Speech/Non-Speech discrimination combining advanced feature extraction and SVM learning

Javier Ramírez, Pablo Yélamos, Juan Manuel Górriz, José C. Segura and Luz García

Departamento de Teoría de la Señal Telemática y Comunicaciones
University of Granada, Spain
javierrp@ugr.es

10.21437/Interspeech.2006-463

Abstract

This paper shows an effective speech/non-speech discrimination method for improving the performance of speech processing systems working in noisy environment. The proposed method uses a trained support vector machine (SVM) that defines an optimized non-linear decision rule over different sets of speech features. Two alternative feature extraction processes based on: *i*) subband SNR estimation after denoising, and *ii*) long-term SNR estimation were compared. Both methods show the ability of the SVM-based classifier to learn how the signal is masked by the acoustic noise and to define an effective non-linear decision rule. However, it is shown that a feature vector incorporating contextual information yielded better speech/non-speech discrimination even when no denoising is applied. 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.

Index Terms: voice activity detection, support vector machine learning, speech enhancement.

1. Introduction

With the advent and development of wireless communications, the emerging applications in the field of speech communication are demanding increased levels of performance in many areas. An important obstacle affecting most of these applications is the environmental noise and its harmful effect on the system performance. Most of the noise reduction algorithms often require to estimate the noise statistics by means of 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. During the last decade numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD decision on speech processing systems. Most of them have focussed on the development of robust algorithms, with special attention on the study and derivation of noise robust features and decision rules [1, 2, 3, 4].

Since their introduction in the late seventies [5], 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 [5] developed by Vladimir Vapnik at AT&T. As an example, SVMs have been used for isolated handwritten digit recognition, object recognition, speaker identification or text categorization. Enqing [6] applied SVMs to VAD showing promising results when the standardized ITU-T G.729 VAD [7] speech features were used as the inputs to the classification module. Later, this VAD was incorporated to a variable low bit-rate speech codec [8] using the local cosine transform. Recently, Qi *et al.* [9] has extended these ideas to the problem of classifying speech into voiced, unvoiced and silence frames. This paper shows an improved SVM-based VAD for robust speech recognition. The proposed method combines noise robust feature extraction processes together with a trained SVM model for classification. The results show improvements in speech/pause discrimination when compared to standardized VADs [7, 10, 11] and other recently published methods [1, 2, 3, 4].

2. Support vector machines

Detecting the presence of speech in a noisy signal is a two-class classification problem requiring a rule, which, based on external observations, assigns an object to one of the classes. A possible formalization of this task is by means of SVMs that enable building a function $f : R^N \longrightarrow \{\pm 1\}$ using training data that is, N-dimensional patterns 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 non observed test data $(\mathbf{x}, y)$.

Hyperplane classifiers are based on the class of decision functions:

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

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 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:

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

In addition, by means of kernels, SVM enables to redefine the classification problem into some other potentially much higher dimensional feature space F via a nonlinear transformation $\Phi : R^N \longrightarrow F$ and perform the above algorithm in F. The kernel is

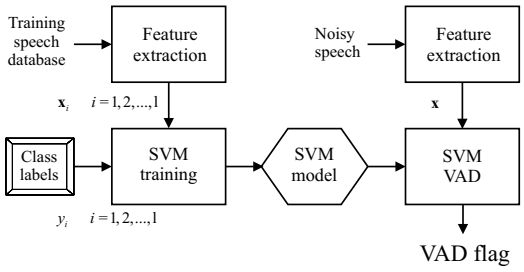

Figure 1: Block diagram of the proposed SVM-based VAD

related to the Φ function by $k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$ and the decision function becomes nonlinear in the input space:

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

Finally, the weights ν_i are determined as the solution of a quadratic programming optimization problem by means of the well known Sequential Minimal Optimization (SMO) algorithm [12].

3. Voice activity detection

A block diagram of the proposed VAD is shown in Fig. 1. The fist 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*) feature extraction, and *ii*) SVM-based classification using the decision function f defined in equation 5.

3.1. Feature extraction

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 the power spectral magnitude $X_l(\omega)$ of the *l*-th frame is computed through the discrete Fourier transform (DFT). Two different feature extraction schemes are compared in this paper:

3.1.1. Subband SNR extraction through denoising

A noise reduction process similar to the used in [13] is applied and the power spectrum of the filtered signal $X_l^f(\omega)$ and the residual noise $N_l^r(\omega)$ is obtained. 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

$$E_l^B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} X_l^f(\omega); \quad N_l^B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} N_l^r(\omega) \quad \omega_k = \tfrac{\pi}{K}k \qquad k = 0, 1, ..., K-1 \quad (6)$$

and the subband SNRs are computed as

$$\text{SNR}_l(k) = 20\log_{10}\left(\frac{E_l^B(k)}{N_l^B(k)}\right) \qquad k = 0, 1, ..., K-1 \quad (7)$$

3.1.2. Contextual subband feature extraction

Alternatively, a contextual feature extraction process was evaluated and compared to the procedure described above. It consists of a measure of the contextual deviation of the power spectrum from the background noise and is defined in terms of the long-term spectral envelope [14]:

$$\hat{X}_l(\omega) = \max\{X_k(\omega)\} \quad k \in \{l-L, ..., l, ..., l+L\} \quad (8)$$

It is then transformed to a wide *K*-band spectral representation:

$$E_l^B(k) = 10\log_{10}\left(\frac{2K}{N}\sum_{\omega=\omega_j}^{\omega_{j+1}-1} \hat{X}_l(\omega)\right) \quad (9)$$

where $\omega_j = 2\pi j/NFFT$, $j = \lfloor NFFTk/(2K) \rfloor$ and k= 0, 1, ..., $K-1$. Finally, the feature vector $\mathbf{x}$ for classification consists of the K subband SNRs defined to be:

$$SNR_l(k) = E_l^B(k) - N_l^B(k) \quad (10)$$

where the spectral representation of the noise, $N_l^B(k)$, is estimated during a short initialization period at the beginning of the process and constantly updated during non-speech periods.

3.2. Training

The SVM model is trained using LIBSVM software tool [15]. The AURORA-3 Spanish SpeechDat-Car database 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 training set consists of 12 utterances recorded at variable SNR conditions.

The SVM formulation is based on *C*-Support Vector Classification [5] and the decision rule is defined by equation 5. The subband SNRs given by equations 7 or 10 are used as discriminative speech features while an RBF kernel is used in the training process that consists of finding the solution of a primal problem

$$\min_\alpha \tfrac{1}{2}\alpha^T\mathbf{Q}\alpha - \mathbf{e}^T\alpha \quad ; \quad 0 \le \alpha_i \le C, \quad i = 1, 2, ..., \ell \quad \text{subject to} \quad \mathbf{y}\alpha = 0 \quad (11)$$

by using LIBSVM [15], where **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 α_i required to evaluate the decision rule 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. This is crucial for the application being considered since a miss of speech frames strongly affects to the performance of most speech processing systems. Next section illustrates the behavior of the SVM-based decision rule as the threshold is modified from the trained value.

4. Experimental analysis

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

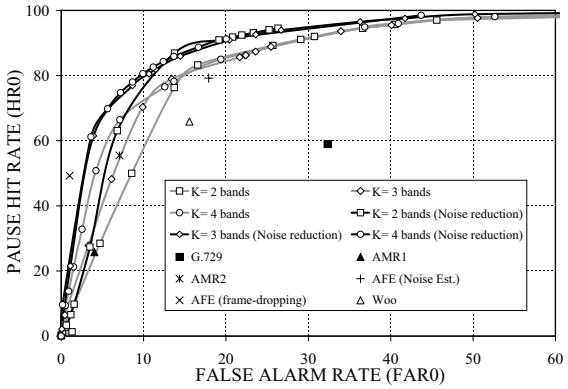

Figure 2: Selection of the number of subbands (High: high speed, good road, 5 dB average SNR).

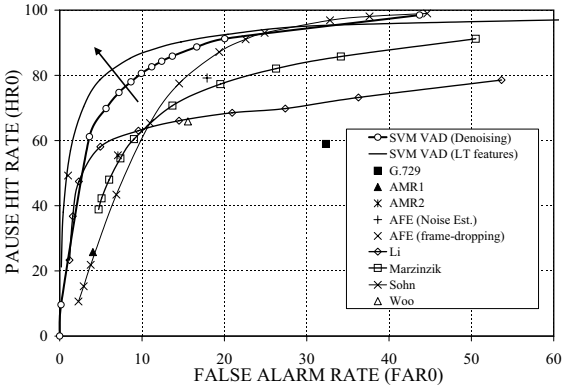

Figure 3: Comparative results to other VAD methods

describe the VAD error rate. The AURORA subset of the original Spanish SDC database [16] 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.

4.1. Selection of the optimum number of subbands

Before showing comparative results, the selection of the optimal number of subbands for the first feature extraction process is addressed. Fig. 2 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 applied to better show the influence of the number of subbands. In this way. 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 reports an additional shift of the ROC curve as shown in Fig. 2.

4.2. Comparative results

Fig. 3 compares the two feature extraction processes for SVM-based VAD and other frequently referred algorithms [1, 4, 2, 3] for recordings from the distant microphone in high noisy conditions. The working points of the ITU-T G.729, ETSI AMR and AFE VADs are also included. It was found that increasing L from 1 to 8 frames also leads to a shift-up and to the left of the ROC curve. The optimal parameters for the proposed VAD are then K= 4 subbands and L= 8 frames. These improvements are mainly achieved by: *i*) including contextual information in the feature vector defined by equation 10, and *ii*) defining a SVM-based classifier that is able to learn how the speech signal is masked by the acoustic noise present in the environment. The results also show improvements in detection accuracy over standardized VADs and over other recently published VADs [1, 4, 2, 3]. 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 Marzinzik's VAD [2] that tracks the power spectral envelopes, and the Sohn's VAD [3], that formulates the decision rule by means of a model-based statistical likelihood ratio test.

5. Analysis and improvements

This section analyzes the decision rule in the input space and suggests a fast algorithm for SVM classification. Fig. 4.a shows the training data set in the 3-band input space. It is shown that the two classes can not be separated without error in the input space. Fig. 4.b shows a slice plot of the SVM decision rule that is obtained after the training process. Note that, *i*) the non-speech and speech classes are clearly distinguished in the 3-D space, and *ii*) the SVM model learns how the signal is masked by the noise and automatically defines the decision rule in the input space.

Fig. 4.b also suggests a fast algorithm for performing the decision rule defined by equation 5 that becomes computationally expensive when the number of support vectors and/or the dimension of the feature vector are high. Note that all the information needed for deciding the class a given feature vector **x** belongs resides in figure 4.b. Thus, the input space can be discretized over the different components of the feature vector **x** as

$$\begin{array}{lllll} \mathbf{x}(1) & m_{\mathbf{x}(1)}, & m_{\mathbf{x}(1)} + \Delta_{\mathbf{x}(1)}, & ..., & M_{\mathbf{x}(1)} \\ \mathbf{x}(2) & m_{\mathbf{x}(2)}, & m_{\mathbf{x}(2)} + \Delta_{\mathbf{x}(2)}, & ..., & M_{\mathbf{x}(2)} \\ & ... & & & \\ \mathbf{x}(N) & m_{\mathbf{x}(N)}, & m_{\mathbf{x}(N)} + \Delta_{\mathbf{x}(N)}, & ..., & M_{\mathbf{x}(N)} \end{array} \quad (12)$$

and the decision rule $f(\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N))$ can be precomputed for the previously defined data grid and stored in an N-dimensional look-up table. Given a feature vector **x**= [**x**(1), **x**(2), ..., **x**(N)], the first step is to find the nearest point in the grid defined previously and then perform a table look-up to assign a class (speech or non-speech) to the feature vector **x**.

6. Conclusions

This paper showed an effective voice activity detector combining noise robust feature extraction processes 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 instantaneous or contextual subbands SNRs. Increasing the number of subbands up to four improved the performance of the pro-

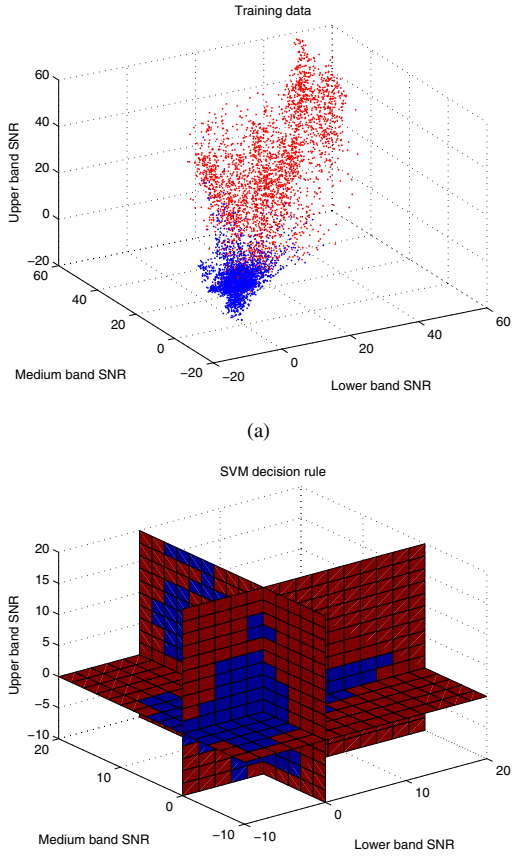

Figure 4: Classification rule in the input space after training a 3-band SVM model. a) Training data set, b) SVM classification rule.

posed VAD by shifting the ROC curve in the ROC space. Moreover, an advanced feature extraction process including contextual information also reported significant improvements in speech/non-speech discrimination yielding the best tradeoff between computational cost and performance. With these and other innovations the proposed methods have shown to be more effective than VADs that define the decision rule in terms of an average SNR values. The proposed algorithms also outperformed ITU G.729, ETSI AMR1 and AMR2 and ETSI AFE standards and recently reported VAD methods in speech/non-speech detection performance.

7. Acknowledgements

This work was supported by the European Commission (HIWIRE, IST No. 507943) and by the Spanish Government under the SR3-VoIP MEC project (TEC2004-03829/FEDER).

8. References

[1] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[2] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3, 1999.

[4] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[5] V.N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, Inc., New York, 1998.

[6] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *6th International Conference on Signal Processing*, 2002, vol. 2, pp. 1124–1127.

[7] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[8] D. Enqing, Z. Heming, and L. Yongli, "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)*, 2002, vol. 1, pp. 423–426.

[9] F. Qi, C. Bao, and Y. Liu, "A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech," in *International Symposium on Chinese Spoken Language Processing*, 2004, pp. 77–80.

[10] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[11] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 Recommendation*, 2002.

[12] J.C. Platt, *Advances in Kernel Methods - Support Vector Learning*, chapter Fast Training of Support Vector Machines using Sequential Minimal Optimization, pp. 185–208, MIT Press, 1999.

[13] J. Ramírez, P. Yélamos, J. M. Górriz, C. G. Puntonet, and J. C. Segura, "Svm-enabled voice activity detection," *Lecture Notes in Computer Science*, vol. 3972, pp. 676–681, 2006.

[14] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[15] C.C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," Tech. Rep., Dept. of Computer Science and Information Engineering, National Taiwan University, 2001.

[16] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proceedings of the II LREC Conference*, 2000.