International Journal of Engineering

Journal Homepage: www.ije.ir

A Voice Activity Detection Algorithm Using Sparse Non-negative Matrix Factorization-based Model Learning in Spectro-Temporal Domain

S. Mavaddati*

Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran

PAPER INFO

Paper history:
Received 02 January 2023
Received in revised form 13 May 2023
Accepted 21 May 2023

Keywords:
Voice Activity Detector
Spectro-temporal Domain
Sparse Structured Principal Component Analysis
Sparse Non-negative Matrix Factorization

ABSTRACT

Voice activity detectors are presented to extract silence/speech segments of the speech signal to eliminate different background noise signals. A novel voice activity detector is proposed in this paper using spectro-temporal features extracted from the auditory model of the speech signal. After extracting the scale, rate, and frequency features from this feature space, a sparse structured principal component analysis algorithm is used to consider the basic components of these features and reduce the dimension of learning data. Then these feature vectors are employed to learn the models by the sparse non-negative matrix factorization algorithm. The model learning procedure is performed to represent each feature vector with a proper sparse rate based on the selected atoms. Voice activity detection of the input frames is performed by computing the energy of the sparse representation for each input frame over the composite model. If the calculated energy exceeds a specified threshold, it indicates that the input frame has a structure similar to the atoms of the learned models and concludes that the observed frame has voice content. The results of the proposed detector were compared with other baseline methods and classifiers in this processing field. These results in the presence of stationary, non-stationary and periodic noises were investigated and they are shown that the proposed method based on model learning with spectro-temporal features can correctly detect the silence/speech activities.

***doi**: 10.5829/ije.2023.36.08b.08*

1. INTRODUCTION

One of the research fields in the speech signal processing is detection of silence/speech areas of the speech signal performed by a voice activity detector (VAD). The VAD block has an important role to eliminate the background noise from the speech signals. So far, different feature domains have been used to determine voice activities since the performance of VAD is closely related to the type of these extracted features. In these methods, an attempt is made to separate the speech frames from the silent sections of the speech signal. The energy of speech signal frames and the calculation of the zero-crossing rate (ZCR) are the most advanced features in this processing area [1]. Since various detectors have been introduced in many fields, this paper only deals with the methods presented based on the model learning technique. Ahmadi, and Joneidi [2] proposed a VAD algorithm based on the sparse representation technique using an orthogonal matching pursuit algorithm (OMP) followed by the K-singular value decomposition (K-SVD) dictionary learning method. The detection criterion of voice activity was based on the energy in the sparse representation of the input frame over the learned voice dictionary. You et al. [3] proposed a VAD algorithm based on the sparse representation technique using the Bergman iteration method and online dictionary learning. In this algorithm, the sparse power spectrum criterion was defined to calculate two types of features and decide on the label of the input frames. This criterion was achieved by averaging over the different signal segments that include the short segment average spectrum and long segment average spectrum. The labels of the different parts of the input frame are determined by calculating the energy in these frames. You et al. [4] optimized algorithm for learning speech and noise dictionaries. The

*Corresponding Author Email: s.mavaddati@umz.ac.ir (S. Mavaddati)

Please cite this article as: S. Mavaddati, A Voice Activity Detection Algorithm Using Sparse Non-negative Matrix **Factorization**-based Model Learning in Spectro-Temporal Domain, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1478-1488

goal of this optimization procedure was to reduce the coherence value between the learned dictionaries to obtain a robust VAD algorithm in the different noise conditions. The features used in this method were the modified versions of the features presented by You et al. [3] and include the long-time average energy and the long-time dynamic threshold. Also, Teng and Jia [5] designed a VAD algorithm using a non-negative sparse coding method with a noise reduction procedure. In this method, the input noise signal is first represented in the combined dictionary which contains the atoms associated with the speech and noise signals. The coefficients related to voice segments are then used as the desirable features in the conditional random field (CRF) method to model the correlation between the feature sequences and detect the speech and noise labels for each input frame. Mavaddaty et al. [6] used the spectro features of speech signal spectrograms to learn the models using the concepts of sparse representation and the K-SVD algorithm. In this work, two supervised and semi-supervised methods were presented to eliminate the background noise from the speech signal. The main part of each method was the presented voice activity detector in the wavelet packet transform domain.

The purpose of this paper is to increase the detection accuracy as much as possible based on the proposed model-based method by applying the spectro-temporal features. In this paper, scale, rate, and frequency characteristics extracted from the auditory model of the speech signal were used to learn models that show the structure of active parts of speech signals. In the following, the dimension of the mentioned features was reduced by the parse structured principal component analysis (SSPCA) algorithm and then the sparse non-negative matrix factorization (SNMF) algorithm is employed to learn the dimensionless feature sets.

In the second part of this paper, the auditory model and its extractive features are introduced. Section 3 introduces the SNMF model and the proposed VAD algorithm. In section 4, the performance of the proposed method is evaluated and the paper is concluded in section 5.

2. Spectro-Temporal Representation Using Auditory Model

As stated, the recognition process to detect speech areas of the speech signal and separate these frames from the silence frames has a great importance in many speech processing applications. In this paper, the spectro-temporal features are used to identify the speech segments of the speech signals that can be described using the auditory model. In this model, the auditory spectrum related to each speech is calculated. Then, the spectro-temporal features are extracted using this spectrogram and the auditory cortex model [7]. The features of the auditory cortex model have four dimensions: scale Ω, speech rate ω, frequency f, and time or frame number t. The auditory part of the cortical model is implemented by a time-frequency filter bank. Each filter can operate at different rates and scales to simulate the cochlear of the human ear and the first layer of the auditory brainstem. This procedure of filtering at different rates and scales is performed linearly in the spectro-temporal space by the wavelet transform function or the two-dimensional Gabor filter [7-9].

The block diagram of the auditory cortex model is shown in Figure 1. Initially, the acoustic signal enters the filter bank that consist of 128 uniformly distributed bandpass filters along the frequency-logarithmic axis that models the performance of the outer membrane of the human ear. The output of this filter bank with a time-frequency structure passes from three steps: a derivative high pass filter, a nonlinear compressor, and a low pass filter to simulate the inner portion of the human ear. In the following, the auditory spectrogram of the speech signal is obtained by the first-order derivative, half-wave rectifier, and integrator. Then, the spectro-temporal content of the auditory spectrum is achieved by a filter bank consisting of a two-dimensional Gabor filter. Then, a four-dimensional speech cortical signal including Ω scale in cycles/octave, speech velocity or rate ω in Hertz, frequency f, and the frame number of the input speech signal t is yielded.

3. THE PROPOSED VOICE ACTIVITY DETECTOR

In this section, the proposed VAD algorithm is presented using the extracted spectro-temporal features and SNMF-based model learning. The proposed method employs model learning technique to represent the structure of the input frame. Model learning in this paper is performed by the sparse non-negative matrix factorization algorithm, which is the non-negative matrix factorization (NMF) procedure that has been added to the nonlinearity constraint.

The combination of the sparse and NMF coding algorithms results in a model learning method called SNMF [10-12]. This technique results in a sparser representation than the NMF algorithm to apply the sparse constraints. In the SNMF algorithm, which is more robust than the NMF algorithm, the generalized Kolbeck-Leibler divergence method used to determine lower approximation error in the data representation. In the sparse encoding technique, each input signal frame can be represented as a linear combination of the dictionary atoms. In this procedure, it is determined which set of atoms and coefficients represent the data frame with the least approximation error. These sparse coefficients for all input signal frames constitute the H sparse coefficient matrix, which is one of the outputs of the SNMF algorithm. Many coefficients in the sparse matrix H have

a zero value and indicate that each data frame can be represented only by a limited number of dictionary atoms. The sparsity or cardinality parameter determines the number of atoms in each representation procedure. The data matrix containing signal frames S can be modeled as follows by sparse coding:

$$S = WH \tag{1}$$

where $W \in R^{N\times L}$ is a learned model or dictionary in which the columns are atoms. The W dictionary matrix contains L columns or atoms $\{W_l\}_{l=1}^{L}$ with the unit norm $\|W_{(:,l)}\|_2 = 1, \forall l = 1,\ldots,L$. Also, the K-sparse coefficient matrix H with L≫K includes the representation coefficients related to the input data matrix [13]. The sparse representation problem that consists of the approximation error and sparse constraint parts is formulated as follows [13]:

$$H^* = \underset{H}{\operatorname{argmin}} \|S - WH\|_2^2 \quad \text{s.t.} \quad \|H\|_0 \le C \tag{2}$$

where C represents the sparse rate or the number of non-zero coefficients in each row of the sparse matrix H. This parameter must be set correctly to avoid massive coding. If the high value is selected for this parameter, the large numbers of atoms participate in the representation of the input data frame that is improper. On the other hand, if the low value is selected for this parameter, the atoms are not enough to represent the data structure, and then the approximation error increases. The NMF algorithm performs a linear analysis on the observed data matrix $S \in R^{N\times M}$ and factorizes the input data matrix into two dictionary matrix $W \in R^{N\times L}$ and the coefficient matrix $H \in R^{L\times M}$ as $I = WH$ with non-negative values, which L is smaller than M and N [13]. These matrices are employed to solve the following optimization problem:

$$\min F(W,H) = \sum_{i,j}(S_{i,j}\log(S_{i,j}[WH]_{i,j}) - S_{i,j} + [WH]_{i,j}) \quad s.t. \quad W,H \ge 0, \sum_l W_{(:,l)} = 1 \tag{3}$$

The optimization of this cost function is based on the generalized Kullback-Leibler divergence method. However, solving this problem with other cost functions yields different versions of the NMF algorithm.

As stated, the SNMF algorithm will obtain a sparser representation to consider a specified constraint than the NMF algorithm [11-13]. The generalized Kullback-Leibler divergence algorithm is then used to determine the approximation error in the analysis of non-negative coefficients, which results in the following optimization problem:

$$\min F(W,H) = \sum_{i,j}(S_{i,j}\log(S_{i,j}[WH]_{i,j}) - S_{i,j} + [WH]_{i,j}) + \alpha\sum_{k,j} h_{k,j} \quad s.t. \quad W,H \ge 0, \sum_l W_{(:,l)} = 1 \tag{4}$$

The α parameter determines the weight coefficient of the sparsity part. The update of atoms in the W dictionary matrix is as follows:

$$h^*_{k,j} = (h_{k,j}\sum_i I_{i,j}\, w_{i,k} \Big/ \sum_l w_{i,l}h_{l,j}) / (1+\alpha),$$
$$w^*_{i,k} = (w_{i,k}\sum_j I_{i,j}\, h_{k,j} \Big/ \sum_l w_{i,l}h_{l,j}) \Big/ \sum_j h_{k,j}\ ,$$
$$w^{**}_{i,k} = (w^*_{i,k} \Big/ \sum_i w^*_{i,k}) \tag{5}$$

The NMF algorithm is obtained when the α parameter is omitted in Equation (5) [11]. Then, the dimensionality of the data matrix is reduced by the SSPCA algorithm to learn comprehensive models for the representation of the input data structure. The principal component analysis algorithm (PCA) is a commonly used statistical method to reduce data dimension and is used to convert the input data sets into a new set of the independent variables that include the maximum changes in the original data [13]. This algorithm presented to develop the SSPCA method which is used to estimate the basic components by applying a sparsity constraint [14]. The benefits of using this method include reducing computation time, extracting the components with more variance, and obtaining appropriate values for important variables of each problem. Further, by generalizing this algorithm, the SSPCA algorithm is obtained, which can extract the data with more variance using the sparsity and some structural constraints [15]. The non-convex form of the SSPCA algorithm is presented by Jenatton et al. [16] to solve the problem of structured sparse dictionary learning. The SSPCA is a robust algorithm to solve the occlusion problem using the block-coordinate descent algorithm for better data analysis.

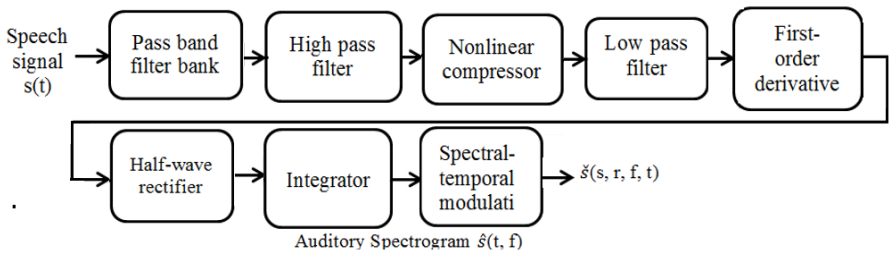

Figure 1. Block diagram of the cortical model of the speech signal

The block diagram of the proposed method to determine the labels of input speech frames using spectro-temporal properties is shown in Figure 2.

4. DETAILS OF SIMULATION

In this paper, the TIMIT dataset is used to determine the efficiency of the proposed method. This comprehensive speech dataset contains a large number of speakers and expressions that is suitable to consider the performance of a VAD algorithm [17]. The sampling rate of speech signals is set to 16kHz. The train and test scenarios contain 200 and 100 spoken expressions, respectively. In the training step, the phrases are uttered by 10 female and 10 male speakers. In the test step, the phrases uttered by 3 male and 3 female speakers were employed in the speaker-independent test. The data frame length is equal to 20 ms and the frame overlap is 50%. The parameter settings in the learning procedure are the same for all spoken data sources in the train and test scenarios.

4. 1. Simulation Results In the proposed method, the model learning procedure using the SNMF algorithm was used to identify silence/speech speech frames. The sparsity rate for the dictionary and the coefficients matrices in the SNMF algorithm are set to 0.9 and 0.7, respectively. These parameter values are achieved based on the experimental simulations to result in a lower approximation error. Also, the number of iterations and the sparsity parameter in the SSPCA as employed in dimension reduction are 250 and 0.6, which leads to stability in the solving procedure. The performance evaluation of the algorithms is determined by the classification accuracy rate, which is calculated by the percentage of voice and silence frames that have the correct labels for the entire test data. In the first step of the proposed algorithm, 100 speech signals with silence/speech labels selected from the TIMIT dataset are used to learn the model of scale, rate, and frequency features extracted from the auditory cortex model. The auditory model of these signals is computed and then applied to the model learning after employing the SSPCA dimension reduction algorithm. Finally, the the learned models that represent the structural features of the silence/speech segments are considered in the representation of the test input signal. The sparsity parameter in this algorithm means that each input data frame can only be represented by a linear combination of a small number of learned atoms. This parameter value is determined by the cardinality rate. Input data classification in this paper is not performed by the usual classifiers such as neural networks, support vector machine or decision trees, but it is suggested to design and use a model-based classifier based on the calculated energy of the extracted features from the sparse coefficients matrix. In the proposed detection procedure, the input signal is sparsely represented by the SNMF algorithm on the combinational dictionary D= [Ds Dr Df]. This composite model D consists of the learning models related to the scale, rate, and frequency features with the same parameter values in the training step. Then, the energy of the sparse representation coefficients obtained on each dictionary is computed as:

$$H_S^*, H_R^*, H_F^* = \mathrm{SNMF}\left(\boldsymbol{Y},\ [D_s\ D_r\ D_f]\right) \tag{6}$$

$$E_s = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{s,l}^{*2}\ ,\ E_r = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{r,l}^{*2},\ E_f = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{f,l}^{*2} \tag{7}$$

where L represents the length of the frame and E_s, E_r, and E_f are the energy of the sparse representing related to scale, rate, and frequency features. $\boldsymbol{Y}$ is the observation matrix. Also, H_S^*, H_R^* and H_F^* are sparse coefficient matrices related to scale, rate and frequency features of the speech signal. The sum of energies is calculated and if this energy is more than half the energy of the input frame then it can result that the input frame contains the voice structure. If the difference between the calculated energy in the sparse coding procedure over the speech model and the energy of the input frame is less than a specified value of ε_1=0.04, then the average energy of the SNMF coding coefficients for one frame before and one frame after the input frame is calculated as the short-term energy. The value ε_1 has been obtained experimentally in various simulations. If the short-term energy is higher than half the energy of the input frame, the input frame has a speech label otherwise it will have a silence label.

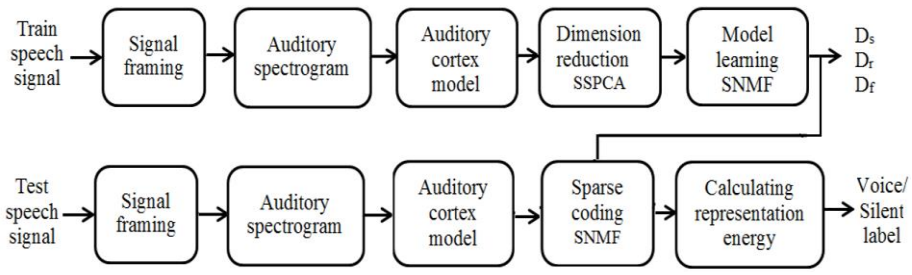

Figure 2. Block diagram of the proposed method to determine silence/speech speech frames using spectro-temporal properties

The time-frequency energy plot of the learned atoms based on the proposed SNMF-based VAD algorithm using the elliptical plots presented by Jafari and Plumbley [18] which is shown in Figure 3. This procedure determines how much of the time-frequency energy of frames is sparsely represented by the learned atoms. These plots show that the learned atoms according to the proposed method have been able to cover the entire time-frequency space of the considered speech signals. The elliptical plot of the proposed method based on the frequency features is concentrated in the center of the time axis and it does not include the entire frequency content at different times. The proposed method consists of a wide time-frequency range caused by a proper matching with the content of observation signal and the dictionary atoms.

The spectrogram plots of the atoms learned by the proposed method, the frequency features and the spectro-temporal features are shown in Figure 4. These plots show that the learned atoms according to the proposed method have the highest energy coverage in the time-frequency space and can precisely display the structure of the speech and silence frames.The parameters setting procedure was done according to the experimental simulations to have a proper decision on the input frame label. Since the input data frame with voice content has more energy in the sparse representation on the related dictionary so the energy criterion of the resulting sparse coefficients is used to determine the appropriate label. As a result, there is no need to use other classifiers, and the labeling procedure of the input frame can only be estimated using the SNMF algorithm. The results of the proposed method to detect the silence/speech frames are reported in Tables 1 and 2 for the speaker-independent and speaker-dependent detection scenarios, respectively. It is noteworthy that this paper has tried to evaluate the performance of the proposed VAD algorithm with the methods presented in the field of sparse representation technique. The results show that the proposed method has the ability to correctly identify the input area by applying the comprehensive learning models based on the structural content of the input frames. These results are slightly higher in the speaker-dependent scenario than in the speaker-independent scenario, which may be due to the overlap between the train and test data speakers.

The results of the proposed algorithm were compared with the other voice activity detection methods introduced in this processing field. These methods include the algorithm presented by Sharma1 and Rajpoot [19] that utilizes a zero-crossing rate and clustering procedure and also the VAD method which uses a clustering method based on the Gaussian mixture model. Mavaddaty et al. [6] presented a VAD algorithm based on the energy of the sparse coefficient matrices extracted from the wavelet packet transform features of speech and noise signals.

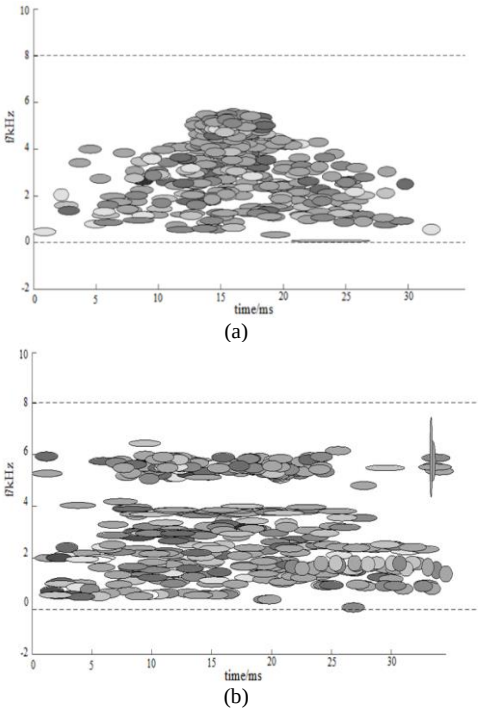

Figure 3. The elliptical plots of the time-frequency energy of the atoms learned by: a) the SMF-based VAD algorithm based on frequency features. b) the proposed method based on spectro-temporal features

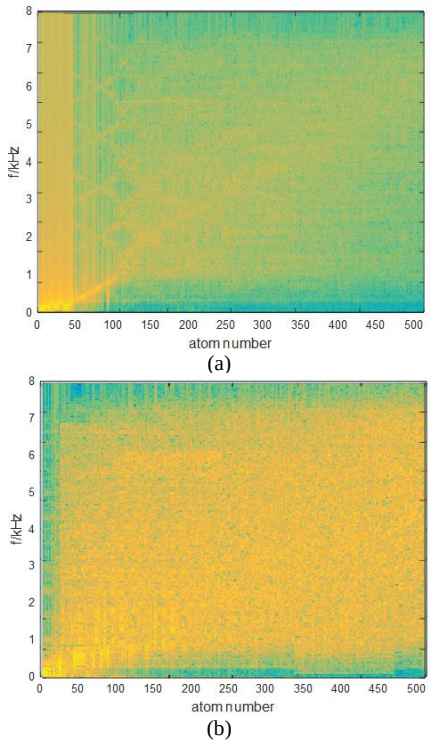

Figure 4. The spectrogram plot of the atoms learned by: a) the SMF-based VAD algorithm based on frequency features. b) the proposed method based on the spectro-temporal features

TABLE 1. The average accuracy of the proposed VAD algorithm in a speaker-independent scenario

Speaker	#Sentences	Voice	Silent	Average accuracy
Woman	25	97.43	98.25	97.84
Man	25	97.62	98.11	97.86

TABLE 2. The average accuracy of the proposed VAD algorithm in a speaker-dependent scenario

Speaker	#Sentences	Voice	Silent	Average accuracy
Woman	25	98.21	98.36	98.28
Man	25	97.89	98.49	98.19

The sparse coding procedure was based on a combination of orthogonal matching pursuit algorithm (OMP) and coherence criterion.

A VAD algorithm with a combination of convolutional recurrent neural network and a recurrent neural network was proposed by Wang and Zhang [20]. Also, a speech enhancement module was designed to improve the performance of VAD system in low signal-noise ratio conditions. Jordán et al. [21] introduced a VAD system to identify correctly the speech frames based on recurrent neural networks. The model defined in this paper was learned using bidirectional long short-term memory.

A comparison is also made with the method presented by Ahmadi and Joneidi [2], which is based on a sparse representation using the orthogonal matching pursuit (OMP) algorithm and K-SVD dictionary learning algorithm. As mentioned before the presented VAD algorithm employed SNMF learning method with a sparse-based statistical structure as a model learning method that has been widely used in signal processing in recent [22, 23].

These results are presented in Tables 3 and 4. The results show that the proposed method correctly identifies the voice and silent regions of the input speech signal. This success and superiority over other methods can be due to the use of appropriate learned models and the dimension reduction algorithm to eliminate the outlier data during the training step. In these simulations, the results of the speaker-dependent scenario are better than the speaker-independent test, which can be due to the similarity between the speakers in the train and test steps. The results show that employing spectro-temporal features and speech signal processing through the auditory model is a desirable approach to identify the speech frames. The combination of these two techniques has many applications as a pre-processing step in speech signal analysis. The first two rows in Tables 3 and 4 are the same since the methods proposed by Sharma1 and Rajpoot [19], they did not employ the learning-based technique and the detection procedure for them occurs in one step, not in the different scenarios.

To investigate more the performance of the proposed method, the ROC curve obtained from the results of the proposed method and other comparable methods in the speaker-independent and speaker-dependent scenarios are shown in Figures 5 and 6, respectively, which emphasize the capability of the proposed method to achieve high accuracy in detection procedure.

4. 2. Simulation Results in The Presence of Different Noise Signals The quality of the speech signal can be significantly reduced in the presence of environmental noise signals and lead to the malfunction of hearing aids, automatic speech recognition systems, cell phones, etc. In this paper, a single-channel speech

TABLE 3. The average accuracy percentage of the proposed algorithm and the compared methods to detect the silence/speech sections of the speech signal in the speaker-independent scenario

	#Sentences	Voice	Silent	Average accuracy
Zero crossing-based method [19]	50	95.56	96.67	96.11
GMM-based method	50	97.41	97.78	97.59
Sparse representation-based method [2]	50	97.23	97.89	97.56
sparse dictionary learning-based method [6]	50	97.92	97.97	97.94
NN-based method [20]	50	97.72	97.91	97.81
CRNN-based method [21]	50	97.86	97.90	97.88
Proposed method	50	**98.05**	**98.19**	**98.12**

TABLE 4. The average accuracy percentage of the proposed algorithm and the compared methods to detect the silence/speech areas of the speech signal in the speaker-dependent scenario

	#Sentences	Voice	Silent	Average accuracy
Zero crossing-based method [19]	50	95.56	96.67	96.11
GMM-based method	50	97.41	97.78	97.59
Sparse representation-based method [2]	50	97.69	97.93	97.81
sparse dictionary learning-based method [6]	50	97.98	98.01	97.99
NN-based method [20]	50	97.81	97.99	97.90
CRNN-based method [22]	50	97.93	98.09	98.01
Proposed method	50	**98.14**	**98.24**	**98.19**

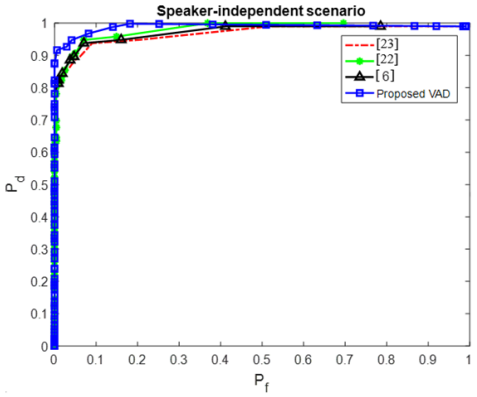

Figure 5. The ROC curve obtained from the results of the proposed method and other compared methods in the speaker-independent scenario

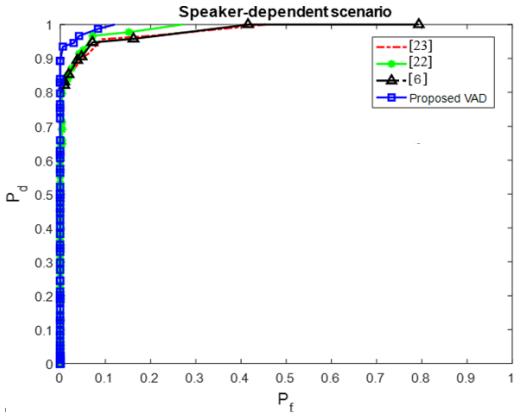

Figure 6. The ROC curve obtained from the results of the proposed method and other compared methods in the speaker-dependent scenario

processing corrupted by additive noise is considered. When the speech signal is exposed to non-stationary noise signals, the performance of the VAD algorithm decreases. This is especially for speech-like noise signals, that have fundamental overlap between the components in the spectro-temporal domain. Although the evaluation of the VAD algorithm is usually not performed in the presence of noise signals, in this paper, the performance of the proposed VAD algorithm in different noise conditions is investigated. In none of the references in which the proposed method has been compared with them, such as [2, 19-23], the performance evaluation in the presence of noise has not been done, so the results of this method have not been reviewed in the noise conditions and only the proposed method has been evaluated.

In this paper, a variety of noise signals consisting of white and babble noises from Noisex92 [24] car and train noises from Aurora2 [25] as well as piano noise from the piano society website[1] have been considered to have a proper investigation about the performance of the proposed method.

The block diagram of the proposed VAD algorithm to determine the labels of the input frames in the presence of the mentioned noise signals is shown in Figure 7. The learning procedures for speech and different noise signals were carried out with the same parameters in the SNMF coding algorithm and the dimension reduction technique.

In recent years, the use of sparse representation techniques for voice activity detector algorithms in a noisy condition has increased. An ideal VAD is used to acquire the data frames needed to learn the noise signal dictionary as reported by Sigg et al. [26]. The data frames obtained by the non-speech frames of the noisy signal are not usually enough to learn a dictionary with low approximation error. The noise dictionary learning algorithm in this approach is performed in the speech enhancement step and leads to a significant increase in the computation time. Also, Sigg et al. [27] presented a generative coherence-based dictionary learning method using the pure noise data to train noise dictionary models. The offline learning process was performed with enough noise signals. In this paper, the advantages of the SNMF technique were utilized to learn the dictionaries for scale, rate, and frequency features. The model learning procedure for the noise signals is done without any problems since adequate noise data is available. This learning process for speech and noise signals is carried

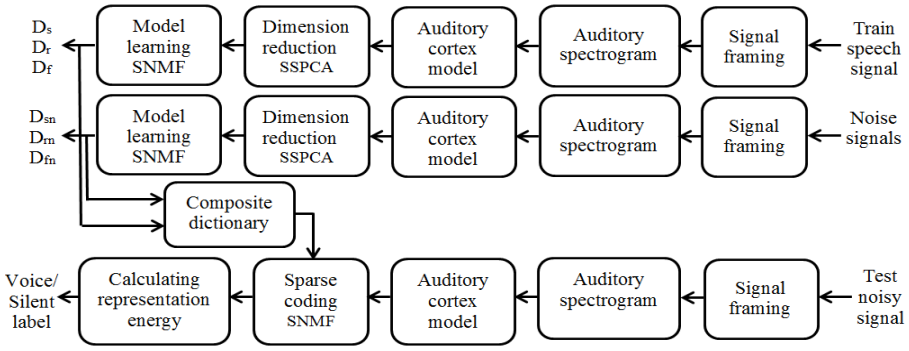

Figure 7. Block diagram of the proposed VAD based on spectro-temporal properties in the presence of noise signals

[1] http://pianosociety.com

out precisely in the same way. The sparse representation in the presence of noise is carried out over a composite dictionary that includes speech and noise models as:

$$H_S^*, H_R^*, H_F^*, H_{Sn}^*, H_{Rn}^*, H_{Fn}^* = \text{SNMF}\left(\boldsymbol{Y},\ \left[D_s\ D_r\ D_f\ D_{sn}\ D_{rn}\ D_{fn}\right]\right) \quad (8)$$

$$E_s = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{s,l}^{*2}\ ,\ E_r = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{r,l}^{*2},\quad E_f = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{f,l}^{*2}$$
$$E_{sn} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{sn,l}^{*2}\ ,\ E_{rn} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{rn,l}^{*2},\quad E_{fn} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{H}_{fn,l}^{*2} \quad (9)$$

where E_{sn}, E_{rn}, and E_{fn} are the energy of the sparse representation corresponding to scale, rate, and frequency features of each noise signal. Also, H_{Sn}^*, H_{Rn}^* and H_{Fn}^* are sparse coefficient matrices related to scale, rate, and frequency features of each noise data class.

This procedure in the train and test steps should be carried out for each noise signal. The learning and dimension reduction procedures for all kinds of noise signals are done the same as a speech signal. According to Equation (8), the input noisy frame is sparsely coded over the composite dictionary $\left[D_s\ D_r\ D_f\ D_{sn}\ D_{rn}\ D_{fn}\right]$. In this test step, the sum of the energies calculated from the sparse coefficient matrices for speech and noise signals is considered. The total energy calculated based on speech and noise model determines the label of the input noisy frames. If this calculated energy over the speech signal model is greater then the calculated energy over the noise model, the input frame is detected as speech frame, otherwise, a noise label is assigned to this frame. Also, if the energy difference calculated on the speech and noise models is less than a certain limit of ε_2=0.08, then the total energy of the sparse coding coefficients for one frame before and one frame after the input frame is calculated over the speech and noise models to obtain the short-term energy of this representation. If the average of these calculated energies on the speech model is higher than the noise model, the speech label is assigned to the input frame, otherwise the noise label.

The average results of the proposed method to assign the proper labels in a speaker-independent scenario in the presence of various noise signals with 10dB SNR are shown in Table 5. Also, these results in a speaker-dependent scenario are reported in Table 6. For more evaluation of the performance of the proposed VAD in different conditions, the average results of the proposed VAD in the speaker-independent and speaker-dependent scenarios in the presence of various noise signals with 5dB SNR are shown in Tables 7 and 8, respectively.

From the reported values in Tables 5-8, it can be concluded that the accuracy of the proposed method decreases as the SNR value decreases. Also, the accuracy of labeling to silence/speech in the presence of noise signals with stationary content such as white noise is higher than other noise signals. The accuracy in the presence of periodic piano noise signal with harmonic structure is more accurate than in other conditions. In

TABLE 5. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-independent scenario and the presence of noise signals with SNR=10dB

	#Sentences	Voice	Silent	Average accuracy
Without noise	100	98.05	98.19	98.12
White noise	100	97.20	97.01	97.10
Car noise	100	94.98	95.47	95.22
Piano noise	100	97.02	97.21	97.11
Babble noise	100	94.20	94.22	94.21
Train noise	100	94.59	95.23	94.91

TABLE 6. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-dependent scenario and the presence of noise signals with SNR=10dB

	#Sentences	Voice	Silent	Average accuracy
Without noise	100	98.14	98.24	98.19
White noise	100	97.51	97.18	97.34
Car noise	100	95.23	95.76	95.49
Piano noise	100	97.21	97.33	97.27
Babble noise	100	94.28	94.36	94.32
Train noise	100	94.76	95.32	95.04

TABLE 7. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-independent scenario and the presence of noise signals with input SNR=5dB

	#Sentences	Voice	Silent	Average accuracy
Without noise	100	98.05	98.19	98.12
White noise	100	94.11	94.26	94.18
Car noise	100	92.22	93.31	92.76
Piano noise	100	94.47	94.63	94.55
Babble noise	100	92.13	92.47	92.30
Train noise	100	92.65	93.06	92.85

TABLE 8. The average accuracy percentage of the proposed algorithm to detect the silence/speech frames in a speaker-dependent scenario and the presence of noise signals with SNR=5dB

	#Sentences	Voice	Silent	Average accuracy
Without noise	100	98.14	98.24	98.19
White noise	100	94.73	94.41	94.57
Car noise	100	92.51	93.49	93.00
Piano noise	100	94.70	94.88	94.79
Babble noise	100	92.53	92.66	92.59
Train noise	100	92.84	93.30	93.07

addition, the results of the proposed VAD algorithm has been considered in the presence of the car noise signal that has a stationary structure. But in the presence of babble noise, which is very similar to the speech signal, accuracy is greatly reduced. It should be noted that in the speaker-dependent scenario, the results are slightly higher than in the speaker-independent scenario in different situations because there is an overlap between the speakers in the train and test steps. Therefore, it can be concluded that the best results are obtained in the high SNR value, in the presence of white and piano noise signals, and the speaker-dependent scenario. Also, the performance of labeling in the case of speech frames that consist of consonant letters such as fricatives that have a similar structure to the noise signal may be decreased. The average accuracy values in the speaker-dependent and independent scenarios evaluated at two SNR values of 10dB and 5dB are represented in Figure 8. These results are obtained for a clean speech signal case and five stationary, non-stationary and periodic noises: white, car, train, babble, and piano signals. In general, it can be concluded that the reported results emphasize that the proposed VAD has an appropriate performance in different noisy conditions.

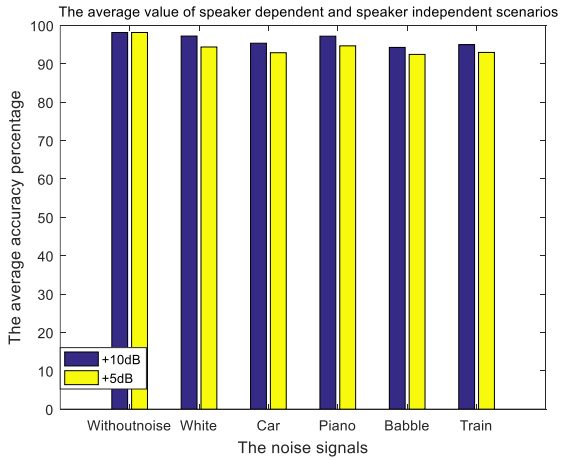

Figure 8. Performance comparison of the proposed method in terms of average accuracy in speaker-dependent and independent cases in 10dB and 5dB SNRs

5. CONCLUSION

Voice activity detection methods are very effective in the various fields of signal analysis and speech processing as a pre-processing block. In this paper, this detection procedure is performed in the space of spectro-temporal features. The features extracted from this space are used to learn comprehensive models of the input data structure. The dimension of these feature matrices is reduced by the SSPCA algorithm. Then the resulted data are used to learn models using the SNMF method which has a sparse-based statistical structure. In the following, by computing the energy derived from the representation of the input frame features on the composite model, the label of the input frame is identified. Also, these results have been examined for an extensive range of noise types including the stationary, non-stationary, and periodic noise signals in two SNR values of 5dB and 10dB. The simulation results in both speaker-independent and speaker-dependent scenarios indicate the superior performance of the proposed method compared to the other methods presented in this processing field.

6. CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7. REFERENCES

1. Park, J.-S., Yoon, J.-S., Seo, Y.-H. and Jang, G.-J., "Spectral energy based voice activity detection for real-time voice interface", ***Journal of Theoretical & Applied Information Technology***, Vol. 95, No. 17, (2017).
2. Ahmadi, P. and Joneidi, M., "A new method for voice activity detection based on sparse representation", in 2014 7th International Congress on Image and Signal Processing, IEEE. (2014), 878-882.
3. You, D., Han, J., Zheng, G. and Zheng, T., "Sparse power spectrum based robust voice activity detector", in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. (2012), 289-292.
4. You, D., Han, J., Zheng, G., Zheng, T. and Li, J., "Sparse representation with optimized learned dictionary for robust voice activity detection", ***Circuits, Systems, and Signal Processing***, Vol. 33, (2014), 2267-2291. doi: 10.1007/s00034-014-9748-y.
5. Teng, P. and Jia, Y., "Voice activity detection via noise reducing using non-negative sparse coding", ***IEEE Signal Processing Letters***, Vol. 20, No. 5, (2013), 475-478. doi: 10.1109/LSP.2013.2252615.

6. Mavaddaty, S., Ahadi, S.M. and Seyedin, S., "Speech enhancement using sparse dictionary learning in wavelet packet transform domain", ***Computer Speech & Language***, Vol. 44, (2017), 22-47. doi: 10.1016/j.csl.2017.01.009.
7. Chi, T., Ru, P. and Shamma, S.A., "Multiresolution spectrotemporal analysis of complex sounds", ***The Journal of the Acoustical Society of America***, Vol. 118, No. 2, (2005), 887-906. doi: 10.1121/1.1945807.
8. Elhilali, M., Chi, T. and Shamma, S.A., "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility", ***Speech Communication***, Vol. 41, No. 2-3, (2003), 331-348. doi: 10.1016/S0167-6393(02)00134-6.
9. Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z. and Shamma, S.A., "Dynamics of precise spike timing in primary auditory cortex", ***Journal of Neuroscience***, Vol. 24, No. 5, (2004), 1159-1172. doi: 10.1523/JNEUROSCI.3825-03.2004.
10. Hoyer, P.O., "Non-negative matrix factorization with sparseness constraints", ***Journal of Machine Learning Research***, Vol. 5, No. 9, (2004). doi: 10.48550/arXiv.cs/0408058.
11. Ullah, R., Islam, M.S., Ye, Z. and Asif, M., "Semi-supervised transient noise suppression using omlsa and snmf algorithms", ***Applied Acoustics***, Vol. 170, (2020), 107533. doi: 10.1016/j.apacoust.2020.107533.
12. Ullah, R., Islam, M.S., Hossain, M.I., Wahab, F.E. and Ye, Z., "Single channel speech dereverberation and separation using rpca and snmf", ***Applied Acoustics***, Vol. 167, (2020), 107406. doi: 10.1016/j.apacoust.2020.107406.
13. Jolliffe, I.T., "Principal component analysis for special types of data, Springer, (2002).
14. Zou, H., Hastie, T. and Tibshirani, R., "Sparse principal component analysis", ***Journal of Computational and Graphical Statistics***, Vol. 15, No. 2, (2006), 265-286. doi: 10.1198/106186006X113430.
15. Jenatton, R., Obozinski, G. and Bach, F., "Structured sparse principal component analysis", in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings. Vol., No. Issue, (2010), 366-373.
16. Jenatton, R., Audibert, J.-Y. and Bach, F., "Structured variable selection with sparsity-inducing norms", ***The Journal of Machine Learning Research***, Vol. 12, (2011), 2777-2824. doi: 10.48550/arXiv.0904.3523.
17. Kapadia, S., Valtchev, V. and Young, S.J., "Mmi training for continuous phoneme recognition on the timit database", in 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE. Vol. 2, (1993), 491-494.
18. Jafari, M.G. and Plumbley, M.D., "Speech denoising based on a greedy adaptive dictionary algorithm", in 2009 17th European Signal Processing Conference, IEEE. (2009), 1423-1426.
19. Sharma, P. and Rajpoot, A.K., "Automatic identification of silence, unvoiced and voiced chunks in speech", ***Journal of Computer Science & Information Technology (CS & IT)***, Vol. 3, No. 5, (2013), 87-96. doi: 10.5121/csit.2013.3509.
20. Wang, G.-B. and Zhang, W.-Q., "An rnn and crnn based approach to robust voice activity detection", in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE. (2019), 1347-1350.
21. Jordán, P.G., Bailo, I.V., Giménez, A.O., Artiaga, A.M. and Solano, E.L., "Vivovad: A voice activity detection tool based on recurrent neural networks", ***Jornada de Jóvenes Investigadores del I3A***, Vol. 7, (2019). doi: 10.26754/jji-i3a.003524.
22. Mavaddati, S., "Voice-based age and gender recognition using training generative sparse model", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 31, No. 9, (2018), 1529-1535. doi: 10.5829/ije.2018.31.09c.08.
23. Sabzalian, B. and Abolghasemi, V., "Iterative weighted non-smooth non-negative matrix factorization for face recognition", ***International Journal of Engineering, Transactions A: Basics***, Vol. 31, No. 10, (2018), 1698-1707. doi: 10.5829/ije.2018.31.10a.12.
24. Varga, A. and Steeneken, H.J., "Assessment for automatic speech recognition: Ii. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems", ***Speech Communication***, Vol. 12, No. 3, (1993), 247-251. doi: 10.1016/0167-6393(93)90095-3.
25. Hirsch, H.-G. and Pearce, D., "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW). (2000).
26. Sigg, C.D., Dikk, T. and Buhmann, J.M., "Speech enhancement with sparse coding in learned dictionaries", in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE. (2010), 4758-4761.
27. Sigg, C.D., Dikk, T. and Buhmann, J.M., "Speech enhancement using generative dictionary learning", ***IEEE Transactions on Audio, Speech, and Language Processing***, Vol. 20, No. 6, (2012), 1698-1712. doi: 10.1109/TASL.2012.2187194.

Persian Abstract

چکیده

آشکارسازهای فعالیت صوتی برای استخراج بخش‌های سکوت/صوت سیگنال گفتار برای حذف سیگنال‌های مختلف نویز پس‌زمینه ارائه شده‌اند. در این مقاله یک آشکارساز فعالیت صوتی جدید با استفاده از ویژگی‌های طیفی-زمانی استخراج شده از مدل شنیداری سیگنال گفتار پیشنهاد شده است. پس از استخراج ویژگی‌های مقیاس، نرخ و فرکانس از این فضای ویژگی، از یک الگوریتم تجزیه و تحلیل مولفه‌های اساسی ساختارمند تُنُک برای در نظر گرفتن اجزای اصلی این ویژگی‌ها و کاهش ابعاد داده‌های یادگیری استفاده می‌شود. سپس این بردارهای ویژگی برای یادگیری مدل توسط الگوریتم فاکتورسازی ماتریس تُنُک غیرمنفی استفاده می‌شوند. روش یادگیری مدل برای نشان دادن هر بردار ویژگی با نرخ تُنُک مناسب براساس اتم‌های انتخاب شده انجام می‌گیرد. تشخیص فعالیت صوتی فریم‌های ورودی با محاسبه انرژی نمایش تُنُک برای هر فریم ورودی بر روی مدل ترکیبی انجام می‌شود. اگر انرژی محاسبه شده از یک آستانه مشخص فراتر رود، نشان می‌دهد که قاب ورودی ساختاری مشابه اتم‌های مدل آموخته شده دارد و نتیجه گیری می‌شود که قاب مشاهده شده دارای محتوای صوتی است. نتایج آشکارساز پیشنهادی با سایر روش‌ها و طبقه‌بندی‌کننده‌های پایه در این زمینه از پردازش سیگنال گفتار مقایسه می‌شود. این نتایج در حضور نویزهای ایستا، غیرایستا و متناوب بررسی شده و نشان داده می‌شود که روش پیشنهادی مبتنی بر یادگیری مدل با ویژگی‌های طیفی-زمانی می‌تواند به درستی فعالیت‌های سکوت/گفتار را تشخیص دهد.