SVM-based speech endpoint detection using contextual speech features

J. Ramírez, P. Yélamos, J.M. Górriz and J.C. Segura

Shown is an effective speech endpoint detection algorithm using a trained support vector machine (SVM) and a feature vector including contextual information speech features. With this and other innovations the proposed algorithm yields high discrimination and reports significant improvements over standard methods and algorithms defining the decision rule in terms of averaged subband speech features.

Introduction: The deployment of new wireless speech communication services finds a serious implementation barrier in the harmful effect of acoustic noise present in the operating environment. This challenge has motivated continuous research and development in robust speech processing and real-time performance. On the other hand, since their introduction in the late 1970s, support vector machines (SVMs) [1] marked the beginning of a new era in the learning from examples paradigm. Enqing *et al.* [2] applied SVMs to voice activity detection (VAD) showing promising results on the ITU-T G.729 speech codec features. This Letter extends these ideas and shows an improved speech endpoint detection algorithm enabling an SVM to define a nonlinear decision rule involving contextual speech features [3]. The proposed method's results are more effective than ITU-T and ETSI standards and methods that define the decision rule in terms of averaged subband speech features [4–7].

SVM-based speech endpoint detection: 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 formalisation of this task is by means of SVMs that enable building a function f: $R^N \rightarrow \{\pm 1\}$ using training data, that is N-dimensional patterns $\mathbf{x}_i$ and class labels y_i:

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

such that f will classify unseen examples $(\mathbf{x}, y)$ according to the structural risk minimisation (SRM) principle [1]. An example $\mathbf{x}$ is assigned to the class $+1$ if $f(\mathbf{x}) \geq 1$ and to the class -1 otherwise. Statistical learning theory [1] shows that it is crucial to restrict the class of functions that the learning machine can implement. Hyperplane classifiers are defined by the class of decision functions $f(\mathbf{x}) = \text{sign}\{(\mathbf{w} \cdot \mathbf{x}) + b\}$, where $\mathbf{w}$ and b are selected to define the maximal margin hyperplane. Moreover, it can be shown that $\mathbf{w}$ can be expanded in terms of a subset of the training patterns $\mathbf{x}_i$ called support vectors that lie on the margin. In addition, SVM enables to redefine the classification problem into some other potentially much higher dimensional feature space F via a nonlinear transformation $\Phi(\mathbf{x})$: $R^N \rightarrow F$ and perform the above algorithm in F:

$$f(\mathbf{x}) = \text{sign}\left\{\sum_{i=1}^{\ell} v_i(\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) + b\right\} \tag{2}$$

where the dot product is efficiently computed according to Mercer's theorem by means of kernels defined to be $k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$, and the weights v_i are the solution of a dual optimisation problem [1].

In this work, the well-known radial basis function (RBF) kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \cdot \|\mathbf{x} - \mathbf{y}\|^2)$ is used instead of linear or polynomial kernels since it yields better speech/pause classification results. Once the SVM model is trained, the speech features $\mathbf{x}$ are classified according to the SVM decision function $f(\mathbf{x})$ defined by (2). Note that b can be used as a detection threshold for the VAD in order to tune its working point and meet the application requirements. This is crucial for the application being considered since a miss of speech strongly affects the performance of most speech communication systems.

Feature extraction: The noisy speech signal is pre-processed and a feature vector $\mathbf{x}$ is extracted for training and testing on a frame by frame basis. A measure of the long-term spectral divergence between speech and noise is used as a discriminative speech feature. 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 l consisting of 200 samples is zero padded to $N = 256$ samples and the power spectral magnitude $X_l(\omega_m)$ is computed through the N-point discrete Fourier transform (DFT), where $\omega_m = 2\pi m/N$ and $m = 0, 1, \ldots, N/2$. Then, the long-term spectral envelope as defined in [7], that includes contextual information of the speech signal, is computed as: $\hat{X}_l(m) \equiv \hat{X}_l(\omega_m) = \max\{X_j(\omega_m),\ j = l-L, \ldots, l-1, l, l+1, \ldots, l+L\}$, and its dimensionality is reduced to a wide K-band spectral representation:

$$E_B(k, l) = 10 \log_{10}\left(\frac{2K}{N} \sum_{m=m_k}^{m=m_{k+1}-1} \hat{X}_l(m)\right) \tag{3}$$

where $m_k = \lfloor NFFT \cdot k/(2K) \rfloor$ and $k = 0, 1, \ldots, K-1$. Finally, the feature vector $\mathbf{x}$ for classification consists of the K subband SNRs defined to be $\text{SNR}(k, l) = E_B(k, l) - N_B(k, l)$, where the spectral representation of the noise, $N_B(k, l)$, is estimated during a short initialisation period at the beginning of the process and constantly updated during non-speech periods according to:

$$N_B(k, l) = \begin{cases} N_B(k, l-1) & \text{VAD flag for the } l\text{th frame} = 0 \\ \alpha N_B(k, l-1) + (1-\alpha) E_B(k, l) & \text{otherwise} \end{cases}$$

Fig. 1 clarifies the motivations for using contextual speech features. The 2-D feature space defined for $K = 2$ is represented for 12 speech utterances of the AURORA 3 Spanish SpeechDat-Car database. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The files are categorised into three noisy conditions defined by different driving conditions with average SNR ranging from 25 to 5 dB. It is clearly shown that increasing the size of the analysis window from $L = 0$ to 8 frames leads to a better separability of the data in the feature space and enables a better defined SVM-based classifier.

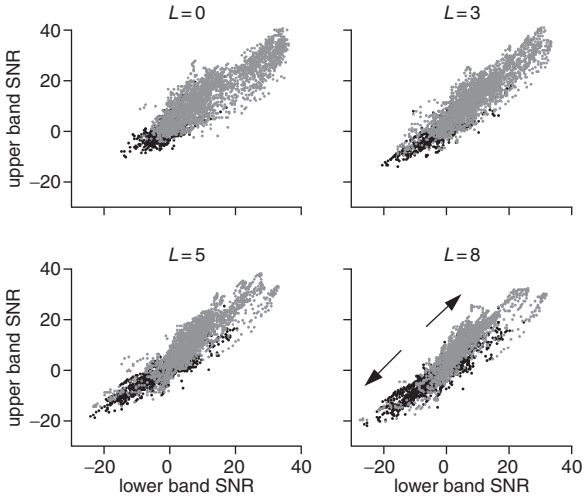

Fig. 1 *Separation of speech features in input space when increasing window size L*

Analysis and results: This Section analyses the proposed VAD and compares its performance to standard and recently published algorithms. The analysis is based on the receiver operating characteristics (ROC) curves [5], a frequently used methodology to completely describe the VAD error rate as the decision threshold b varies. The non-speech hit rate (HR0) and speech hit rate (HR1) are used to assess the performance of the VAD. They are defined as the ratio of the detected non-speech or speech frames to the total number of non-speech or speech frames, respectively. Complementary to these values are the false alarm rates defined by $\text{FAR1} = 1 - \text{HR0}$ and $\text{FAR0} = 1 - \text{HR1}$. For the analysis, the actual speech frames and actual speech pauses were determined by hand-labelling the database on the close-talking microphone.

Before showing comparative results, the selection of the optimal number of subbands (K) is considered. Fig. 2 shows the influence of the number of subbands on the ROC curves in high noisy conditions (high speed over good road, 5 dB) for $L = 8$. The working points of the ITU-T G.729, ETSI AMR and ETSI AFE VADs are also included as well as ROC curves of other VAD methods. 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. Thus, $K=4$ subbands yields the best trade-off between computational cost and performance.

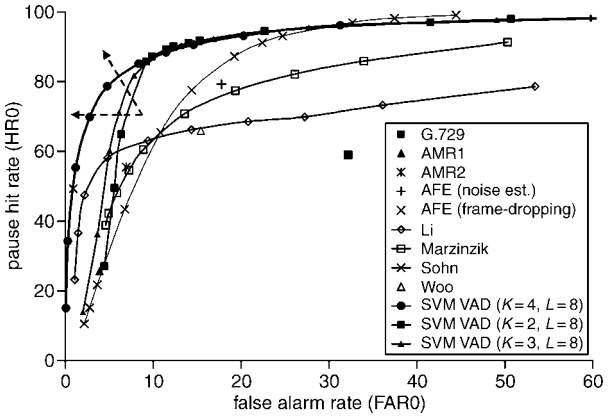

Fig. 2 *ROC curves of proposed VAD for different number of subbands K (high speed, good road, 5 dB average SNR)*

Fig. 3 shows the ROC curves of the proposed VAD against L and $K=4$. It is shown that increasing L from 1 to 8 frames also leads to a shift-up and to the left of the ROC curve. These results are consistent with Fig. 1 that predicted a better data separability for $L=8$. The optimal parameters for the proposed VAD are then $K=4$ subbands and $L=8$ frames. It can be also concluded that the proposed method yields significant improvements in speech/non-speech discrimination over ITU-T G.729 and ETSI AMR and ETSI AFE standards as well as over a representative set of VAD algorithms [4–7]. These improvements are mainly achieved by: (i) including contextual information in the feature vector, and (ii) defining an SVM-based classifier that is able to learn how the speech signal is masked by the acoustic noise present in the environment.

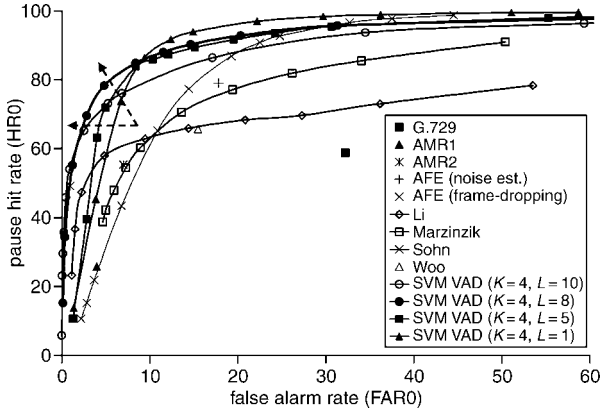

Fig. 3 *Influence of size of analysis window (L) on ROC curves; comparison to ITU-T and ETSI standards and recently published VAD methods (high speed, good road, 5 dB average SNR)*

Conclusion: We have shown the effectiveness of SVM learning concepts for robust speech endpoint detection. The proposed method defines a nonlinear decision rule in terms of a feature vector including contextual information speech features. The analysis conducted on well-known speech databases unveils significant improvements over ITU-T G.729, ETSI AMR and ETSI AFE standards as well as over VADs that define the decision rule in terms of averaged subband speech features.

Acknowledgments: This work has been funded by the European Commission (HIWIRE, IST Project No. 507943) and the SR3-VoIP Spanish MEC project (TEC2004-03829/FEDER).

© IEE 2006
21 November 2005
Electronics Letters online no: 20064068
doi: 10.1049/el:20064068

J. Ramírez, P. Yélamos, J.M. Górriz and J.C. Segura (*Department of Signal Theory, Networking and Communications, Periodista Daniel Saucedo Aranda, 18071 Granada, Spain*)

E-mail: javierrp@ugr.es

References

1 Vapnik, V.N.: 'Statistical learning theory' (John Wiley & Sons, Inc., New York, 1998)
2 Enqing, D., Guizhong, L., Yatong, Z., and Xiaodi, Z.: 'Applying support vector machines to voice activity detection'. 6th Int. Conf. on Signal Processing, 2002, Vol. 2, pp. 1124–1127
3 Ramírez, J., Segura, J.C., Benítez, C., De La Torre, A., and Rubio, A.: 'Efficient voice activity detection algorithms using long-term speech information', *Speech Commun.*, 2004, **42**, (3–4), pp. 271–287
4 Sohn, J., Kim, N.S., and Sung, W.: 'A statistical model-based voice activity detection', *IEEE Signal Process. Lett.*, 1999, **16**, (1), pp. 1–3
5 Marzinzik, M., and Kollmeier, B.: 'Speech pause detection for noise spectrum estimation by tracking power envelope dynamics', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (6), pp. 341–351
6 Woo, K., Yang, T., Park, K., and Lee, C.: 'Robust voice activity detection algorithm for estimating noise spectrum', *Electron. Lett.*, 2000, **36**, (2), pp. 180–181
7 Li, Q., Zheng, J., Tsai, A., and Zhou, Q.: 'Robust endpoint detection and energy normalization for real-time speech and speaker recognition', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (3), pp. 146–157