# Speech Event Detection Using Support Vector Machines

P. Yélamos[1], J. Ramírez[1], J.M. Górriz[1],
C.G. Puntonet[2], and J.C. Segura[1]

[1] Dept. of Signal Theory, Networking and Communications,
University of Granada, Spain
javierrp@ugr.es
[2] Dept. of Architecture and Computer Technology,
University of Granada, Spain

**Abstract.** An effective speech event detector is presented in this work for improving the performance of speech processing systems working in noisy environment. The proposed method is based on a trained support vector machine (SVM) that defines an optimized non-linear decision rule involving the subband SNRs of the input speech. It is analyzed the classification rule in the input space and the ability of the SVM model to learn how the signal is masked by the background noise. The algorithm also incorporates a noise reduction block working in tandem with the voice activity detector (VAD) that has shown to be very effective in high noise environments. The experimental analysis carried out on the Spanish SpeechDat-Car database shows clear improvements over standard VADs including ITU G.729, ETSI AMR and ETSI AFE for distributed speech recognition (DSR), and other recently reported VADs.

## 1 Introduction

With the advent of wireless communications, new speech services are being deployed with the development of modern robust speech processing technology. An important obstacle affecting these systems is the environmental noise and its harmful effect on the system performance. Most of the noise reduction algorithms often require a precise voice activity detector (VAD). The detection task is not as trivial as it appears since the increasing level of background noise degrades the classifier effectiveness.

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, 3] developed by Vladimir Vapnik at AT&T. Enqing [4] applied SVMs to the VAD problem showing promising results when the standardized ITU-T G.729 VAD [5] speech features were used as the inputs to the classification module. Later, this VAD was incorporated to a variable low bit-rate speech codec [6] using the local cosine transform. Recently, Qi *et al.*

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part I, LNCS 3991, pp. 356–363, 2006.

[7] has extended these ideas to the problem of classifying speech into voiced, unvoiced and silence frames. Again the SVM-based classifier operated on the G.729 speech features including the full-band energy difference, the low-band energy difference, the spectral distortion and the zero-crossing rate. This paper shows an effective SVM-based speech event detector for low-delay speech processing. The proposed method combines a noise robust speech processing feature extraction process together with a trained SVM model for classification. The results show improvements in speech/pause discrimination when compared to standardized VADs [5, 8, 9] and other recently published VAD methods [10, 11, 12, 13].

## 2 Support Vector Machines

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 $f : R^N \longrightarrow \{\pm 1\}$ using training data that is, $N$-dimensional patterns $\mathtt{x}_i$ and class labels $y_i$:

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

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

Hyperplane classifiers are based on the class of decision functions:

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

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:

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

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

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

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 $\Phi : R^N \longrightarrow F$ 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:

– Polynomial

$$k(\mathbf{x}, \mathbf{y}) = [\gamma(\mathbf{x} \cdot \mathbf{y}) + c]^d \tag{5}$$

– Radial basis function (RBF)

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2) \tag{6}$$

– Sigmoid

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\gamma(\mathbf{x} \cdot \mathbf{y}) + c) \tag{7}$$

Thus, the decision function is nonlinear in the input space

$$f(\mathbf{x}) = \text{sign}\{\sum_{i=1}^{\ell} \nu_i k(\mathbf{x}_i, \mathbf{x}) + b\} \tag{8}$$

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 [14]. 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 [15, 16].

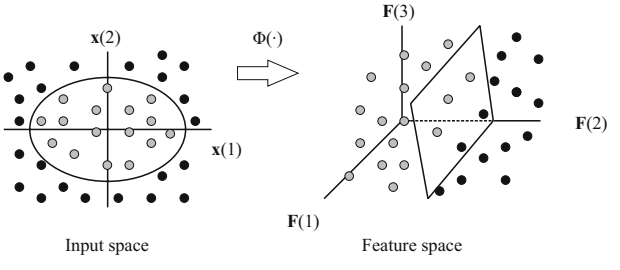



**Fig. 1.** Effect of the map from input to feature space where the separation boundary becomes linear

## 3 Speech Event Detection

The first step in defining the classification rule is the training process on the training data set and the associated class labels. The signal is preprocessed and a feature vector is extracted for training. Once the SMV model has been trained, the proposed speech event detector is described as follows: *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 8.

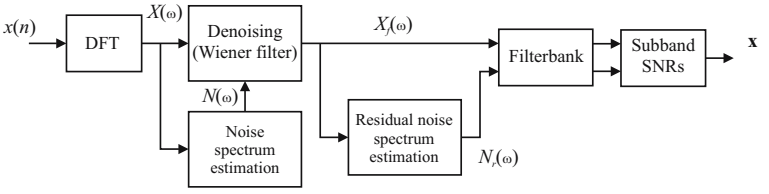



**Fig. 2.** Block diagram of the proposed SVM-based VAD

### 3.1 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. The current frame consisting of 200 samples is zero padded to 256 samples and power spectral magnitude $X(\omega)$ is computed through the discrete Fourier transform (DFT). A denoising process based on a two-stage Wiener filter is applied to improve the performance of the VAD in high noise environments. It is described as follows:

1. Spectral subtraction.

$$S_1(\omega) = L_s X_f(\omega) + (1 - L_s) \max(X(\omega) - \alpha N(\omega), \beta X(\omega)) \tag{9}$$

2. First WF design and filtering.

$$\begin{gathered} \mu_1(\omega) = S_1(\omega)/N(\omega) \\ W_1(\omega) = \mu_1(\omega)/(1 + \mu_1(\omega)) \\ S_2(\omega) = W_1(\omega)X(\omega) \end{gathered} \tag{10}$$

3. Second WF design and filtering.

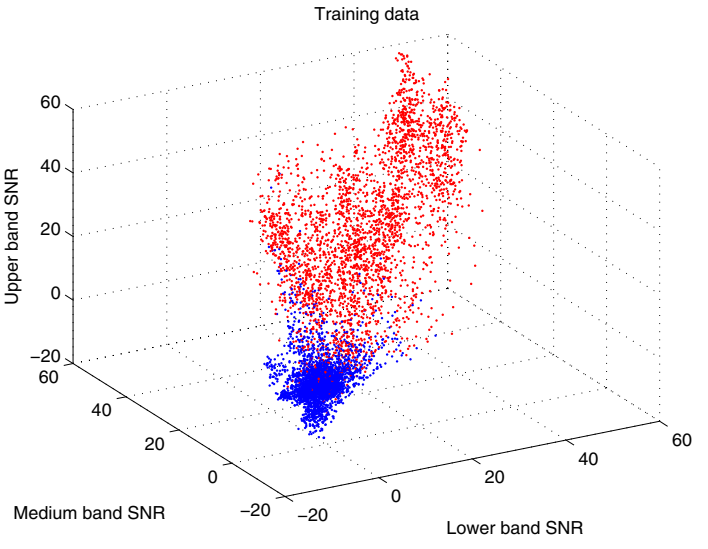
$$\begin{gathered} \mu_2(\omega) = S_2(\omega)/N(\omega) \\ W_2(\omega) = \max(\mu_2(\omega)/(1 + \mu_2(\omega)), \beta) \\ X_f(\omega) = W_2(\omega)X(\omega) \end{gathered} \tag{11}$$

where $L_s = 0.99$, $\alpha = 1$ and $\beta = 10^{(-22/10)}$ is selected to ensure a -22dB maximum attenuation for the filter in order to reduce the high variance musical noise that normally appears due to rapid changes across adjacent frequency bins. 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. Thus, the signal and the residual noise is passed through a $K$-band filterbank which is defined by

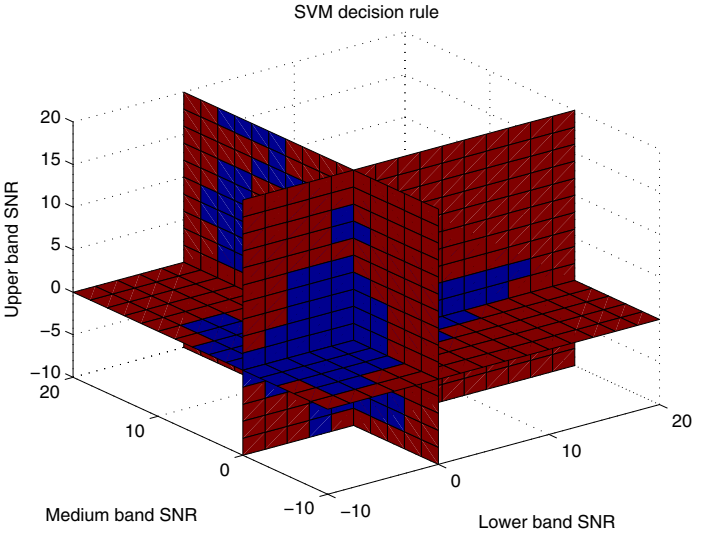
$$\begin{gathered} E_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} X_f(\omega); \quad N_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} N(\omega) \\ \omega_k = \tfrac{\pi}{K}k \qquad k = 0, 1, ..., K-1 \end{gathered} \tag{12}$$

and the subband SNRs are computed as

$$\mathrm{SNR}(k) = 20 \log_{10} \left( \frac{E_B(k)}{N_B(k)} \right) \qquad k = 1, 2, ..., K-1 \tag{13}$$




(a)




(b)

**Fig. 3.** Classification rule in the input space after training a 3-band SVM model. a) Training data set, b) SVM classification rule.

### 3.2 Training

The SVM model has been trained using LIBSVM software tool [17]. A training set consisting of 12 utterances of the AURORA 3 Spanish SpeechDat-Car (SDC) 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. 3.1.a shows the training data set in the 3-band input space. After the training process, the SVM decision rule defined by equation 8 is graphically shown in Fig 3.1.b where the non-speech and speech classes are clearly distinguished in the 3-D space. Note that, the SVM model learns how the signal is masked by the noise and automatically defines the decision rules in the input space.

The SVM formulation is based on $C$-Support Vector Classification [18, 3] while the decision rule is defined by equation 8. An RBF kernel is used and the training process consists in finding the solution of a primal problem

$$\begin{array}{c} \min_{\alpha} \quad \frac{1}{2}\alpha^T \mathbf{Q}\alpha - \mathbf{e}^T\alpha \\ 0 \leq \alpha_i \leq C, \quad i = 1, 2, ..., \ell \\ \text{subject to} \quad \mathbf{y}\alpha = 0 \end{array} \tag{14}$$

by using LIBSVM [17], where $\mathbf{e}$= [1 1 ... 1], $C > 0$ is the upper bound and $Q_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$. After this process, the support vectors $\mathbf{x}_i$ and coefficients $\alpha_i$ required to evaluate the decision rule defined in equation 8 are selected where $\nu_i = y_i \alpha_i$. 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 [19] was used again in this analysis. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined for each noisy condition being the actual speech frames and actual speech pauses determined by hand-labelling the database on the close-talking microphone. Thus, the objetive is to work as close as possible to the ideal [0%,100%] point where both speech and non-speech are determined with no error. Preliminary experiments determined that increasing the number of subbands up to four subbands improved the performance of the proposed VAD by shifting the ROC curves in the ROC space.

Fig. 4 compares the ROC curve of the proposed VAD to frequently referred algorithms [11, 12, 13, 10] for recordings from the distant microphone high noisy

conditions. The working points of the ITU-T G.729, ETSI AMR and AFE VADs are also included. The results show improvements in detection accuracy. Thus, 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 [5], which is used along with a speech codec for discontinuous transmission, and over the Li's algorithm [12], that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik's VAD [13] that tracks the power spectral envelopes, and the Sohn's VAD [10], that formulates the decision rule by means of a model-based statistical likelihood ratio test.

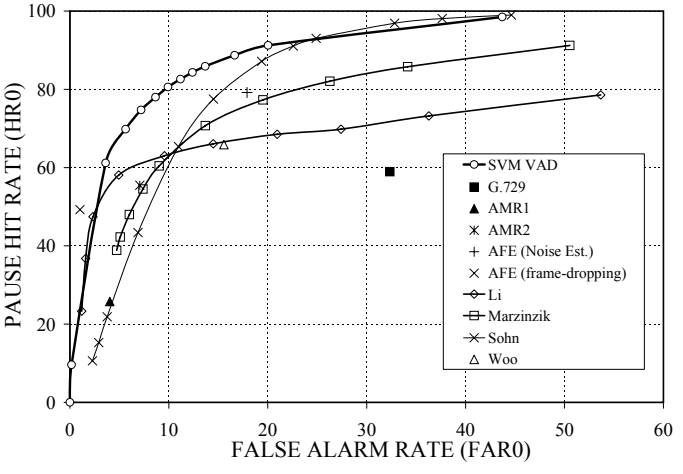



**Fig. 4.** Comparative results to other VAD methods

## 5 Conclusions

This paper has shown an effective speech event detector combining spectral noise reduction and support vector machine learning tools. The use of kernels enables defining a non-linear decision rule in the input space which is defined in terms of subbands SNRs. It is also shown the ability of SVM tools to learn how the speech is masked by the acoustic noise. With these and other innovations the proposed method has shown to be more effective than VADs that define the decision rule in terms of an average SNR values. The proposed algorithm also outperformed ITU G.729, ETSI AMR1 and AMR2 and ETSI AFE standards and recently reported VAD methods in speech/non-speech detection performance.

## Acknowledgements

This work has been funded by the European Commission (HIWIRE, IST No. 507943) and the Spanish MEC project TEC2004-03829/FEDER.

## References


1. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer-Verlag, New York (1982)
2. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, Berlin (1995)
3. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, Inc., New York (1998)
4. Enqing, D., Guizhong, L., Yatong, Z., Xiaodi, Z.: Applying support vector machines to voice activity detection. In: 6th International Conference on Signal Processing. Volume 2. (2002) 1124–1127
5. ITU: A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70. ITU-T Recommendation G.729-Annex B (1996)
6. Enqing, D., Heming, Z., Yongli, L.: Low bit and variable rate speech coding using local cosine transform. In: Proc. of the 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCON '02). Volume 1. (2002) 423–426
7. Qi, F., Bao, C., Liu, Y.: A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech. In: International Symposium on Chinese Spoken Language Processing. (2004) 77–80
8. ETSI: Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels. ETSI EN 301 708 Recommendation (1999)
9. ETSI: Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. ETSI ES 201 108 Recommendation (2002)
10. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Processing Letters **16** (1999) 1–3
11. Woo, K., Yang, T., Park, K., Lee, C.: Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters **36** (2000) 180–181
12. 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
13. 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
14. 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
15. 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
16. Ganapathiraju, A., Hamaker, J., Picone, J.: Applications of support vector machines to speech recognition. IEEE Transactions on Signal Processing **52** (2004) 2348–2355
17. 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)
18. Cortes, C., Vapnik, V.: Support-vector network. Machine Learning (1995)
19. 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)