

# SVM-enabled Voice Activity Detection

J. Ramírez<sup>1</sup>, P. Yélamos<sup>1</sup>, J.M. Górriz<sup>1</sup>, C.G. Puntonet<sup>2</sup> and J.C. Segura<sup>1</sup>

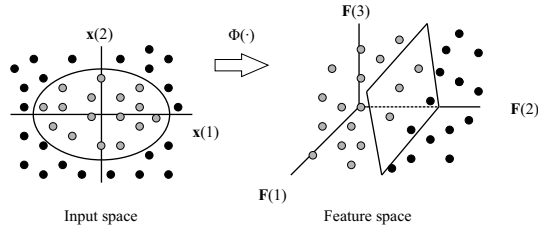
<sup>1</sup> Dept. of Signal Theory, Networking and Communications,  
University of Granada, Spain  
javierrp@ugr.es,

<sup>2</sup> Dept. of Architecture and Computer Technology,  
University of Granada, Spain

**Abstract.** Detecting the presence of speech in a noisy signal is an unsolved problem affecting numerous speech processing applications. This paper shows an effective method employing support vector machines (SVM) for voice activity detection (VAD) in noisy environments. The use of kernels in SVM enables to map the data into some other dot product space (called feature space) via a nonlinear transformation. The feature vector includes the subband signal-to-noise ratios of the input speech and a radial basis function (RBF) kernel is used as SVM model. It is shown the ability of the proposed method to learn how the signal is masked by the acoustic noise and to define an effective non-linear decision rule. The proposed approach shows clear improvements over standardized VADs for discontinuous speech transmission and distributed speech recognition, and other recently reported VADs.

## 1 Introduction

Currently, there are technology barriers inhibiting speech processing systems that work in extremely noisy conditions from meeting the demands of modern applications. These systems often require a noise reduction system working in combination with a precise voice activity detector (VAD). The classification task is not as trivial as it appears and its performance is strongly affected by the increasing background noise level. Since their introduction in the late seventies [1], Support Vector Machines (SVMs) marked the beginning of a new era in the learning from examples paradigm. SVMs have attracted recent attention from the pattern recognition community due to a number of theoretical and computational merits derived from the Statistical Learning Theory [2] developed by Vladimir Vapnik at AT&T. This paper shows an effective SVM-based VAD for improving the performance of speech processing systems that need to operate in noisy environment. The proposed method combines a noise robust speech processing feature extraction process together with a trained SVM model for classification. The results are compared to standardized techniques and a representative set of VAD methods.



**Fig. 1.** Effect of the map from input to feature space where the separation boundary becomes linear

## 2 Background on SVM learning

SVMs have recently been proposed for pattern recognition in a wide range of applications by its ability for learning from experimental data. The reason is that SVMs are much more effective than other conventional parametric classifiers. In SVM-based pattern recognition, the objective is to build a function  $f : R^N \rightarrow \{\pm 1\}$  using training data that is,  $N$ -dimensional patterns  $\mathbf{x}_i$  and class labels  $y_i$ :

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_\ell, y_\ell) \in R^N \times \{\pm 1\} \quad (1)$$

so that  $f$  will correctly classify new examples  $(\mathbf{x}, y)$ .

Hyperplane classifiers are based on the class of decision functions:

$$f(\mathbf{x}) = \text{sign}\{(\mathbf{w} \cdot \mathbf{x}) + b\} \quad (2)$$

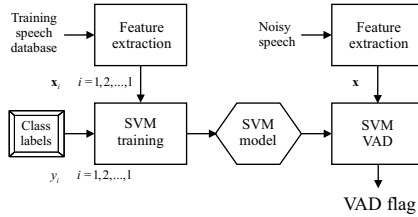
It can be shown that the optimal hyperplane is defined as the one with the maximal margin of separation between the two classes. The solution  $\mathbf{w}$  of a constrained quadratic optimization process can be expanded in terms of a subset of the training patterns called support vectors that lie on the margin:

$$\mathbf{w} = \sum_{i=1}^{\ell} \nu_i \mathbf{x}_i \quad (3)$$

Thus, the decision rule depends only on dot products between patterns:

$$f(\mathbf{x}) = \text{sign}\left\{\sum_{i=1}^{\ell} \nu_i (\mathbf{x}_i \cdot \mathbf{x}) + b\right\} \quad (4)$$

The use of kernels in SVM enables to map the data into some other dot product space (called feature space)  $F$  via a nonlinear transformation  $\Phi : R^N \rightarrow F$  and perform the above linear algorithm in  $F$ . Figure 1 illustrates this process where the 2-D input space is mapped to a 3-D feature space where the data is linearly separable. The kernel is related to the  $\Phi$  function by  $k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$ . In the input space, the hyperplane corresponds to a nonlinear decision function whose form is determined by the kernel. There are three common kernels that are used by SVM practitioners for the nonlinear feature mapping: *i*) polynomial, *ii*)



**Fig. 2.** Block diagram of the proposed SVM-based VAD

Radial basis function (RBF), and *iii*) sigmoid kernels. Thus, the decision function is nonlinear in the input space

$$f(\mathbf{x}) = \text{sign}\left\{\sum_{i=1}^{\ell} \nu_i k(\mathbf{x}_i, \mathbf{x}) + b\right\} \quad (5)$$

and the parameters  $\nu_i$  are the solution of a quadratic programming problem that are usually determined by the well known Sequential Minimal Optimization (SMO) algorithm [3]. Many classification problems are always separable in the feature space and are able to obtain better results by using RBF kernels instead of linear and polynomial kernel functions [4, 5].

### 3 Proposed SVM-based VAD

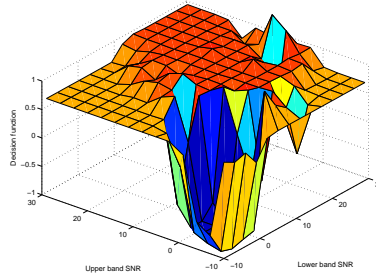
A block diagram of the proposed VAD is shown in Fig. 2. The first step is the training process on the training data set and its associated class labels. The signal is preprocessed and a feature vector is extracted for training. Once the SVM model has been trained, the proposed SVM-based algorithm consists of the following stages: *i*) the input signal is decomposed into speech frames and feature extraction is conducted for classification, and *ii*) the speech features  $\mathbf{x}$  are processed by the SVM decision function  $f$  defined in equation 5.

#### 3.1 Preprocessing and feature extraction

The algorithm for feature extraction is stated as follows. The input signal  $x(n)$  sampled at 8 kHz is decomposed into 25-ms overlapped frames with a 10-ms window shift. A denoising process based on a Wiener filter is applied to improve the performance of the VAD in high noise environments. Once the input signal has been denoised, a filterbank reduces the dimensionality of the feature vector to a representation including broadband spectral information suitable for detection.

#### 3.2 Training the SVM classification rule

The SVM model has been trained using LIBSVM software tool [6]. A training set consisting of 12 utterances of the AURORA 3 Spanish SpeechDat-Car (SDC)



**Fig. 3.** Decision function of a 2-band trained SVM model.

was used. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions with average SNR values between 25dB, and 5dB. The recordings used for training the SVM are selected to deal with different noisy conditions.

Fig. 3 shows the decision function of a 2-band trained SVM model. Note that,  $b$  can be used as a decision threshold for the VAD in the sense that the working point of the VAD can be shifted in order to meet the application requirements.

## 4 Experimental framework

This section analyzes the proposed VAD and compares its performance to other algorithms used as a reference. The analysis is based on the ROC curves, a frequently used methodology to describe the VAD error rate. The AURORA subset of the original Spanish SDC database [7] was used again in this analysis. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined as a function of the decision threshold being the actual speech frames and actual speech pauses determined by hand-labelling the database on the close-talking microphone.

Before showing comparative results, the selection of the optimal number of subbands is addressed. Fig. 4 shows the influence of the noise reduction block and the number of subbands on the ROC curves in high noisy conditions. First, noise reduction is not carried to better show the influence of the number of subbands. Increasing the number of subbands improves the performance of the proposed VAD by shifting the ROC curves in the ROC space. For more than four subbands, the VAD reports no additional improvements. This value yields the best trade-off between computational cost and performance. On the other hand, the noise reduction block included in the proposed VAD reports an additional shift of the ROC curve as shown in Fig. 4.

Fig. 5 shows the ROC curves of the proposed VAD and other frequently referred algorithms [8–11] for recordings from the distant microphone in high noisy conditions. The working points of the ITU-T G.729, ETSI AMR and AFE

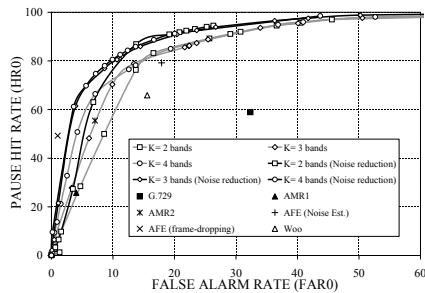


Fig. 4. Subband selection (High: high speed, good road, 5 dB average SNR).

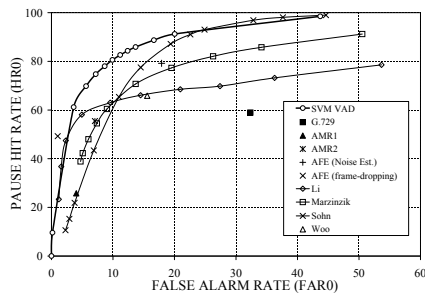


Fig. 5. Comparative results to other VAD methods

VADs are also included. The results show improvements in detection accuracy over standard VADs and over a representative set VAD algorithms [8–11]. Among all the VAD examined, our VAD yields the lowest false alarm rate for a fixed non-speech hit rate and also, the highest non-speech hit rate for a given false alarm rate. The benefits are especially important over ITU-T G.729, which is used along with a speech codec for discontinuous transmission, and over the Li’s algorithm, that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinik’s VAD [10] that tracks the power spectral envelopes, and the Sohn’s VAD [11], that formulates the decision rule by means of a model-based statistical likelihood ratio test.

## 5 Conclusions

An effective algorithm for detecting presence of speech in a noisy signal is proposed in this paper. The proposed strategy combines spectral noise reduction techniques and support vector machine learning tools to derive a non-linear decision rule in the input space defined in terms of the subbands SNRs. With these and other innovations the proposed method has shown to be more effective than VADs that define the decision rule in terms of average SNR values. The non-speech and speech classes can be clearly distinguished in the 3-D space and that the SVM model learns how the signal is masked by the noise. On the other hand,

increasing the number of subbands up to four improves the performance of the proposed VAD by shifting the ROC curve in the ROC space. Finally, the experiments conducted on the Spanish SpeechDat-Car database showed that the proposed algorithm outperforms ITU G.729, ETSI AMR1 and AMR2 and ETSI AFE standards as well as other recently reported VAD methods in speech/non-speech detection performance.

## 6 Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SESIBONN and SR3-VoIP projects (TEC2004-06096-C03-00, TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

## References

1. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer-Verlag, New York (1982)
2. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, Inc., New York (1998)
3. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Advances in Kernel Methods - Support Vector Learning. MIT Press (1999) 185–208
4. Clarkson, P., Moreno, P.: On the use of support vector machines for phonetic classification. In: Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing. Volume 2. (1999) 585–588
5. Ganapathiraju, A., Hamaker, J., Picone, J.: Applications of support vector machines to speech recognition. IEEE Transactions on Signal Processing **52** (2004) 2348–2355
6. Chang, C., Lin, C.J.: LIBSVM: a library for support vector machines. Technical report, Dept. of Computer Science and Information Engineering, National Taiwan University (2001)
7. Moreno, A., Borge, L., Christoph, D., Gael, R., Khalid, C., Stephan, E., Jeffrey, A.: SpeechDat-Car: A Large Speech Database for Automotive Environments. In: Proceedings of the II LREC Conference. (2000)
8. Woo, K., Yang, T., Park, K., Lee, C.: Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters **36** (2000) 180–181
9. Li, Q., Zheng, J., Tsai, A., Zhou, Q.: Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing **10** (2002) 146–157
10. Marzinzik, M., Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Transactions on Speech and Audio Processing **10** (2002) 341–351
11. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Processing Letters **16** (1999) 1–3

## Javier Ramirez

---

**De:** ISSN2006 [isnn@cqu.edu.cn]  
**Enviado:** jueves, 15 de diciembre de 2005 18:06  
**Para:** Dr. Ramirez  
**Asunto:** ISSN2006 Paper Code: 3-15-0087, Title: Svm-enabled Voice Activity Detection

Dear Dr. Javier Ramirez:

Congratulations! On behalf of the Program Committee of the Third International Symposium on Neural Networks (ISNN 2006) to be held in Chengdu during May 28-31, 2006, we are pleased to inform you that your paper showing in the E-mail subject has been accepted for presentation at ISSN2006. ISSN2006 received 2472 submitted papers from authors in 44 countries/regions. Based on reviews, only 630 papers are selected for publication in the symposium proceedings to be published by Springer as three volumes of Lecture Notes in Computer Science (LNCS). To include your paper in the proceedings, we need your full cooperation in the following aspects: 1. Finalize your paper based on the comments in the reviews listed below (if any). You may also view the reviewers' ratings, comments, and/or suggestions about the paper in you author' account of the ISSN2006 Online Submission System. Please improve your writing if possible. 2. Prepare your paper in the exact format as the sample for LNCS. Failure to do so will result in the exclusion of the paper in the proceedings. See [http://cilab.uestc.edu.cn/isnn2006/Instructions\\_LNCS.pdf](http://cilab.uestc.edu.cn/isnn2006/Instructions_LNCS.pdf) for the sample. The standard paper templates for both Word and LaTeX files can be found at <http://cilab.uestc.edu.cn/isnn2006/submission.htm> 3. Fill in the copyright form at <http://www.acae.cuhk.edu.hk/~isnn2006/Copyright.pdf>. The title of proceedings is Advances in Neural Networks - ISSN2006. The volume editors are Jun Wang, et al. 4. Fill in the registration form at <http://cilab.uestc.edu.cn/isnn2006/registration.htm> and make sure that your payment of the registration fee be received by January 15, 2006. Please note that each paper must be accompanied by at least one registration. Otherwise the paper cannot be included in the Proceedings. Please note that the nominal length of a paper is six pages and pages beyond six are subject to surcharge. The maximum number of pages is ten. 5. Upload all materials including scanned copyright form, pdf or rtf file of the paper, and source files (i.e., rtf file or LaTeX file and postscript files of figures) via online submission system at the ISSN2006 Online Submission System <http://isci.cqu.edu.cn/isnn/contribution/author/regupload.php>.

It is crucial for you to follow the above instructions and deadline to avoid leaving your paper out of the Proceedings. We will compile all the files and ship to Springer Germany by in Feb. Thank you very much in advance for your cooperation. If you need a hard copy of the acceptance letter or have any question in this regard, please contact the ISSN2006 secretariat at [isnn2006@uestc.edu.cn](mailto:isnn2006@uestc.edu.cn). Sincerely, Jun Wang, Bao-Liang Lu, and Hujun Yin ISSN2006 General Chair and Program Chairs

==Review Comments==

Reviewer 1:

General Interest: Good  
Originality of the Work: Average  
Technical Contents: Average  
Clarity in writing, tables, graphs & illustrations: Good  
Recommendation: Acceptable  
Comments:

This paper presents an interesting application of SVM in voice signal detection from nosiy environment. The experimental results are good. If the authors can make a more detailed discussion on how to select better features for SVM to capture the characters of voice signal with higher reliablity, the paper would be more complete.

Reviewer 2:

General Interest: Good  
Originality of the Work: Average  
Technical Contents: Good  
Clarity in writing, tables, graphs & illustrations: Good

Recommendation: Acceptable  
Comments:  
Nothing.