하여 비교하였다 [5]. 실험에 사용된 데이터는 기존의 음성 검출 알고리즘에서 성능 비교를 위해 사용된 음성 데이타의 길이를 고려하여 각각 4명의 남성, 여성화자가 말한 총 349초의 음성을 8 kHz로 샘플링하였다. 또한 평가를 위해 깨끗한 음성 데이터에 음성과 비음성 부분을 10 ms마다 수동으로 표시하였다. 분류된 음성 데이터의 음성 구간은 총 56.74% (19,796 frames)로 유성음 44.0% (15,346 frames), 무성음 12.7% (4,450 frames)로 구성되었다. 또한 일반적인 음성 검출기의 성능 실험에 사용되는 vehicular, babble, street, white 및 훈련 데이터와의 종속성을 보기 위해 훈련 데이터에서 사용하지 않은 F16 잡음이 5, 10 dB 그리고 20 dB SNR로 부과되었다.

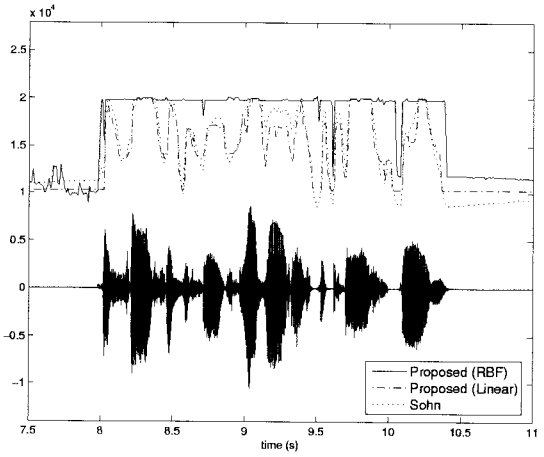

그림 3. Vehicular 잡음 5 dB SNR에서의 각 음성 검출기의 출력 결과 비교
Fig. 3. Comparison of VAD output for vehicular noise (SNR = 5 dB).

그림 3은 vehicular 잡음 SNR 5 dB에서의 각 VAD 출력 결과를 나타내고 있다. 비교의 편의를 위해 vehicular 잡음 5 dB SNR이 부가된 음성 파형 대신 깨끗한 음성 파형 도시되어 있다. 선형 SVM의 경우 음성 구간에서의 VAD 결과는 기존의 VAD와 유사한 결과를 보여주는 반면,

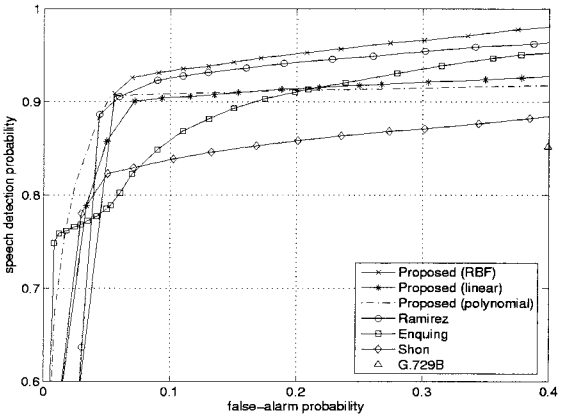

그림 4. Vehicular 잡음 5 dB SNR에서의 ROC 곡선
Fig. 4. ROC curves for vehicular noise (SNR = 5 dB).

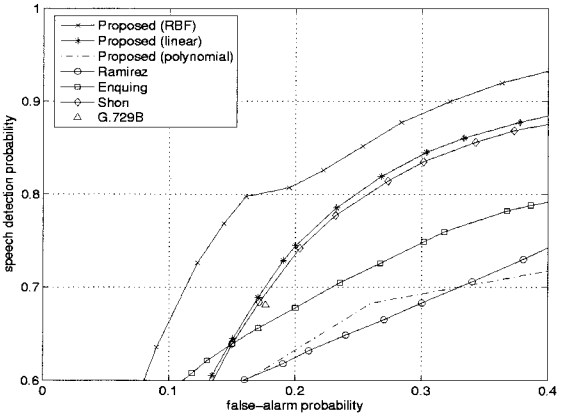

그림 5. Babble 잡음 5 dB SNR에서의 ROC 곡선
Fig. 5. ROC curves for babble noise (SNR = 5 dB).

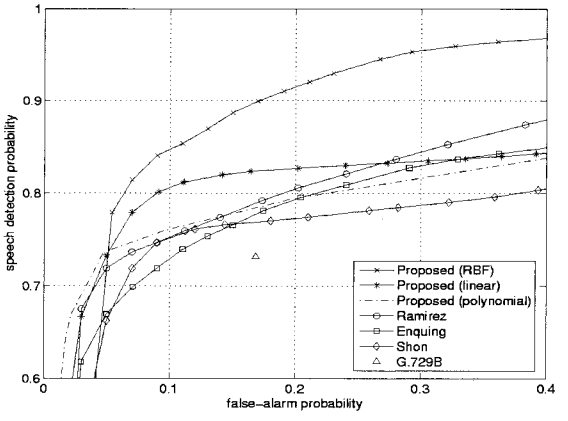

그림 6. Street 잡음 5 dB SNR에서의 ROC 곡선
Fig. 6. ROC curves for street noise (SNR = 5 dB).

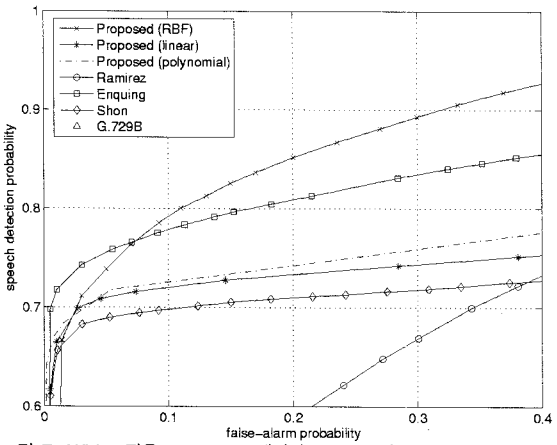

그림 7. White 잡음 5 dB SNR에서의 ROC 곡선
Fig. 7. ROC curves for white noise (SNR = 5 dB).

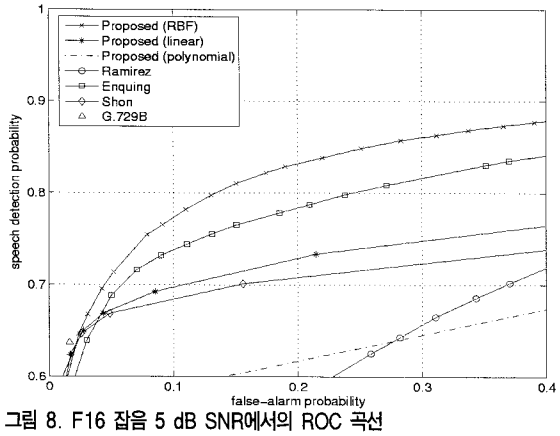

그림 8. F16 잡음 5 dB SNR에서의 ROC 곡선
Fig. 8. ROC curves for F16 noise (SNR = 5 dB).

비음성 구간과 음성의 onset과 offset 구간에서 일정한 값을 유지함으로써 비음성에 대한 안정적인 출력 결과를 보여주며, 비선형 SVM의 경우 비음성 구간에서 출력 결과의 변동이 있는 반면 음성 구간에서는 다른 VAD에 비해 견고한 출력값을 유지하여 잡음에 강인한 VAD 임을 보여준다.

그림 4-8은 각각 vehicular, babble, street, white 그리고 F16 잡음 환경 5 dB SNR에서 제안된 음성 검출기와 기존의 SVM 및 LRT를 기반으로 한 음성 검출기의 문턱값을 변경하면서 실제 음성을 음성이라고 판단한 음성 검출 확률 (Speech detection probability, Pd)과 비음성에 대해 음성이라고 판단한 오경보 확률 (False-alarm probability, Pf)을 측정한 ROC 곡선이다. 또한 G.729B 음성 검출기의 동작점을 그림 상에 표시하였다 [2], [5], [11-12].

모든 실험 조건에서 제안된 선형 kernel을 사용한 SVM 기반의 음성 검출기는 LRT 기반의 음성 검출기보다 우수한 성능을 보여주며, 이것은 단순한 기하 평균을 취하는 결정식 보다 SVM이 통계적 모델 기반의 음성 검출기를 위한 결정식에 더욱 적합한 방법임을 보여준다.

제시된 RBF 커널을 사용한 SVM 기반의 음성 검출기의 경우 훈련 데이터에 사용된 잡음 환경에 관계없이 기존의 SVM 기반의 음성 검출기 보다 다양한 잡은 환경에서 낮은 Pf 구간을 제외하고 가장 우수한 성능을 보이며, 특히 babble 잡음과 street 잡음에서 다른 음성 검출기에 비해 월등히 뛰어난 음성 검출 능력을 보여준다.

RBF 커널을 적용한 SVM의 경우 그림 3에 나타나듯이 음성 신호의 변화에도 일정한 성능을 유지하지만 음성의 onset과 offset과 같이 검출 오류가 발생하기 쉬운영역에서 비음성을 음성 구간으로 판단할 경우 문턱값이 일정값 이상이 될 때까지 오분류 상태로 판단함에 따라 낮은 Pf 구간에서는 다른 음성 검출기에 비해 저하된 성능을 보여준다.

각각의 잡음 환경 SNR 10 dB 및 20 dB에서도 5 dB SNR 처럼 제안된 음성 검출기가 기존의 음성 검출기에 비해 향상된 성능을 보였다. 이는 제안된 SVM을 이용한 통계적 모델 기반의 음성 검출기가 다른 SVM 기반의 음성 검출기 보다 다양한 잡음 환경에서도 강인한 음성 검출 방법임을 확인할 수 있다.

V. 결 론

본 논문에서는 음성의 존재와 부재에 대한 통계적 모델에 기반한 각 주파수 채널별 우도비를 단순히 기하 평균을 취하여 문턱값과 비교하는 대신, 이진 분류에 우수한 성능을 보이는 SVM을 도입함으로써 더욱 견고한 분류 방법을 바탕으로 기존의 방식보다 향상된 음성 검출기를 제시하였다. 선형 SVM 이외에 입력 특징 벡터의 비선형 특성을 고려하여 RBF 커널을 적용한 비선형 SVM을 실험에 적용하였으며, 다양한 실험 결과에서 얻어낸 ROC 곡선으로부터 제안된 음성 검출기의 성능이 우수함을 알 수 있다. 특히 RBF 커널을 적용한 비선형 SVM은 다양한 잡음에서 낮은 Pf 영역을 제외하면 가장 우수한 성능을 보여주었다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음. [2005-S096- 02, 신체장애인을 위한 착용형 단말 인터페이스 기술]

참 고 문 헌

1. K. Srinivasant and Allen Gersho, "Voice activity detection for cellular networks," Proc. IEEE Speech Coding Workshop, 85-86, Oct. 1993.
2. ITU, "A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70,", ITU-T Rec. G. 729, Annex B, 1996.
3. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoustics, Speech, Sig. Process., ASSP-32 (6), 1190-1121, Dec. 1984.
4 J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," Proc. Int. Conf. Acoustics, Speech, and Sig. Process., 1, 365-368, May 1998.
5. J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Sig. Process. Lett., 6 (1), 1-3, Jan. 1999.
6. Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," IEEE Sig. Process. Lett., 8 (10), 276-278, Oct. 2001.
7. J.-H. Chang, J. W. Shin, and N. S. Kim, "Voice activity detector employing generalized gaussian distribution," Electron. Lett., 40 (24), 1561-1563, Nov. 2004.
8. J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," IEEE Trans. Sig. Process., 54 (6), 1965-1976, June 2006.
9. Y. C. Lee and S. S. Ahn, "Statistical model-based VAD algorithm with Wavelet Transform," IEICE Trans. Fundamentals., E89-A (6), 1594-1600, June 2006.
10. J. Ramirez, J. M. Gorriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," IEEE Sig. Process. Lett., 13 (8), 497-500, Aug. 2006.
11. D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," Proc. Int. Conf. Sig. Process., 2, 1124-1127, Aug. 2002.
12. J. Ramirez, J. M. Gorriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, "Speech/Non-speech discrimination based on contextual information integrated bispectrum LRT," IEEE Sig. Process. Lett., 13 (8), 497-500, Aug. 2006.
13. V. N Vapnik, "An overview of statistical learning theory," IEEE Trans. Neural Networks, 10 (5), 988-999, Sep. 1999.
14. N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. (Cambridge Univ. Press, 2000)

저자 약력

• 조 규 행 (Q-Haing Jo)

2004년 2월: 인하대학교 전자공학과 학사
2004년 7월~2006년 7월: LG.Philips LCD 연구원
2006년 9월 ~현재: 인하대학교 전자공학부 석사

• 장 준 혁 (Joon-Hyuk Chang)

1998년 2월: 경북대학교 전자공학과 학사
2000년 2월: 서울대학교 전기공학부 석사
2004년 2월: 서울대학교 전기컴퓨터공학부 박사
2000년 3월~2005년 4월: ㈜넷더스 연구소장
2004년 5월~2005년 4월: 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월~2005년 8월: 한국과학기술연구원 (KIST) 연구원
2005년 9월~현재: 인하대학교 전자전기공학부 조교수

• 강 상 기 (Sangki Kang)

2002년 2월: 서울대학교 전기공학부 박사
2002년 3월~현재: 삼성전자 정보통신총괄 통신연구소 책임연구원