

# 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\*, 감 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 데이터에 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.

핵심어: 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. 서론

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)





식 (16)에서 사상함수  $\phi$ 가 존재할 수 있는 커널 함수  $K$ 가 주어진 경우 판별 함수는 아래와 같이 최종적으로 구한다.

$$f(A) = \sum_{i=1}^l \alpha_i^* z_i K(A_i^*, A) + b^* \begin{cases} H_1 & > \mu \\ H_0 & < \mu \end{cases} \quad (17)$$

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 데이터의 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 데이

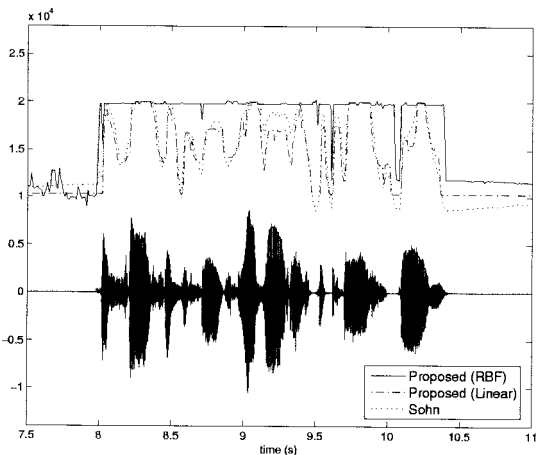


그림 3. Vehicular 잡음 5 dB SNR에서의 각 음성 검출기의 출력 결과 비교  
Fig. 3. Comparison of VAD output for vehicular noise (SNR = 5 dB).

터에서 사용하지 않은 F16 잡음이 5, 10 dB 그리고 20 dB SNR로 부과되었다.

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

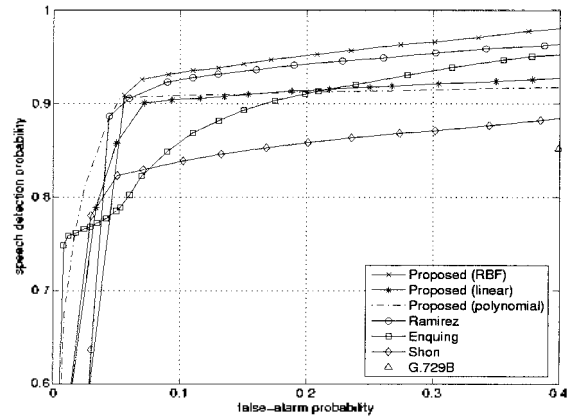


그림 4. Vehicular 잡음 5 dB SNR에서의 ROC 곡선  
Fig. 4. ROC curves for vehicular noise (SNR = 5 dB).

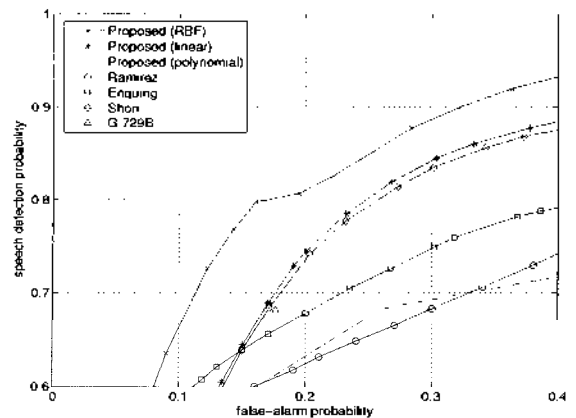


그림 5. Babble 잡음 5 dB SNR에서의 ROC 곡선  
Fig. 5. ROC curves for babble noise (SNR = 5 dB).

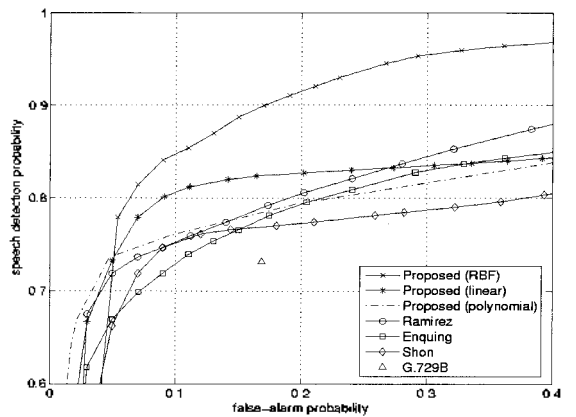


그림 6. Street 잡음 5 dB SNR에서의 ROC 곡선  
Fig. 6. ROC curves for street noise (SNR = 5 dB).

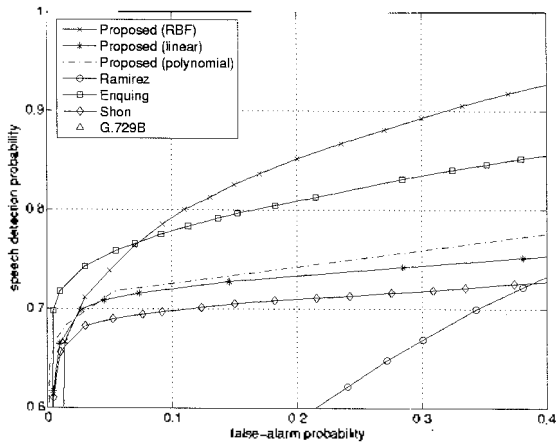


그림 7. White 잡음 5 dB SNR에서의 ROC 곡선  
Fig. 7. ROC curves for white noise (SNR = 5 dB).

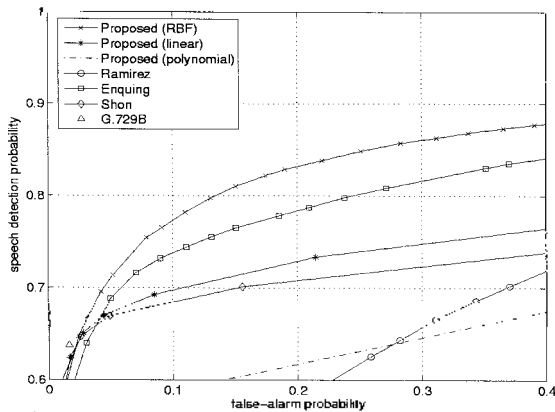


그림 8. F16 잡음 5 dB SNR에서의 ROC 곡선  
Fig. 8. ROC curves for F16 noise (SNR = 5 dB).

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 강한 VAD 임을 보여준다.

그림 4-8은 각각 vehicular, babble, street, white 그리고 F16 잡음 환경 5 dB SNR에서 제안된 음성 검출기와 기존의 SVM 및 LRF를 기반으로 한 음성 검출기의 문턱값을 변경하면서 실제 음성을 음성이라고 판단한 음성 검출 확률 (Speech detection probability, Pd)과 비음성에 대해 음성이라고 판단한 오정보 확률 (False-alarm probability, Pf)을 측정하여 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 LRF 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 잡음 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 쉬운 영역에서 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 강한 음성 검출 방법임을 확인할 수 있다.

## V. 결론

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]

## 참고 문헌

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. Gorrioz, 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. Gorrioz, 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

## • 조규행 (Q-Haing Jo)



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

## • 장준혁 (Joon-Hyuk Chang)



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

## • 강상기 (Sangki Kang)

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