# Using the Support Vector Algorithm to Detect Voiced Segments and Eliminate Unvoiced Frames

Damyan Damyanov [1], Zlatka Nikolova [2]

*Abstract* – **This paper introduces a new algorithm for the detection of voiced speech, which functions as a classifier for frames of recorded speech. In this context, that means that it decides whether or not the frame contains good quality voice data as the frame could contain silence, unvoiced speech, or degraded speech which would be unusable without additional processing. The problem considered in this paper is a very important one and many classical methods exist for solving it. Regardless of the possible advantages of these methods, they all suffer from one major drawback – their efficiency is low, which necessitates some degree of post-processing in order to achieve a high recognition ratio. The Support Vector Algorithm achieves an excellent classification of voiced and other speech frames by the use of nonparametric classification and training.**

*Keywords* – **support vector machine, speech recognition, classifying of voiced speech frames.**

## I. Introduction

There are two main types of algorithms used in classifiers: parametric ones, in which a priori knowledge of data distributions is assumed, and non-parametric ones, in which no such a priori knowledge is assumed.

Neural networks, fuzzy systems, and support vector machines (SVMs) are typical non-parametric classifiers. Through training using input-output pairs, classifiers acquire decision functions that classify an input datum into one of the given classes [1].

SVM are preferred to neural-networks, because of their better generalization ability in speech processing. Three-layer neural networks are universal classifiers in that they can classify any labeled data correctly if there are no identical data in different classes. In training multilayer neural network classifiers, network weights are usually corrected, so that the sum-of-squares error between the network outputs and the desired output is minimized. However since the decision boundaries between classes acquired by training are not directly determined, classification performance for the unknown data, i.e. the generalization ability, depends on the training method. And it degrades greatly when the number of training data is small and there is no class overlap.

Conversely, in training support vector machines the decision boundaries are determined directly from the training data so that the separating margins of decision boundaries are maximized in the high-dimensional space called feature space. This learning strategy minimizes the classification errors of the training data and the unknown data.

On the other hand, speech recognition by humans does not pose a problem [2]: One can understand very well that something has been said, even if its information is lost, for example due to loud ambient noise, or when a language, not known by the listener, is spoken. In practical applications this detection of a signal's voiced frames is very useful. It can be employed in speech databases, where the type of a speech should be automatically determined for additional information generation and context-based indexing, or in full automatic speech recognition systems. In the latter case, the room where the acoustic signal propagates is scanned with the aim of recognizing one particular speaker.

Early work in speech/non speech signal classification was carried out, for example, by Hoyt and Wechsler [3]. This research uses a common architecture of signal classifiers: The incoming signal is segmented into frames; discriminative features are extracted and finally classified. Classification of the latter with support vector machines reportedly achieved a recognition rate in excess of 99% [4].

## II. Support vector algorithm

Speech is a phenomenon which constantly changes with time. One approach to registering the non-stationarity of the speech signal is to process it as a temporal sequence of an alphabet of states. The concept beyond the suggested speech detection algorithm is that non-speech signals, which are non-stationary, produce sequences different from those found in speech [5, 6]. This is true, when using the same alphabet for speech and non-speech signals. It is well known [2] that non-speech signals which are stationary are very easy to recognize. Some new efficient methods suggested in [7, 8] can be used.

The parts of speech considered here are voiced frames, the most important components of which are phonemes. In normal speech these are combined to form syllables, words, phrases and more complicated structures. The most important property of voiced-frame detection is the syllabic structure: Normal speech changes constantly between vowels, which are contained in frames with high frame energy, and other phonemes with lower frame energy. Both the vowels and all other types of phonemes possess different spectral features. If a classification algorithm is based on phonemes, then vowels are most conveniently incorporated into a frame model: The support vector machine will discriminate between voiced-like and unvoiced-like frames in certain temporal sequences.

The experiment to derive the speech recognition algorithm was composed of four stages: (1) frame extraction, (2) cepstral Linear Prediction Coefficients (LPC) extraction, (3) segmentation of voiced-unvoiced frames and (4) training and classifying with the support vector algorithm – (Fig.1).

[1] Damyan Damyanov is with the Dept. of Radiocommunications and Video Technology, Technical University of Sofia, Bulgaria, e-mail: ellov@abv.bg

[2] Zlatka Nikolova is with the Dept. of Communication Networks, Technical University of Sofia, Bulgaria, e-mail: zvv@tu-sofia.bg

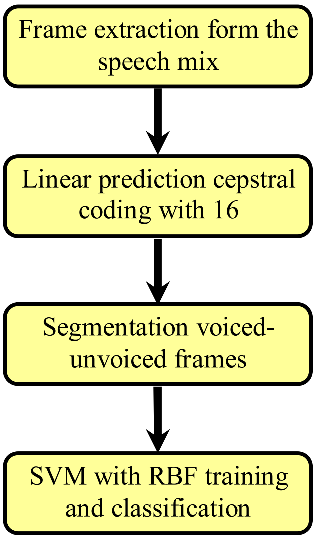
Frame extraction form the speech mix
Linear prediction cepstral coding with 16
Segmentation voiced-unvoiced frames
SVM with RBF training and classification

Fig.1: Support Vector Algorithm

## III. Classification

In the initial frame extraction step, the signal is recorded at a sampling frequency of 11025 Hz. This is a very good sampling frequency for speech signals, since the frequency spectrum of these seldom exceeds 5 kHz. It also provides good speaker differentiation, compact and de-correlated representation of the signal and computational efficiency.

Next, the signal is windowed into frames of 23 ms or 253 samples. The frames are then multiplied by a rectangular window. Whichever window is taken for frame smoothing, there is always a problem with the samples located on both edges of the frame; their spectrum, and hence their energy, is attenuated by the window. For this reason a half-time overlap of the frames has been chosen. This means approx. 11 ms, or 126 samples of the signal. As classification features, linear prediction cepstral coefficients of order 16 were chosen. This employs the well-known formula:

$$M = F_s + (4\ or\ 5) \tag{1}$$

where $M$ is the number of the cepstral LPC coefficients, $F_s$ is the sampling frequency in kHz. The first term in the equation accounts for proper representation of $F_s$ poles from the all-pole of the vocal tract, and the 4 or 5 poles are added for better modelling of the glottal pulse in the voiced case.

The cepstral LPC coefficients have been evaluated, using the recursive formula:

$$y(n,m) = \begin{cases} \log\widehat{\Theta}_0(m), & n = 0 \\ a(n;m) + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right) y(k;m)\, a(n-k;m), & n > 0 \end{cases} \tag{2}$$

where $m$ is the size of the window used, $n$ is the current number of the coefficient. $\widehat{\Theta}_0$ is the first sample of the all-pole filter model of the vocal tract, $a$ is the LPC coefficient, and y is the cepstral LPC coefficient.

Next, points of maximal acoustic change (acoustic landmarks) should be evaluated.

As a basis for the detection of acoustic landmarks, a Euclidean distance measure between the cepstral coefficients, $\bar{c}_i$, of frames with a certain time difference, is employed:

$$d_i = \left\| \bar{c}_i - \bar{c}_{i-k} \right\|_2 \tag{3}$$

where $k$ is the time difference between the frames, and $i$ is the current frame. In this paper, the time difference is set to approximately 11 ms, or 126 samples (i.e., $k$=1). The peaks of $d_i$ can be interpreted as points where the signal spectrum exhibits more significant changes compared to the vicinity, such as phoneme boundaries or instants of rapid spectral change in interferences. Measures like $d_i$ in general produce a huge number of peaks which are not easy to analyse. In this case, smoothing might be preformed: The frame-to-frame Euclidean distance is convolved with a "Mexican hat" function:

$$\Psi(x) = \frac{1}{\left(\sqrt{2\pi\sigma^3}\right)\left(1 - \frac{x^2}{\sigma^2}\right)} \exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{4}$$

This is the popular second derivative of a Gaussian, with standard deviation $\sigma$ set to 3,6 ms. The peaks of the resulting acoustic change function are interpreted as acoustic landmarks and used for segmenting the waveform if they exceed a certain threshold (see Fig.2).

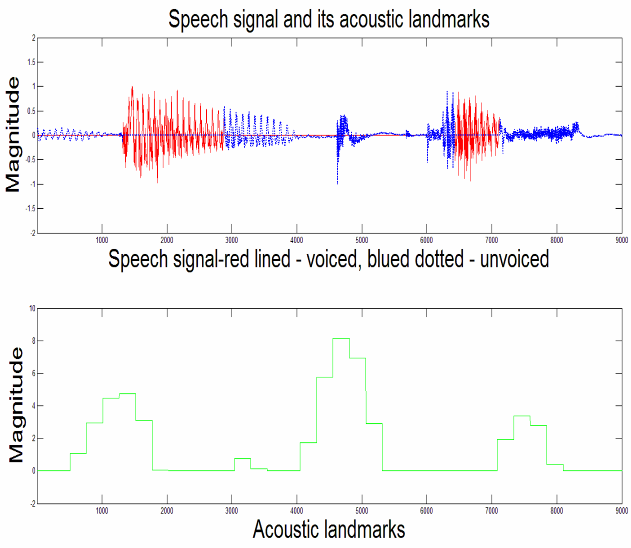

Fig.2: Speech Signal and Acoustic Landmarks

Classification of each frame is performed with a support vector machine. A kernel, based on Radiant Based Function (RBF), has been used for the classification. The analytical expression of the latter is:

$$H(x,x') = \exp(-\gamma \| x - x' \|^2) \tag{5}$$

where the operator $\| \ \|$ denotes the Euclidean distance, and $x'$ means in general different variable than $x$. The parameter $\gamma$ controls the radius of the function. Its typical values are from $0.9 \div 1$ [1]. In our case it is set to 1. The method for the evaluation of the separating hyper plane has been set to Least Squares. Thus the SVMs are able to discriminate between classes via topologically complex hyper planes and are computationally efficient at the test stage. The target classes are {voiced, unvoiced}.

## IV. Experiments and Discussions

For the experiment a private database has been used. 3000 frames from 30 different speakers have been classified. 300 frames were used for training the support vector machine, and the other 2700 were used in classification.

One utterance from the upper is shown in Fig.3.

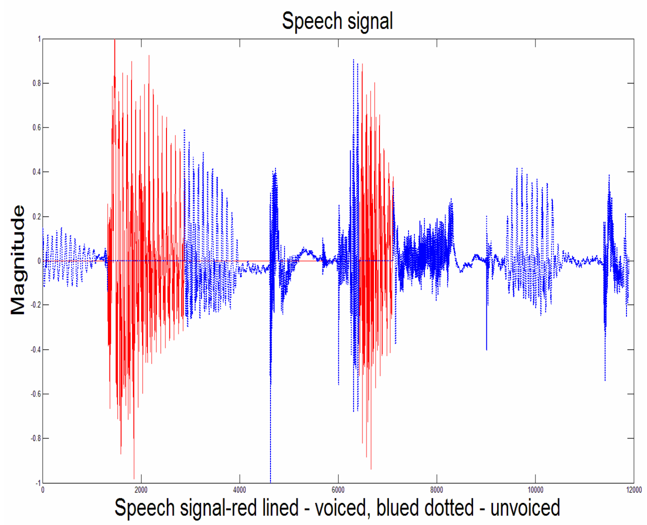

Fig.3: Example for a male utterance. The red straight line represents the voiced frames, and the blue dotted line – the unvoiced ones

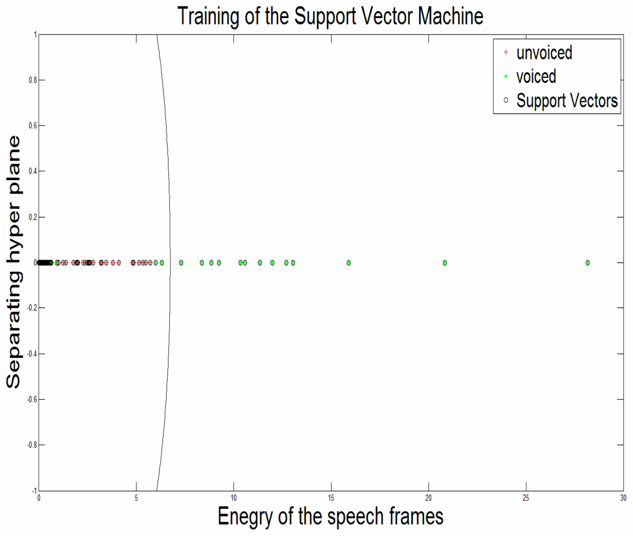

Fig.4: Training of the support vector machine. The red crosses on the left side represent the unvoiced frame energies, and the green crosses on the right side represent the voiced frame energies

The training phase produced the results shown in Fig 4. As can be seen from the latter, this first phase of the classification algorithm has been very effective. The frame energies lie far enough apart from each other to draw a parabola between the two classes and to separate them easily. The classification phase for the utterance from Fig.3 is shown in Fig.5. As expected, the separating parable performs well in separating the two target classes. This can be seen if we show how a classical algorithm performs against the support vector one. As a classical approach, the Spectral Auto-Correlation Peak-to-Valley Ratio (SAVPR) has been chosen, for its simplicity of implementation. It is a parametric approach, meaning that it doesn’t need training [5, 6].

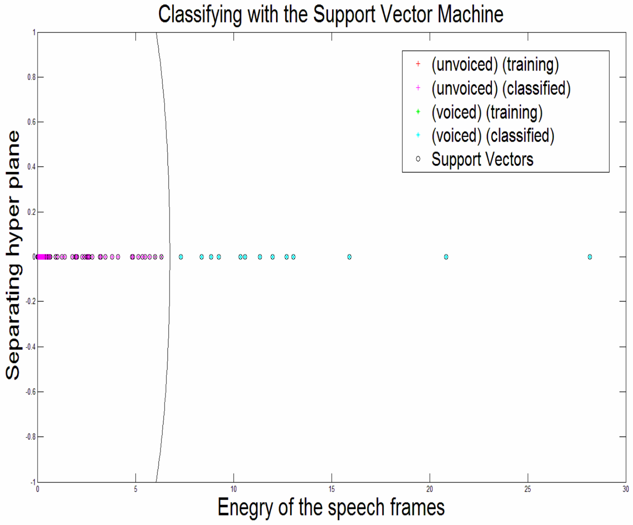

Fig.5: Classifying with the support vector machine. The red and magenta crosses on the left side represent the unvoiced frame energies, and the green and cyan crosses on the right side represent the voiced frame energies

The classification with the SAVPR method for the utterance is shown on Fig. 6, the one with the SVM is shown in Fig. 7, and both are plotted together on Fig. 8. As can be seen, the performance of the support vector algorithm is much better than that shown by the spectral auto-correlation peak-to-valley ratio algorithm. After the classification, using the database mentioned at the beginning of the chapter, the following results were obtained, demonstrating that the overall performance of the SVM is better [6] (Table I).

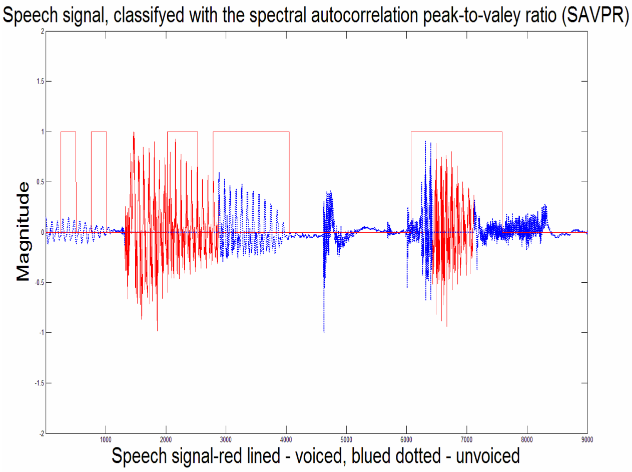

Fig.6: Classification with the spectral auto-correlation peak-to-valley ratio

TABLE I

| | Classification with the SAVPR algorithm | Classification with the SVM algorithm |
|---|---|---|
| Male speech | 61.70 % | 92.40 % |
| Female speech | 60.10 % | 91.50 % |

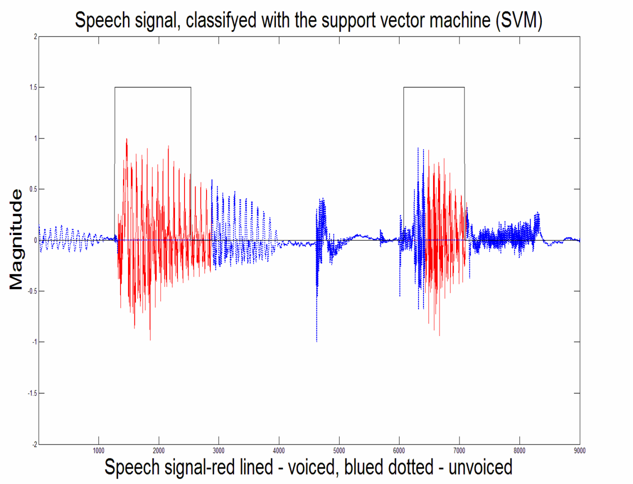

Fig.7: Classification with the spectral SVM

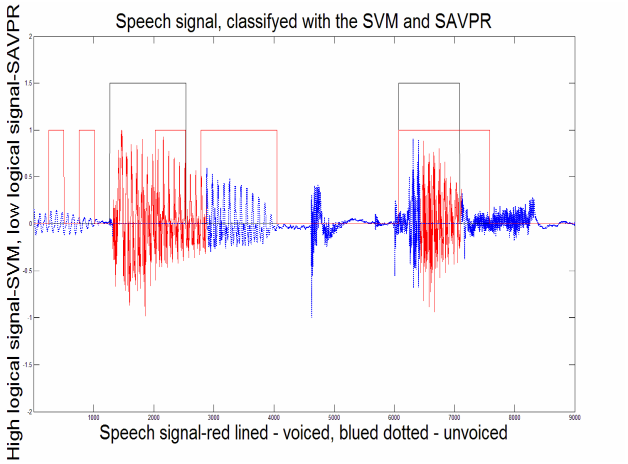

Fig.8: Classification with the SAVPR (low red logical signal) and SVM (high black logical signal)

## V. CONCLUSIONS

The support vector algorithm has shown some very good features in the recognition of voiced speech segments – it is non-parametric, requires a small amount of training data and is computationally very effective. In comparison with classical approaches it performs much better, suggesting, that is can be used in "speech overlap detection". This is where two or more people speak simultaneously and recognition of one of the speakers is required. The future work of the authors is to utilize the support vector machine in the field of speech overlap detection.

## REFERENCES

[1] Shiego Abe, "Support vector machines for pattern classification", *Springer,* 2005.

[2] John Proakis "Discrete-Time Processing of Speech Signals", *IEEE press*, 2000.

[3] J.D.Hoyt, and H. Wechsler "Detection of human speech using hybrid recognition models", *Proc. ICASSP-94,* Vol.I, pp. 330-333. 1994.

[4] G. Heinrich, "Speech identification using a sequence-based heuristic", *47th International Symposium ELMAR-2005*, 08-10 June 2005, Zadar, Croatia.

[5] Bekiarski Al, and L. Docheva, "Investigation of back Propagation Algorithm Implementation in Analog Neural Networks ", *ICEST,* 2005, Nis Serbia and Montenegro, pp. 637-640.

[6] S. G. Pleshkova-Bekiarska and D. A. Damyanov, "Speech overlap Detection Algorithms Simulation", *ICEST*, Ohrid pp. 495–499, 2007.

[7] Iliev G, Z. Nikolova, V. Poulkov, G. Stoyanov, "Noise Cancellation in OFDM Systems Using Adaptive Complex Narrowband IIR Filtering", *IEEE International Conference On Communications (ICC-2006)*, Istanbul, Turkey, 11-15 June 2006, pp. II-451 - II-455.

[8] Nikolova Z., V. Poulkov, G. Iliev, G. Stoyanov, "Narrowband Interference Cancellation in Multiband OFDM Systems", *3rd Cost 289 Workshop "Enabling Technologies For B3g Systems"*, Aveiro, Portugal, 12-13 July 2006.