Journal of Applied Research and Technology

www.jart.ccadet.unam.mx
Journal of Applied Research and Technology 17 (2019) 8-19

Original

Voice activity detection using smoothed-fuzzy entropy (*sm*FuzzyEn) and support vector machine

R. Johny Elton [a, *], P. Vasuki [a], J. Mohanalin [b], J. S. Gnanasekaran [c]

[a] *Department of Electronics and Communication Engineering, K.L.N college of Information Technology, Madurai – 630612, Tamilnadu, India*

[b] *Department of Electrical and Electronics Engineering, College of Engineering, Pathanapuram, Kerala – 689696, India*

[c] *Department of Mechanical Engineering, K.L.N college of Information Technology, Madurai – 630612, Tamilnadu, India*

Received dd mm aaaa; accepted dd mm aaaa
Available online dd mm aaaa

Abstract: In this paper a novel voice activity detection approach using smoothed fuzzy entropy (*sm*FuzzyEn) feature using support vector machine is proposed. The proposed approach (*sm*FESVM) uses total variation filter and Savitzky-Golay filter to smooth the FuzzyEn feature extracted from the noisy speech signals. Also, convolution of the first order difference of TV filter and noisy fuzzy entropy feature (*con*FETV') is also proposed. The obtained smoothed feature vectors are further normalized using min-max normalization and the normalized feature vectors train SVM model for speech/non-speech classification. The proposed *sm*FESVM method shows better discrimination of noise and noisy speech when tested under various nonstationary background noises of different signal-to-noise ratio levels. 10 – fold cross validation was used to validate the efficacy of the SVM classifier. The performance of the *sm*FESVM is compared against various algorithms and comparison suggests that the results obtained by the *sm*FESVM is efficient in detecting speech under low SNR conditions with an accuracy of 93.88%.

Keywords: Voiced Activity Detection, Fuzzy Entropy, Support Vector Machine, Savitzky-Golay filter, Total variation filter

1. INTRODUCTION

Voice activity detection (VAD) tries to detect speech segments from background noises. It is an important speech processing technique which is used in various applications such as automatic speech recognition (ASR) (Hernández-Mena, Meza-Ruiz, & Herrera-Camacho, 2017; Karray & Martin, 2003), mobile communications (Freeman, Cosier, Southcott, & Boyd, 1991), VoIP (Sangwan et al., 2002; Zhang, Gao, Bian, & Lu, 2005), and noise suppression in digital hearing aids (Itoh & Mizushima, 1997). The outcome of VAD is usually binary decisive, which indicate absence of speech (noise only segments indicated as HRns) or presence of speech (noisy speech segments, indicated as HRs). The challenge to the VAD is to detect speech under low signal-to- noise ratio

* Corresponding author.
E-mail address: erjohnyelton@gmail.com (R. Johny Elton).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

(SNR) scenarios and also under the influence of nonstationary noises. During the past few decades, researchers have tried many approaches to improve the VAD performance under low SNRs and for various types of noises. Still there is no unique feature or set of features identified that would improve the performance of the VAD and so the problem remains challenging to explore new robust features and algorithms.

The performance of the VAD algorithm relies on two key stages namely, features to VAD and HRs and HRns classification. Simple features such as short-term energy (Hsieh, Feng, & Huang, 2009), zero crossing rate (Kotnik, Kacic, & Horvat, 2001) were considered in early days. But the performance lacks at low SNR levels as well as with non-stationary noises. These drawbacks are overcome by robust features like spectrum (Davis, Nordholm, & Togneri, 2006), autocorrelation (Shi, Zou, & Liu, 2014), power in the bandlimited region (Marzinzik & Kollmeier, 2002), wavelet coefficients (LEE, 2006), higher-order statistics(Nemer, Goubran, & Mahmoud, 2001), etc. These robust features shows better performance when the SNR levels are above 10dB assuming that the background noise is stationary for certain period but still lacking at low SNR levels, because at low SNR levels the structural and spectral properties of the speech signals get distorted (Khoa, 2012). Few algorithms improve VAD performance by estimating noise during this stationary period which is computationally expensive. So in order to improve the performance of VAD, entropy based features were considered. Entropy is a powerful tool to detect speech from noisy signal which was first introduced by Shannon to estimate the uncertainty in a signal (Wu & Wang, 2005). This can be used in both time and frequency domain called spectral entropy. Similarly fuzzy entropy (FuzzyEn) which is a modified algorithm of sample entropy (SampEn) (Chen, Wang, Xie, & Yu, 2007; Chen, Zhuang, Yu, & Wang, 2009; Richman, & Moorman, 2000) is based on fuzzy set theory and is used to measure the complexity of the time series data. FuzzyEn retains certain characteristics of SampEn like excluding self-matches and also it overcomes the limitations of SampEn by using an exponential function to select or discard the similarities between the two vectors rather using a Heaviside function. Additionally, by inheriting the similarity measurement using fuzzy sets, the limitations cited by SampEn which uses the Heaviside function as the tolerance to select or discard the similarities between the two vectors was overcome by FuzzyEn, as FuzzyEn transits smoothly through varying parameters with the use of the exponential function.

Finally, to classify for HRs and HRns, an adaptive threshold based on the features extracted from the speech signals can be used or with the use of machine learning algorithms (MLA). Different classifiers based on machine learning algorithms (MLA) are also invoked for HRs and HRns classification. Neural networks have been widely used, but its training procedures are cost expensive. Another popularly used MLA is the support vector machine (SVM) (Shabat & Tapamo, 2017; Nazir, Majid-Mirza, & Ali-Khan, 2014). It is a powerful tool used in classification because of its convergence speed in training phase which is faster than that of other classifiers. In this paper, SVM (Cortes & Vapnik, 1995) proposed by Vapnik is used to classify HRs and HRns because of its applications in audio classification (Guo & Li, 2003), pattern recognition (Ganapathiraju, Hamaker, & Picone, 2004), etc. The FuzzyEn feature is computed over the short term analysis frames (usually 20 – 40 ms). Instead of using noisy FuzzyEn feature directly as an input to SVM, the obtained FuzzyEn features are smoothed to attenuate the noisy features using two filters namely, total variation (TV) filter and Savitzky – Golay (SG) filter. Total variation (TV) filtering introduced by Rudin, Osher, and Fatemi (1992) and Chan, Osher, and Shen (2001), produces nonlinear function of the data, which is defined by the minimizing a non-quadratic cost function. Even though the output of the TV filter produces “staircase effect”, it flattens out the rapid fluctuations that’s available in the data. This is mainly due to the regularization parameter (lambda) present in the cost function. The larger the (lambda) value the staircase effect is obtained which removes major noisy information in the data. SG filter (Savitzky & Golay, 1964) which was proposed by Savitzky and Golay, can be generalized as a least-squares smoothing filter, where the filter coefficients are obtained using least-squares fit using a polynomial degree. The effects of smoothing is controlled by two parameters namely the size of the window and the degree of the polynomial. The advantages with SG filter is that it smooths the data to reduce the noisy information by preserving the shape and height of the waveforms.

In this paper, the feature vector consists of TV-filter smoothed FuzzyEn feature, SG filter smoothed FuzzyEn feature and convolution of first order difference and

FuzzyEn. This feature vector is fed as the input into the SVM for VAD, and its performance was investigated under various noisy conditions (airport, babble, car, and train) at different SNR levels (−10 dB, −5 dB, 0 dB, 5 dB, and 10 dB). The structure of the rest of this article is arranged as follows. Section 2 discusses the various stages of proposed algorithm which includes feature extraction, feature vector formation, etc and Section 3 presents the speech and noise database and metrics used in the evaluation along with the experimental results. Finally, a conclusion of this work is given in Section 4.

2. PROPOSED METHODOLOGY

Voice activity detection usually addresses a binary decision to detect the presence of speech for each frame of the noisy speech signal. The noisy speech signal $s(n)$ is obtained by corrupting the clean speech signal $x(n)$ by the additive noise $v(n)$, as in (1),

$$s(n) = x(n) + v(n) \tag{1}$$

The proposed smFVAD block diagram is shown in Fig.1. The motivation for the proposal is to identify robust feature vectors that would improve the accuracy of the VAD and speech detection under low SNR conditions. Since speech signal is nonstationary in nature, the obtained noisy signal is divided into sequence of small frames of size ranging between 20 – 40 ms. In this paper, the size of each frame is 32ms with a frame shift of 10 ms which is windowed by Hanning window. Therefore each frame consists of 512 samples each and the total number of frames varies depending on the size of the speech signal. The major blocks of the proposed VAD are explained in detail in the following subsections.

2.1 FEATURE EXTRACTION - FUZZY ENTROPY (FuzzyEn)

Let $s_k(i)$ be a N sample noisy speech sequence of k^{th} frame, where i = *1, 2, 3, …, N*, which is reconstructed by phase-space with an embedded dimension m, and the reconstructed phase-space speech vector S_i^m is given by in (2),

$$S_i^m = \{s(i), s(i+1), \dots, s(i+m-1)\} - s0(i), \quad i = 1,2, \dots, N-m+1 \tag{2}$$

and is generalized by removing the baseline as in (3),

$$s0(i) = m^{-1} \sum_{j=0}^{m-1} s(i+j) \tag{3}$$

For given vector S_i^m, the similarity degree D_{ij} of its neighboring vector S_j^mthrough its similarity degree is defined by fuzzy function, given in (4),

$$D_{ij} = \mu(d_{ij}^m, r) \tag{4}$$

and d_{ij}^m is the maximum absolute difference of the scalar components of S_i^m and S_j^m, given in (5),

$$d_{ij}^m = d[S_i^m, S_j^m] = \max_{l \in (0,m-1)} [(s(i+l) - s_0(i)) - (s(j+l) - s_0(j))] \tag{5}$$

Here $\mu(d_{ij}^m, r)$ is the fuzzy membership function, which is given by the exponential function, as in (6),

$$\mu(d_{ij}^m, n, r) = exp\left(\frac{-(d_{ij}^m)^n}{r}\right) \tag{6}$$

where n and r are the gradient and width of the fuzzy membership function.

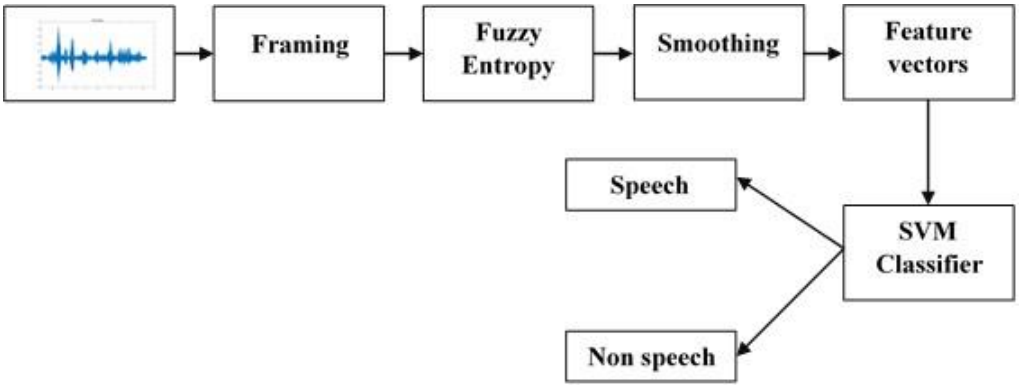

Fig. 1. Block diagram of *sm*FuzzyEn VAD.

For each S_i^m, averaging all similarity degree of the neighboring vectors S_j^m, we get (7),

$$\Phi_i^m = \frac{1}{(N-m-1)} \sum_{j=1\, j\neq i}^{N-m} D_{ij}^m \tag{7}$$

Now construct (8) and (9),

$$\varphi^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \Phi_i^m(r) \tag{8}$$

and

$$\varphi^{m+1}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \Phi_i^{m+1}(r) \tag{9}$$

From this, FuzzyEn(m,r) of the speech, is defined by, given in (10),

$$\text{FuzzyEn}(m,r) = \lim_{N\to\infty} [\ln \varphi^m - \ln \varphi^{m+1}(r)] \tag{10}$$

In this work, the embedding dimension, m is 2 and the exponential function parameters n and r are set as 2 and 0.2 times standard deviation, respectively. Note, FuzzyEn in (10) will be denoted as nFE hereafter.

2.2 TOTAL VARIATION (TV) FILTER

The total variation (TV) of the fuzzy entropy feature ($\hat{n}FE$) is given by the sum of the absolute values of its first order difference, which is given by,

$$TV(\hat{n}FE) = \sum_{n=2}^{LL} |\hat{n}FE(n) - \hat{n}FE(n-1)| \tag{11}$$

where LL is the length of $\hat{n}FE$. Note TV of the signal nFE can be written as,

$$TV(\hat{n}FE) = \|D\hat{n}FE\|_1 \tag{12}$$

Where

$$\mathrm{D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \ddots & \\ & & & & -1 \quad 1 \end{bmatrix}$$

is a matrix of size (LL - 1) x LL. The notation $\|.\|_1$ denotes the l_1 norm. Given noisy fuzzy entropy feature (nFE) from equation (10), the output of TV filter is defined as the solution $\hat{n}FE$ to the minimization problem,

$$arg \min_{\hat{n}FE} \lambda \|D\hat{n}FE\|_1 + \|nFE - \hat{n}FE\|_2^2 \tag{13}$$

where λ is a parameter that controls the trade-off between denoising and signal distortion.

2.3 SAVITZKY – GOLAY (SG) FILTER

The FuzzyEn features (nFE) obtained using (10) are further smoothed using Savitzky-Golay (SG) filter. SG filter can be generalized as a least-squares smoothing filter, where the filter coefficients are obtained using least-squares fit using a polynomial. This filter depends on two main parameters, window size and degree of the polynomial, M. If M is too high, redundancies of data is obtained and when M is too low, the signal gets distorted. Similarly, when the size of the window is larger, valid information may be lost and when the window is smaller, poor denoised signal is obtained. Smoothing the obtained noisy features through SG filter is given by,

$$FES_k = \frac{\sum_{i=-m}^{m} c_i\, nFE_{k+i}}{NN} v \tag{14}$$

where nFE_k is the noisy FE features and FES_k is the smoothed output of the SG filter, c_i is the coefficient for the i-th smoothing, NN is the number of data points in the smoothing window and is equal to 2m + 1, where m is the half-width of the smoothing window. The essence of SG filtering is adopting a polynomial in a sliding window to fit the original signal piece-by-piece depending on the least-squares estimation algorithm. The polynomial can be modelled as:

$$p_M = a_0 + a_1 k + \cdots + a_M k^M \tag{15}$$

2.4 CONVOLUTION OF ABSOLUTE FIRST ORDER DIFFERENCE OF TV FILTER (conFETV')

The output of the TV filter (FETV) is further optimized by computing the absolute first order difference of the TV filter which is given by,

$$FETV' = |FETV(n) - FETV(n-1)| \tag{16}$$

$$FETV' = \begin{cases} FETV', & FETV' > \mu_{FETV'} \\ 0, & FETV' < \mu_{FETV'} \end{cases} \tag{17}$$

where, $\mu_{FETV'} = \sum_{i=1}^{LL} FETV'$.

The absolute first order difference ($FETV'$) obtained in equation (16) is further refined by replacing with zeros for values of $FETV'$ less than average of $FETV'$ ($\mu_{FETV'}$) and the resultant obtained is convoluted with the Fuzzy entropy, FE obtained in equation (10) which is given by,

$$conFETV'[n] = nFE[n] * FETV'[n] \tag{18}$$

$$conFETV'[n] = \sum_{k=0}^{LL-1} FETV'[k]\, nFE[n-k] \tag{19}$$

where * is a convolution operator.

2.5 FORMATION OF FEATURE VECTOR

The nFE obtained using (10) is further processed as explained in section 2.2 to 2.4 to form the feature vector of dimension J x K, where J is the number of frames in the noisy speech signal and K to be 3. The first feature vector is obtained as the output of TV filter which flattens out the rapid changes in the signal due to "staircase-effect" while preserving the slow changes in the signal. The second feature vector obtained as a result of SG filtering preserves the peak information in the signal. The third feature vector which was obtained as a result of convolution shows better discrimination between HRs and HRns regions. The significance of the feature vectors considered is explained in Figure 2.

2.6 FEATURE SCALING – minmax

This minmax scaling method, rescales the given feature vectors from one range of values to a new range of values. More often, the feature vectors are rescaled to lie within a range of [0, 1] or [-1, +1], depending on the output classes. This rescaling is accomplished by using a linear interpretation equation given in (20),

$$sFE' = \frac{smFE - \min(smFE)}{\max(smFE) - \min(smFE)} \tag{20}$$

where $smFE$ is the smoothed feature vectors obtained using TV filter, SG filter and $conFETV'$ and sFE' is the normalized feature vector which is rescaled to fit the range [0, 1].

2.7 CLASSIFIER - SUPPORT VECTOR MACHINE

In this work, SVM is used as a classifier, because of its effectiveness in classification accuracy and computational time than other conventional nonparametric classifiers such neural networks, kNN, etc. SVM constructs an optimal hyperplane $[\langle w, x\rangle + b = 0]$ that maximizes the margin using a known kernel function that accurately predicts the unknown data into two classes, where w and b, shall be derived based on the classification accuracy of the linear problems. Let $(x_i, y_i)_{i=1}^{N}$ be the training set samples, where $y_i \in R^m$ is the corresponding target classes for the input data $x_i \in R^m$. This is achieved by minimizing the error function shown in (21),

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &Subject\ to\ y_i[w^T\varphi(x_i) - b] \geq 1 - \xi_i \\ &\xi_i \geq 0, i = 1,2,\ldots,n \end{aligned} \tag{21}$$

where $\varphi(x_i)$ is a mapping function to map x_i to its higher dimensional feature space, ξ_i is the misclassification error and C controls the tradeoff between the cost of classification and the margin. The classification of the new data as +1 or −1 is obtained by minimizing the error function in (21) based on the decision function,

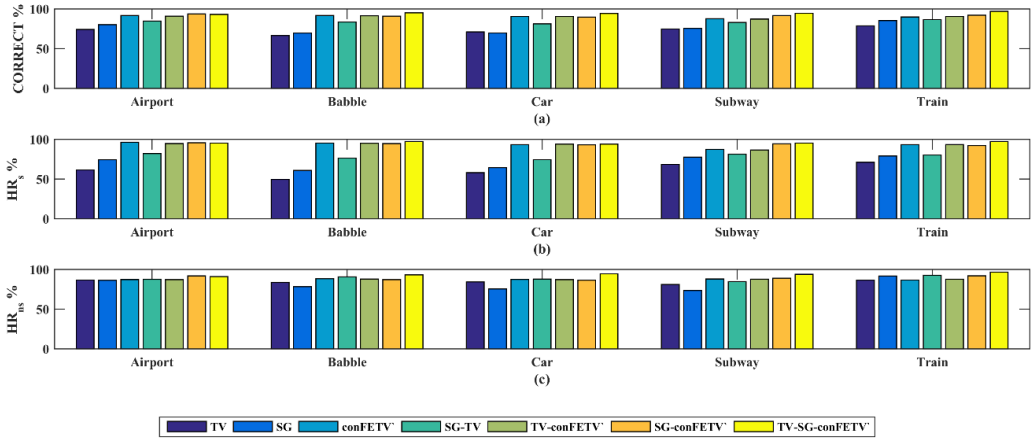

Fig. 2. Performance measures comparison for the feature vectors considered averaged over five SNRs for 5 noises. (a) CORRECT (b) HRs (c) HRns.

Class $i = sign\ (\langle w, x\rangle + b)$, yields 1 or 0 respectively. The mapping of the input training set into a higher dimensional space is done through a kernel function $K\ (x_i\ , y_i)$. In this work, the RBF kernel function is used for its excellent generalization and low computational cost (Hariharan, Fook, Sindhu, Adom, & Yaacob, 2013). The RBF kernel function is given by (22),

$$K\ (x_i\ , y_i) = \exp\left(\frac{-(x_i - y_i)^2}{2\,\sigma^2}\right) \tag{22}$$

where, the parameter σ is the width of the Gaussian function. For this given kernel function, the error function of the classifier is given by (23),

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i\ K(x, x_i) + b\right) \tag{23}$$

3. RESULTS AND DISCUSSION

The actual speech signals which is collected from TIMIT database (Garofolo et al., 1993) is contaminated by adding nonstationary noises collected from AURORA2 database (Hirsch & Pearse, 2000) of different SNR levels (-10 dB to 10 dB). TIMIT signals were preferred because it provides transcriptions down to word and phoneme levels. Each TIMIT sentence is almost around 3.5s long, out of which 90% is the actual speech signal and 10% contains silence (non-speech) regions. To change this ratio of speech and non-speech regions to 40 % and 60 % (Beritelli, Casale, Ruggeri, & Serrano, 2002) respectively, silence is added to the original speech of the TIMIT corpus. Five types of nonstationary noises such as airport, babble, car, subway and train noises are considered for the experiment which is resampled to 16 kHz depending on the need.

3.1 PERFORMANCE EVALUATION

Performance evaluation of VAD algorithm can be performed both subjectively and objectively. In subjective evaluation, a human listener evaluates for VAD errors, whereas, numerical computations are carried out for objective evaluation. However, subjective evaluation alone is insufficient to examine the VAD performance because listening tests like ABC fail to consider the effects of false alarm (Beritelli et al., 2002; Ghosh, Tsiartas, & Narayanan, 2011). Hence numerical computations through the proposed VAD algorithm.

VAD performance is calculated using (24) and (25)

$$HRns = \frac{NS_{ns,ns}}{NS_{ns,ref}} \tag{24}$$

and

$$HRs = \frac{NS_{s,s}}{NS_{s,ref}} \tag{25}$$

where, HR_{ns} and HR_{s}, non-speech frames and speech frames correctly detected among non-speech and speech frames respectively. NS_{ns} and NS_{s} refers to the number of non-speech and speech frames in the whole database, respectively, while $NS_{ns,\ ns}$ and $NS_{s,\ s}$, refers to the number of frames classified correctly as non-speech and speech frames. The overall accuracy rate is given by (26),

$$CORRECT = \frac{NS_{ns,ns} + NS_{s,s}}{NS_{ns,ref} + NS_{s,ref}} \tag{26}$$

The best performance is achieved when three parameters referred in the equations (24), (25) and (26) become maximum.

Feature vector performance metrics:

The performance metrics (CORRECT, HRs, HRns) for the proposed TV – SG – $conFETV'$ feature vector averaged for overall noises considered (Airport, Babble, Car, Subway and Train) for different SNR levels is shown in Table 1. The table clearly shows that the feature proposed $conFETV'$ achieves better performance measures except for HRns. It also detects speech regions (HRs) in an effective way. But when combining all the three features namely TV, SG and $conFETV'$, the performance increases for all three metrics giving better detection of HRs and HRns of about 90% and above when compared to that of features considered as a single one. Figure 2 (a) – (c) shows the performance evaluation metrics averaged over the five SNRs (-10 to 10) for the five nonstationary noises computed for the feature vectors TV, SG, $conFETV'$, TV – SG, TV – $conFETV'$, SG – $conFETV'$, TV – SG – $conFETV'$ (proposed feature vector) considered. From the Figure 2, it is clear that the proposed TV – SG – $conFETV'$ feature vector is the best for all the performance metrics considered for CORRECT, HRs and HRns. The significance of the feature vector formed is

misclassification rate of classifying HRs and HRns. The figure clearly shows that the error rate is minimal while forming the feature vector (TV – SG – *conFETV′*), because the proposed feature *conFETV′* detects speech regions (HRs) well when compared to that of -the non-speech regions (HRns) but when combining improves the overall efficiency of detecting HRs and HRns effectively.

Performance metrics comparison with other VADs (G.729B, VAD-SOHN, VAD-RAMIREZ)

The efficiency of the proposed *sm*FESVM based VAD is examined by comparing the performance metrics with existing VAD algorithms which is explained below:

- G.729B VAD (ITU, 1995) is a standard VAD method which utilizes several traditional features and is used in speech communication systems to improve the bandwidth.

-VAD-SOHN (Sohn, 1999) is based on statistical modelling which estimates spectral SNR and Gaussian model distribution for speech and noise assuming that the Gaussian distribution for speech and noise in Fourier domain are independent.

- VAD-RAMIREZ (Ramírez, Segura, Benítez, De la Torre, & Rubio, 2004) combines multiple-observation technique and statistical models VAD and the False Alarm Rate (FAR) is controlled by the use of contextual global hypothesis.

Figure 4 show the comparison of CORRECT rate performance metrics of the proposed *sm*FESVM based VAD with G.729B, VAD-SOHN and VAD-RAMIREZ for different SNR levels (−10,-5, 0, 5 and 10 dB) for airport noise, babble noise, car noise, subway noise, and train noise. It is inferred from the figure that the *sm*FESVM based VAD outperforms the rest of the VAD algorithms by obtaining best performance in CORRECT rate especially under low SNR levels (< 0 dB). There is a noticeable lag in car noise scenario, because a complex event has been encountered in the car noise scenario which is treated as valuable speech thereby increasing misclassification error.

Figure 5 and Figure 6 show hit rate performance evaluation metrics (HRs and HRns) with five SNRs for different kinds of noises computed for G.729B, VAD-SOHN, VAD-RAMIREZ, and the *sm*FESVM based VAD. It is clear that the VAD methods used in comparison yields very high speech detection rates (HRs) for different SNR levels and for various noises, especially G.729B and VAD-RAMIREZ. But this performance behaviour degrades in non-speech detection (HRns) rates when the noise level increases which makes it less conservative in practical speech processing schemes. This is due to the effect of hangover scheme (Aneeja & Yegnanarayana, 2015; Davis et al., 2006) used in these methods. As the energy towards the end of speech is relatively low, to control the risk of false alarm rates (FAR), hangover schemes are used. On the other hand, the proposed *sm*FESVM based VAD yields better detection rates for speech (HRs) and non-speech (HRns) regions when compared to that of the other VAD algorithms. Its detection rate for non-speech regions is relatively higher and exhibits a very low marginal decay in performance while detection speech regions for certain noisy conditions. Table 2 shows the performance metrics averaged for all the noises with various SNR levels. The table clearly illustrates the consistency of the *sm*FESVM based VAD

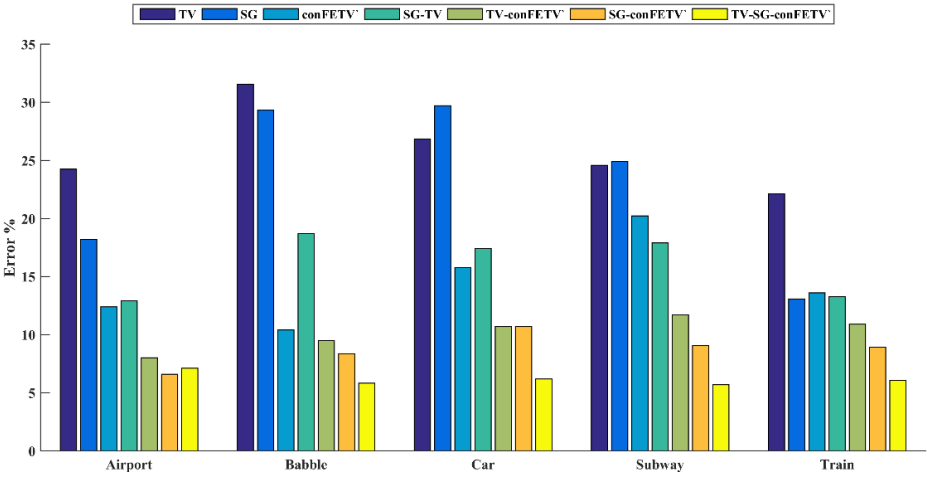

Fig. 3. Misclassification Error % for various feature vectors considered.

in detecting HRs and HRns with different SNR levels. The challenge of VAD is to detect speech under low SNR levels (<0dB) and from the table it is clear that the *sm*FESVM based VAD algorithm outperforms the rest of the VAD algorithms in detecting speech as well as non-speech regions under low SNR levels.

Table 3 shows the comparison of performance metrics under clean conditions, averaged for all noisy scenarios and averaged over both clean and noisy conditions with different SNR levels. It is clearly observed from the table that the *sm*FESVM based VAD excels rest of VAD in every performance metrics especially for non-stationary noisy scenarios. There is a noticeable lag in speech detection rate when compared with other VADs, but the proportion of non-speech detection rate confirms that the proposed *sm*FESVM based VAD is well suited in detecting both speech and non-speech regions effectively. Also, the detection of non-speech regions is very much higher for the *sm*FESVM based VAD which is very useful in major speech applications like speech compression, VoIP, speech enhancement, etc. Therefore, based on these performance metrics, the *sm*FESVM based VAD detects speech and non-speech regions effectively especially under low SNR conditions.

Table 1. Feature vector performance metrics averaged for overall noises.

	SNR	**TV**	**SG**	***conFETV′***	**SG – TV**	**TV – *conFETV′***	**SG – *conFETV′***	**TV – SG – *conFETV′***
CORRECT	-10	71.07	74.14	91.03	84.42	92.85	91.42	94.55
	-5	70.10	75.91	92.47	81.33	90.27	92.59	93.49
	0	72.83	74.49	89.54	83.66	89.45	91.75	92.88
	5	72.40	76.89	89.56	83.04	88.31	91.59	94.26
	10	78.96	79.16	88.96	86.65	89.69	90.86	94.24
HRs	-10	66.61	71.54	93.57	81.03	96.41	95.04	97.24
	-5	61.61	74.30	95.64	79.25	92.91	93.08	96.42
	0	61.69	71.54	91.46	81.25	90..62	93.37	95.07
	5	55.62	68.97	91.87	75.70	91.73	94.69	95.14
	10	62.99	69.72	93.01	77.28	91.99	93.87	94.68
HRns	-10	75.54	76.74	88.49	87.81	89.28	87.80	91.95
	-5	78.60	77.52	89.31	83.41	87.63	92.10	92.63
	0	83.97	77.44	87.63	86.06	88.28	89.53	92.27
	5	89.18	84.81	87.26	90.38	84.89	88.50	92.16
	10	94.92	88.60	84.91	96.03	87.39	87.85	94.66

Table 2. Performance metrics comparison for VAD-SOHN (Sohn, 1999), VAD-RAMIREZ (Ramírez et al., 2004), G.729B (ITU, 1995) and smFESVM averaged over 5 noises for 5 SNR levels.

SNR	**SOHN**			**RAMIREZ**			**G.729B**			***sm*FESVM**		
	COR	**HRs**	**HRns**	**COR**	**HRs**	**HRns**	**COR**	**HRs**	**HRns**	**COR**	**HRs**	**HRns**
-10	80.56	93.77	67.00	79.31	99.00	59.28	65.79	98.80	36.85	94.55	97.24	91.95
-5	82.20	96.97	67.25	80.03	99.00	60.30	66.07	99.40	37.07	93.49	96.42	92.63
0	82.78	98.17	67.20	80.88	99.00	61.88	66.04	99.40	37.05	92.88	95.07	92.27
5	82.92	98.57	67.00	81.49	99.00	63.08	65.85	99.20	36.65	94.26	95.14	92.16
10	82.92	98.77	67.00	81.87	99.00	63.88	65.67	99.20	36.45	94.24	94.68	94.66

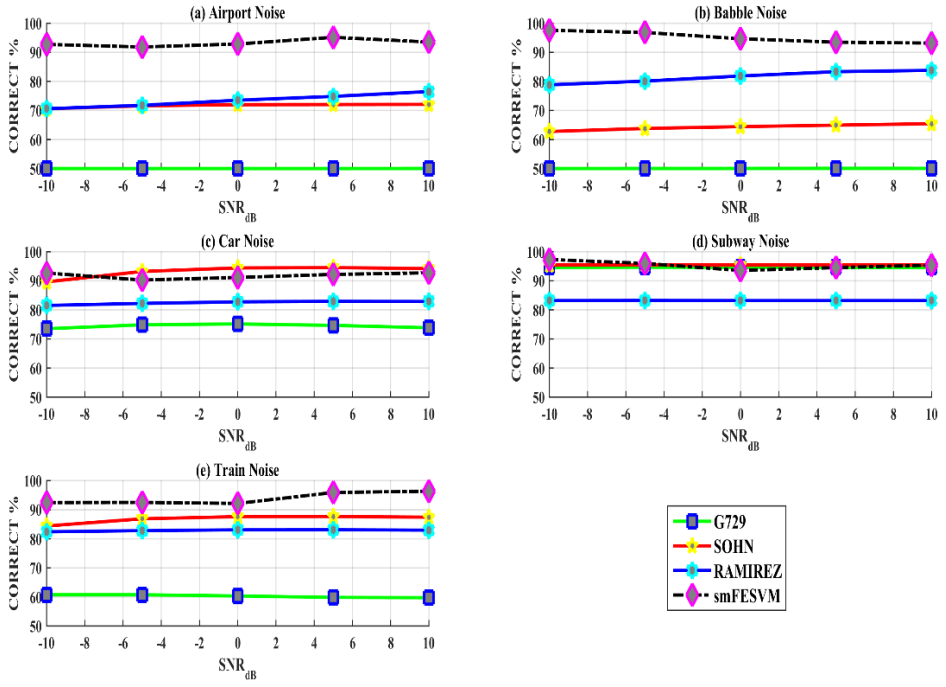

Fig. 4. CORRECT comparisons for G.729B (ITU, 1995), VAD-SOHN (Sohn, 1999), VAD-RAMIREZ (Ramírez et al., 2004) and smFESVM based VAD for SNR values ranging between -10 and 10 dB (a) Airport (b) Babble (c) Car (d) Subway and (e) Train noises.

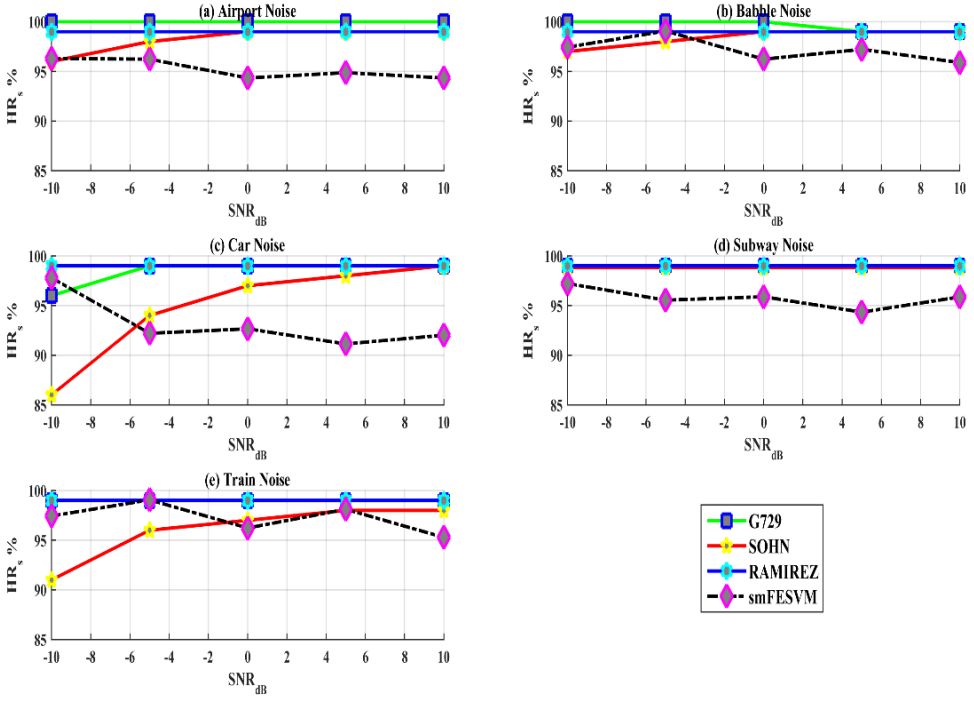

Fig. 5. HRs comparisons for G.729B (ITU, 1995), VAD-SOHN (Sohn, 1999), VAD-RAMIREZ (Ramírez et al., 2004) and smFESVM based VAD for SNR values ranging between -10 and 10 dB (a) Airport (b) Babble (c) Car (d) Subway and (e) Train noises.

Table 3. Overall VAD performance comparison for VAD-SOHN (Sohn, 1999), VAD-RAMIREZ (Ramírez et al., 2004), G.729B (ITU, 1995) and smFESVM averaged for all noises over 5 SNR levels [-10, -5, 0, 5 and 10 dB].

VAD	SOHN	RAMIREZ	G.729B	*sm*FESVM
		NOISES		
CORRECT	78.99	80.09	58.69	93.88
HR$_s$	96.85	99	99.25	95.71
HR$_{ns}$	61.1	60.5	17.78	92.73
		CLEAN		
CORRECT	95.75	83.74	94.98	98.28
HR$_s$	99.78	100	100	97.84
HR$_{ns}$	91.7	67.49	89.97	100
		OVERALL		
CORRECT	87.37	81.92	76.84	96.08
HR$_s$	98.32	99.5	99.63	96.78
HR$_{ns}$	76.4	63.99	53.88	96.37

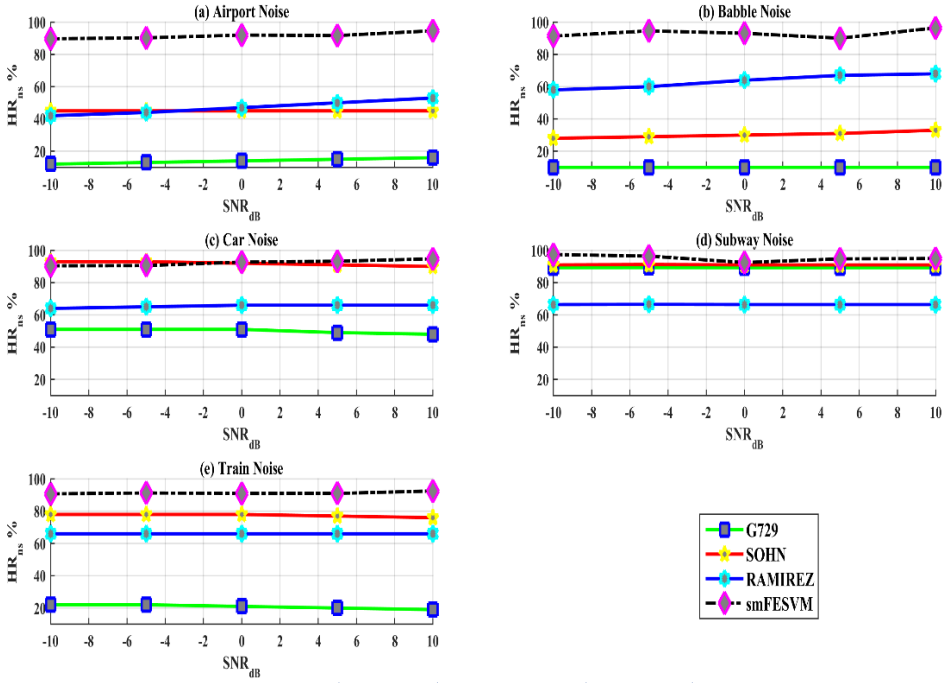

Fig. 6. HRns comparisons for G.729 (ITU, 1995), VAD-SOHN (Sohn, 1999), VAD-RAMIREZ (Ramírez et al., 2004) and *sm*FESVM based VAD for SNR values ranging between -10 and 10 dB (a) Airport (b) Babble (c) Car (d) Subway and (e) Train noises.

4. CONCLUSIONS

In this paper, the *sm*FESVM based VAD is presented. The significance of the feature ***conFETV′*** is discussed experimentally under various non-stationary noises at different SNR levels. The efficacy of the ***conFETV′*** feature is compared against the smoothing features considered and the feature vector formed was also analysed, which proved to be efficient in distinguishing HRs and HRns. The performance of the classifier is analyzed by 10-fold cross validation scheme. The results show that the proposed *sm*FESVM based VAD outperforms the other VAD algorithms considered based on performance metrics. Similarly, for babble noises and for other non-stationary noises at lower SNRs around -5 dB and -10 dB, the proposed algorithm proves its robustness under noisy conditions.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

Aneeja, G., & Yegnanarayana, B. (2015). Single Frequency Filtering Approach for Discriminating Speech and Nonspeech. *IEEE Transactions on Audio, Speech and Language Processing*, *23*(4), 705–717. https://doi.org/10.1109/TASLP.2015.2404035

Beritelli, F., Casale, S., Ruggeri, G., & Serrano, S. (2002). Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors. *IEEE Signal Processing Letters*, *9*(3), 85–88. https://doi.org/10.1109/97.995824

Chan, T. F., Osher, S., & Shen, J. (2001). The digital TV filter and nonlinear denoising. *IEEE Transactions on Image Processing*, *10*(2), 231–241. https://doi.org/10.1109/83.902288

Chen, W., Wang, Z., Xie, H., & Yu, W. (2007). Characterization of surface EMG signal based on fuzzy entropy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *15*(2), 266–272. https://doi.org/10.1109/TNSRE.2007.897025

Chen, W., Zhuang, J., Yu, W., & Wang, Z. (2009). Measuring complexity using FuzzyEn, ApEn, and SampEn. *Medical Engineering and Physics*, *31*(1), 61–68. https://doi.org/10.1016/j.medengphy.2008.04.005

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Davis, A., Nordholm, S., & Togneri, R. (2006). Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. *Audio, Speech, and Language Processing, IEEE Transactions on*, *14*(2), 412–424. https://doi.org/10.1109/TSA.2005.855842

Freeman, D. K., Cosier, G., Southcott, C. B., & Boyd, I. (1991). The voice activity detector for the Pan-European digital cellular mobile telephone service. In *International Conference on Acoustics, Speech, and Signal Processing* (pp. 369–372). IEEE. https://doi.org/10.1109/ICASSP.1989.266442

Ganapathiraju, A., Hamaker, J., & Picone, J. (2004). Applications of Support Vector Machines to Speech Recognition. *IEEE Transactions on Signal Processing*, *52*(8), 2348–2355. https://doi.org/10.1109/TSP.2004.831018

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. Retrieved from https://catalog.ldc.upenn.edu/LDC93S1

Ghosh, P. K., Tsiartas, A., & Narayanan, S. (2011). Robust Voice Activity Detection Using Long-Term Signal Variability. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(3), 600–613. https://doi.org/10.1109/TASL.2010.2052803

Guo, G., & Li, S. Z. (2003). Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Network*, *14*(1), 209–215. https://doi.org/10.1109/TNN.2002.806626

Hariharan, M., Fook, C. Y., Sindhu, R., Adom, A. H., & Yaacob, S. (2013). Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy. *Digital Signal Processing*, *23*(3), 952–959. https://doi.org/10.1016/j.dsp.2012.12.003

Hernández-Mena, C. D., Meza-Ruiz, I. V., & Herrera-Camacho, J. A. (2017). Automatic speech recognizers for Mexican Spanish and its open resources. *Journal of Applied Research and Technology*, *15*(3), 259–270. https://doi.org/10.1016/j.jart.2017.02.001

Hirsch, H. G., & Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *ICSLP 2000 (6th International Conference on Spokel Language Processing)*, *2000* (October),16–19.Retrieved from http://dblp.unitrier.de/db/conf/interspeech/interspeech2000.html#PearceH00

Hsieh, C. H., Feng, T. Y., & Huang, P. C. (2009). Energy-based VAD with grey magnitude spectral subtraction. *Speech Communication*, *51*(9), 810–819. https://doi.org/10.1016/j.specom.2008.08.005

Itoh, K., & Mizushima, M. (1997). Environmental noise reduction based on speech/non-speech identification for hearing aids. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*(Vol.1, pp.419–422). IEEE Comput.Soc.Press. https://doi.org/10.1109/ICASSP.1997.599662

ITU-T Study Group. (1995). Coding of Speech at 8 kbits/s Using Conjugate-structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). *International Telecommunica-tion Union Telecommunication Standardization Sector, Draft Recommendation, Version*, *6*.

Karray, L., & Martin, A. (2003). Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication*, *40*(3), 261–276. https://doi.org/10.1016/S0167-6393(02)00066-3

Khoa, P. C. (2012). Noise robust voice activity detection. *Nanyang Technological University*, *24*.

Kotnik, B., Kacic, Z., & Horvat, B. (2001). A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction. *Proc. 7th Euro. Conf. on Spee. Commn and Tech*, (977126),197–200. Retrieved from http://www.iscaspeech.org/archive/eurospeech_2001/e01_0197.html

LEE, Y.-C. (2006). Statistical Model-Based VAD Algorithm with Wavelet Transform. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E89–A*(6), 1594–1600. https://doi.org/10.1093/ietfec/e89-a.6.1594

Marzinzik, M., & Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, *10*(2), 109–118. https://doi.org/10.1109/89.985548

Nazir, M., Majid-Mirza, A., & Ali-Khan, S. (2014). PSO-GA Based Optimized Feature Selection Using Facial and Clothing Information for Gender Classification. *Journal of Applied Research and Technology*, *12*(1), 145–152. https://doi.org/10.1016/S1665-6423(14)71614-1

Nemer, E., Goubran, R., & Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*,*9*(3),217–231.https://doi.org/10.1109/89.905996

Ramírez, J., Segura, J. C., Benítez, C., De la Torre, A., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, *42*(3–4), 271–287.https://doi.org/10.1016/j.specom.2003.10.002

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol*, *278*(6), H2039-2049. https://doi.org/10.1103/physreva.29.975

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, *60*(1–4), 259–268. https://doi.org/10.1016/0167-2789(92)90242-F

Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R., Venkatesha Prasad, R., & Gaurav, V. (2002). VAD techniques for real-time speech transmission on the Internet. In *5th IEEE International Conference on High Speed Networks and Multimedia Communication (Cat. No.02EX612)* (pp. 46–50). IEEE. https://doi.org/10.1109/HSNMC.2002.1032545

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Shabat, A. M., & Tapamo, J.-R. (2017). A comparative study of the use of local directional pattern for texture-based informal settlement classification. *Journal of Applied Research and Technology*, *15*(3), 250–258. https://doi.org/10.1016/j.jart.2016.12.009

Shi, W., Zou, Y., & Liu, Y. (2014). Long-term auto-correlation statistics based voice activity detection for strong noisy speech. In *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)* (pp. 100–104). IEEE. https://doi.org/10.1109/ChinaSIP.2014.6889210

Sohn, J. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*(1), 1–3. https://doi.org/10.1109/97.736233

Wu, B. F., & Wang, K. C. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 762–774. https://doi.org/10.1109/TSA.2005.851909

Zhang, L., Gao, Y. C., Bian, Z. Z., & Lu, C. (2005). Voice activity detection algorithm improvement in adaptive multi-rate speech coding of 3GPP. *Proceedings - 2005 International Conference on Wireless Communications, Networking and Mobile Computing, WCNM 2005*, *2*(1), 1257–1260. https://doi.org/10.1109/WCNM.2005.1544283