A New VAD Algorithm using Sparse Representation and Updated Dictionary in Spectrogram Domain

Mohadese Eshaghi*

Abstract –This article proposes the new VAD (Voice Activity Detection) method was made using Spectrogram Domain (Spectro-Temporal Response Field) space based on sparse representation. Spectrogram Domain components have two dimensions of time and frequency. On the other hand, using sparse representation in learning dictionaries of speech and noise and updating dictionaries, causes better separation of speech and noise segments. In this algorithm, using auditory spectrogram and sparse representation, an updating dictionaries with different atom sizes and K-SVD (k-means clustering method) and NMF (non-negative matrix factorization) learning methods were constructed and the results indicate that this method works well. For example, the proposed VAD performance was obtained in SNRs greater than 0dB is more than 92.71% and 91.21% in White noise and Car noise respectively, which shows the good performance of the proposed VAD compared to other methods. By comparing the NDS and MSC evaluation parameters with other methods, the results show better performance of the proposed method.

Keywords: Spectro-Temporal Response Field, Voice Activity Detection (VAD), Sparse representation, Updating dictionaries, K-SVD,NMF.

1. Introduction

Voice Activity Detection plays an important role in communication and information transmission systems due to reducing costs and increasing the speed and improving the quality of information transmission [1]. This field has been researched for more than fifty years, but due to the statistical nature of speech and noise, improving VAD is still important for noisy environments, so many methods for VAD have been proposed. Some of these methods include spectral subtraction[2], linear prediction analysis[3] and wavelet transform. [4].

Recent psycho-acoustical and physiological findings in mammalian auditory systems, however, propose that the spectral decomposition is only the first stage of transformations in the representation of speech[5]. Specifically, it is thought that neurons in the auditory cortex analyze the spectrogram further reaches the content of its spectral-temporal modulation. This finding has inspired a multi-scale model representation of speech modulations that has been revealed its usefulness in speech representation, reproduction, intelligibility, discriminating speech from non-speech signals and describing a variety of other psycho-acoustic phenomena [6].

The early stage converts the sound waveform into a time-frequency distribution along a logarithmic frequency axis. The cortical stage works as a two-dimensional filter banks on the auditory spectrogram image to investigate efficient clues of different acoustic phenomena[7]. Each filter has a spectro-temporal impulse response (usually called as Spectro-Temporal Response Field (STRF)) in the form of a Gabor function which effectively is a multi-resolution wavelet filter. It is computationally performed by filter banks that are chosen by different modulation parameters ranging from slow to fast rates temporally and narrow to broad scales spectrally [8].Our main VAD features consist of sparse representation from the first step above.

Representation of signals in sparse become more interesting for various applications like restoration, compression and recognition. Sparse represents signal as a few elements from the dictionary atoms. There are various algorithms proposed by researchers for dictionary learning[9]. The sparse imputation method is able to restore the underlying clean speech even in very low SNR conditions provided a sufficiently accurate spectrographic mask can be created, for a single digit recognition task and whole word exemplars [10].

*** Corresponding Author:** Department of Electrical Engineering, Nowshahr Branch, Islamic Azad University, Nowshahr, Iran. Email: eshaghi@iauns.ac.ir

Received: 2021.06.04 ; Accepted: 2021.08.09

In this paper, we present the new approach to VAD systems; based on sparse representation of information in spectro-temporal domain. In the approach, separation of speech and noise regions is performed using auditory spectrogram and sparse decomposition.

This paper is organized as follows: Section 2 describes the background research including auditory spectrogram, sparse representation model and dictionary learning. The architecture of the proposed algorithm is presented in Section 3. Section 4 evaluates the performance of the proposed algorithm and section 5 concludes the paper.

2. Background

2.1. Auditory Spectrogram

According to mammalian auditory systems, a model for speech processing called Spectro-Temporal Response Field (STRF) was proposed. The Spectro-Temporal Response Field shows the change in the responses of the auditory cortex neurons in the face of sound, and in the next step, it decomposes the spectrogram of auditory speech into two dimensions of scale and rate (as shown in Figure 1).Our VAD key features constitute the sparse representation of the output of the first stage.

The early stages of auditory processing are modeled as a sequence of three operations. The entering acoustic signal produces a complex spatiotemporal pattern of vibrations along the basilar membrane of the cochlea. The basilar membrane outputs are then converted into inner hair cell intra–cellular potentials. This process is modeled as a 3–step operation: a high pass filter, followed by an instantaneous nonlinear compression ($g_{hc}(.)$), and then a low pass filter ($\mu_{hc}(t)$). Finally, a lateral inhibitory network (LIN) detects highly responses across the logarithmic frequency axis. The mathematical formulation for this stage can be summarized as [11]:

$$y_{cochlea}(t,f) = s(t) * h_{cochlea}(t;f) \quad (1)$$

$$y_{an}(t,f) = g_{cochlea}(\partial_t y_{cochlea}(t,f)) * \mu_{hc}(t) \quad (2)$$

$$y(t,f) = \max(\partial_f y_{an}(t,f),0) * \mu_{midbrain}(t,\tau) \quad (3)$$

Where $*$ denotes the convolution operand in time. The impulse response of each filter is denoted by$h_{cochlea}(t;f)$. The cochlear filter output is $y_{cochlea}(t,f)$. This output is then transduced into auditory-nerve patterns $y_{an}(t,f)$, which is the result of a frequency axis high pass filter followed by a nonlinear function $g_{cohlea}(.)$and a low pass time domain integrator $\mu_{hc}(t)$. y(t, f) is the auditory spectrogram.

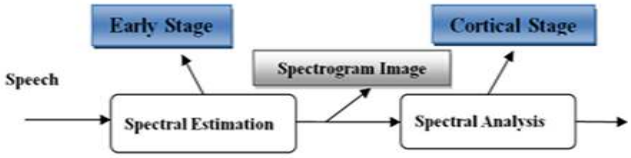

Figure 1. Block diagram of the auditory model.

2.2. Sparse Representation

Sparse representation [12]is adopted in sound recognition, speech denoising and speech enhancement. The main idea in sparse representation is to reconstruct each speech using a small number of base atoms.

If $y(n)$ is a noisy speech, it can be represented as a linear combination of samples a_i where the index i denotes a specific signal and N is the number of exemplars. We can write:

$$y = \sum_{i=1}^{N} A\alpha_i \quad (4)$$

Where $\alpha_i \in R$ are weights and A is a set of N vectors with K elements:

$$A = \begin{bmatrix} a_1^1 & a_2^1 & \dots & a_N^1 \\ a_1^2 & a_2^2 & \dots & a_N^2 \\ \vdots & \vdots & \dots & \vdots \\ a_1^K & a_2^K & \dots & a_N^K \end{bmatrix} \quad (5)$$

We can now express any speech frame y in frequency domain as:

$$y = A\alpha \quad (6)$$

Where $\alpha = [\alpha_1 \alpha_2 \dots \alpha_N]^T$ is an N-dimensional vector with minimum number of non-zero elements.

2.3. Dictionary learning

In some of the applications, the dictionary is pre-defined using methods such as wavelet packets (WP), discrete cosine transforms (DCT). They should be designed analytically for each special class of signals. Another approach is to learn a dictionary, based on a set of training signals, which is called dictionary learning[13].

To determine a high-quality dictionary, various learning algorithms have been proposed. These algorithms are typically composed of two major steps: 1) finding an approximate sparse representation of the training signals 2) updating the dictionary using the sparse representation[14].

Here we use two dictionary learning algorithms, namely

NMF and K-SVD. Two separate dictionaries are created for signal and noise signals.

In the next step, any estimate for each time frame, if the VAD output was speech, the features vector is added to the speech features vector and the atoms of speech dictionary and activation coefficients are updated again. If the VADoutput was noise, the features vector is added to the noise features vector and the components of noise dictionary and activation coefficients are updated again [15].

2.3.1 NMF

Non-negative matrix factorization algorithm (NMF) given a non-negative matrix *V*, finds non-negative matrix factors *A* and *H* such that:

$$V \approx AH \quad (7)$$

NMF can be applied to the statistical analysis of multivariate data in the following manner. Given a set of multivariate n-dimensional data vectors, the vectors are placed in the columns of $n \times m$ matrix V where m is the number of examples in the data set. This matrix is then approximately factorized into $n \times r$ matrix A and $r \times m$ matrix H. Usually r is chosen so that A and H are smaller than the original matrix V. This result is a compressed version of the original data matrix [16][17]

We used the following multiplicative update rules as a good compromise between speed and ease of implementation for NMF dictionary learning to generate speech and noise dictionaries employing SPAM [18].

$$\mathrm{H_{rm}} \leftarrow \mathrm{H_{rm}} \frac{(\mathrm{A^T V})_{\mathrm{rm}}}{(\mathrm{A^T AH})_{\mathrm{rm}}}, \mathrm{A_{nr}} \leftarrow \mathrm{A_{nr}} \frac{(\mathrm{VH^T})_{\mathrm{nr}}}{(\mathrm{AHH^T})_{\mathrm{nr}}} \quad (8)$$

2.3.2 K-SVD

K-SVD is a generalization of the k-means clustering method, and it works by iteratively alternating between sparse coding the input data based on the current dictionary, and updating the atoms in the dictionary to have a better fit to the data[19].

The main purpose in the K-SVD algorithm is to solve the following equation:

$$min_{\varphi,c}\{\|S-\varphi c\|\}_F^2 \ st.\ \forall i, \|c_i\|_0 \le T_0 \quad (9)$$

This algorithm has two basic steps: the first step is related to display the sparse and modify it, and the next is related to modify the columns of the dictionary-matrix columns[20].

Suppose that in the first step, φ is constant, and the previous equation to be solved in relation to matrix c, so we have by using of Frobenius norm:

$$\|S-\varphi c\|_F^2 = \sum_{i=1}^{k} \|S_i - \varphi c_i\|_F^2 \quad (10)$$

Consequently Eq. (9) is divided into k problem as follows:

$$min_{\varphi,c}\{\|S_i-\varphi c_i\|\}_2^2 \ st.\ \forall i, \|c_i\|_0 \le T_0 \quad (11)$$

$T_0$is considered small enough to keep the sparsity of c_i.

Eq. (22) can be solved using various methods, such as l_1 norm minimization, threshold method, norm approximation method, Basis Pursuit (BP), and Matching Pursuit (MP).

For the second step, assume that both the c and φ matrices are constant. Only one column from the φ_k dictionary matrix and its corresponding coefficient means kth row of c, means c_k^T to be variable, the Eq. (10) is rewritten as follows:

$$\|S-\varphi c\|_F^2 = \|S_i - \sum_{j=1}^{m} \varphi_i c_j^T\|_F^2 = \left\|\left(S_i - \sum_{j\neq k} \varphi_i c_j^T\right) - \varphi_k c_k^T\right\|_F^2 = \|E_k - \varphi_k c_k^T\|_F^2 \quad (12)$$

The E_k matrix calculates the error rate for the kth element of the equation to be deleted. To find φ_k and c_j^T, we can use the analysis of single value decomposition (SVD):

$$E_K = U\Lambda V^T \quad (13)$$

φ_k is defined first column of U and the vector coefficient c_K^T is the product of the first column V in the first layer Λ [21].

3. Proposed Algorithm

3.1 The Proposed Algorithm Based on Auditory Spectrogram and Sparse Representation

In this section, firstly, the input speech is converted into two-dimensional components of frequency and time, then with the help of the method of sparse and updated dictionaries of speech and noise, the segments of speech and noise of the input signal are separated. Figure 2 displays the block diagram of this method.

Spectro–temporal features of speech were produced using auditory spectrogram. These features (y), contains two dimensions, frequency and time (frame number). Then sliding window method [22] is used for the continuous speech signal, due to the presence of multiple concatenated phonemes. This two-dimensional signal is converted to windows with TW length. Δ (1 <Δ <TW) shows the level of overlapping of the windows. The window is larger, less calculations and the window is smaller, the more accurate the reconstruction. As a result, in every window, new feature vectors are extracted with smaller dimensions. Then, using sparse representation, speech and noise dictionaries are obtained for each window from the labeled training data.

The sparse formulation for each window subset can be

summarized as:

$$\alpha = \text{argmin}\{\|A\tilde{\alpha} - y_w\| + \lambda\|\tilde{\alpha}\|1\} \tilde{\alpha} \in IR^N \quad (12)$$

$$\widehat{y_w}^s = A^S\alpha^S \quad (13)$$

$$\widehat{y_w}^n = A^n\alpha^n \quad (14)$$

Where, y_w is the two-dimensional matrix of each window, w is the number index of windows, A^s is the matrix of components of the speech dictionary, and A^n is the components matrix of the noise dictionary $\widehat{y_w}^s$ and $\widehat{y_w}^n$ are the speech and noise estimates of each window using speech and noise dictionaries respectively and α is the activation coefficient vector of each window whose most values are zero. The speech and noise frames of speech can be recognized in each window using the minimum Euclidean distance of the sparse reconstructed signal on two dictionaries.

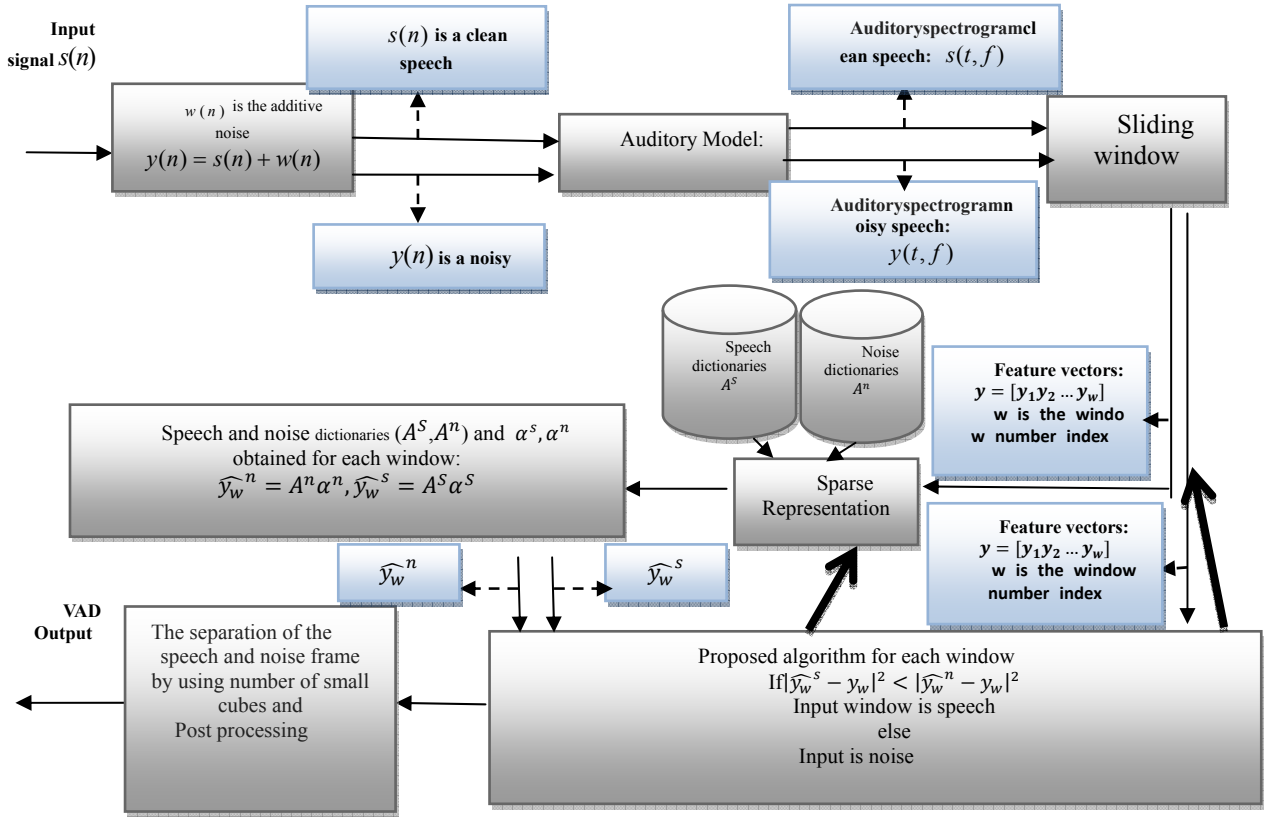

Figure 2. The block diagram of the proposed method.

$V(y_w)$is the VAD output of each window.

Figure 3 shows the VAD output decision algorithm of each window.

Figure 4 illustrates the VAD result of the proposed algorithm (first case of the third experiment) on a sample voice. Fig.4 (a) presents clear speech and the VAD output of the clean speech. Fig.4 (b) shows the speech mixed with white noise and the VAD result of the proposed algorithm, if the output is 0.2 (just as the graph indicator), it is speech, otherwise the output is assumed to be noise. The speech and noise signals are mixed to obtain -5 dB SNR.

Figure 5 displays the results on a different test sample. Fig. 5(a) presents the raw signal which is the same as Fig.4 (a) and shows the speech and noise sections of the clean signal as the ground truth. The speech signal in Fig.5 (b) was corrupted with babble noise to obtain 0dB SNR andthe VAD result of the proposed algorithm. If the output is 0.2, the frame has been recognized as speech; otherwise the output has been noise.

3.2 Post-processing

In the post-processing stage, the temporal nature of human speech is considered where speech (both vowel and consonant phonemes) never takes less than 100ms.

If $\left|\widehat{y_w}^{s} - y_w\right|^2 < \left|\widehat{y_w}^{n} - y_w\right|^2$
$V(y_w) = 1$// window
Otherwise
$V(y_w) = 0$ window
End

Figure 3. The VAD output decision algorithm of each window.

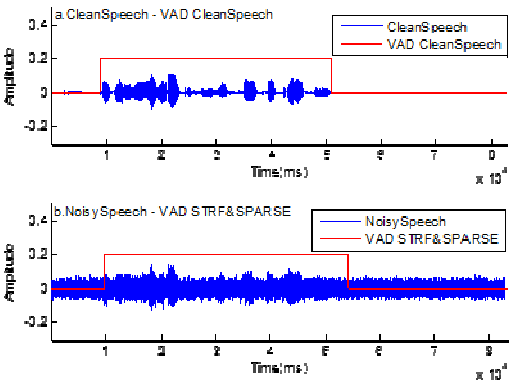

Figure 4. The VAD result of the proposed algorithm

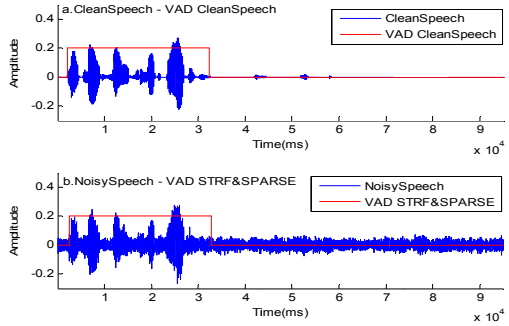

Figure 5. The VAD result of the proposed algorithm

So, the class of small duration portions of speech and noise, surrounded by the opposing class segments are inverted. This is because it never occurs to have a speech signal with the length of 32ms between different noisy frames. Similarly, a silence frame with the length of 32ms will not happen between different speech frames. The post-processing algorithm is applied as presented in Figure 6.

If $X_t \in NonSpeech$ & $X_{\substack{t-i \\ i=-3:3 \\ ,i\neq 0}} \in Speech$
Then $X_t \in Speech$
If $X_t \in Speech$ & $X_{\substack{t-i \\ i=-1:1, \\ i\neq 0}} \in NonSpeech$
Then $X_t \in NonSpeech$

Figure 6.The VAD Post-processing algorithm for any time frame.

4. Simulation Results

4.1. Evaluation of the proposed Algorithm

In this section, the results of simulations of the proposed algorithms are provided. All train and test clean speech utterances were selected from "TIMIT" database. Noise samples were taken from "NOISEX" database. The speech and noise signals were mixed in the test bench in order to control the signal-to-noise ratio (SNR).

The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance. The TIMIT corpus transcriptions have been hand verified. Test and training subsets, balanced for phonetic and dialectal coverage, are specified [23].

NOISEX consist of airport, babble, car, exhibition, office, restaurant, train, subway, street and white noises [24].

Table.1 shows the simulation conditions and parameters of simulated cases. The most important point in the sparse method is the production of a proper dictionary. The more complete the dictionary, the better the performance of this method, so two types of dictionaries are created simultaneously, the first one is created from the clean speech of speech training set and second one is derived from noise in the noise training set by NMF and KSVD methods. Of course, to investigate the effect of the dictionary dimensions, three kinds of speech and noise dictionary with 1000, 500 and 100 atoms were made from 11873 speech data and 9763 noise data, respectively.

There are two categories for research: the first one, that their dictionaries were trained using the NMF method, and the second category, the dictionaries were made using the K-SVD method and in each category, there are three groups that have different dictionary atoms.

Table 2 represents the average percentage of the proposed detector accuracy in the three groups of the first simulation in three different noises and different SNR. A review of the column in this table shows that as the number of atoms in the dictionary decreases, the average accuracy of the speech detector diminishes. Therefore, reducing the number of atoms in the dictionary increases the computational speed, but causes the loss of some valuable atoms in the dictionary, the separation of speech and noise in Babble noise, which has a similar structure to speech,

and Car noise, which is more powerful at low frequencies, is not done well.For example, the first line of Table 2shows the accuracy of this method in separating speech and noise in the presence of white noise and in SNRs greater than -5dB is more than 80%, which indicates the good performance of this method in removing white noise, so the first group with 1000 atoms It has better results in the dictionary than the other two groups. Therefore, decreasing dictionary atoms increases computational speed, but reduces the accuracy of speech and noise separation by removing valuable dictionary atoms. The results of this proposed method, similar to spectral processing methods, show good results in white noise, but do not work well in color noise.

Table 1. Simulation Test bench Parameters.

Parameters	**Description**
Sampling	16 KHz. 16 bits
frame size	32ms
window	5 frames
Δ	4 frames
Training utterances	190 TIMIT sentences
Test utterances	100 TIMIT sentences.
Types of noise	White, babble, car
Noise levels	-20dB to 20dB

Table 2. Performance of the proposed VAD under three types of noise and seven specific SNR values with NMF dictionary.

case	Number of atoms in dictionary	SNR / Noise	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE) 20	15	10	5	0	-5	-10
first	1000	White	93.15%	93.19%	93.02%	92.71%	90.91%	82.19%	74.22%
		Babble	96.05%	92.48%	89.36%	85.33%	81.39%	77.61%	73.17%
		Car	95.84%	94.62%	93.43%	91.21%	86.03%	80.82%	75.53%
second	500	White	93.01%	93.31%	92.06%	91.72%	87.93%	76.84%	70.11%
		Babble	92.45%	90.71%	86.24%	82.36%	79.02%	71.26%	65.93%
		Car	92.13%	92.23%	91.74%	86.69%	81.33%	74.57%	67.34%
third	100	White	98.81%	94.52%	94.04%	87.44%	88.02%	76.58%	69.32%
		Babble	92.98%	91.09%	87.42%	81.05%	73.75%	70.91%	63.22%
		Car	97.63%	91.04%	88.09%	82.53%	79.47%	72.76%	61.48%

Table 3. Performance of the proposed VAD under three types of noise and seven specific SNR values with K-SVD dictionary.

case	Number of atoms in dictionary	SNR / Noise	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE) 20	15	10	5	0	-5	-10
first	1000	White	96.13%	96.13%	92.74%	91.86%	85.19%	80.01%	80.01%
		Babble	95.48%	95.48%	86.33%	82.24%	71.35%	68.62%	68.62%
		Car	96.92%	96.92%	90.87%	85.31%	73.08%	66.53%	66.53%
second	500	White	97.68%	97.68%	94.82%	91.08%	85.44%	76.79%	76.79%
		Babble	94.41%	94.41%	83.95%	78.54%	70.06%	65.51%	65.51%
		Car	96.15%	96.15%	89.26%	83.88%	72.04%	66.04%	66.04%
third	100	White	97.42%	97.42%	91.13%	88.05%	83.90%	75.32%	75.32%
		Babble	94.31%	94.31%	83.41%	77.68%	69.43%	63.24%	63.24%
		Car	96.11%	96.11%	90.00%	81.79%	71.62%	68.14%	68.14%

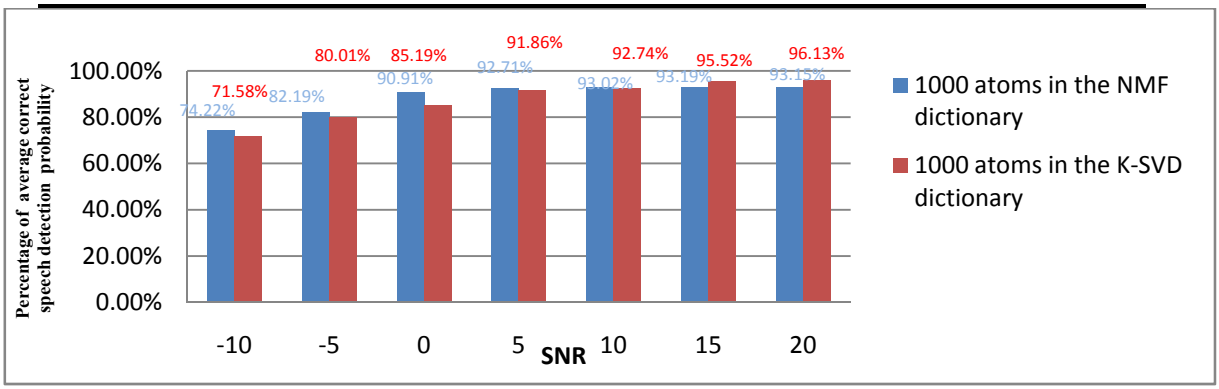

Figure 7. Comparison of the best results of two experiments of proposed VAD in White noise.

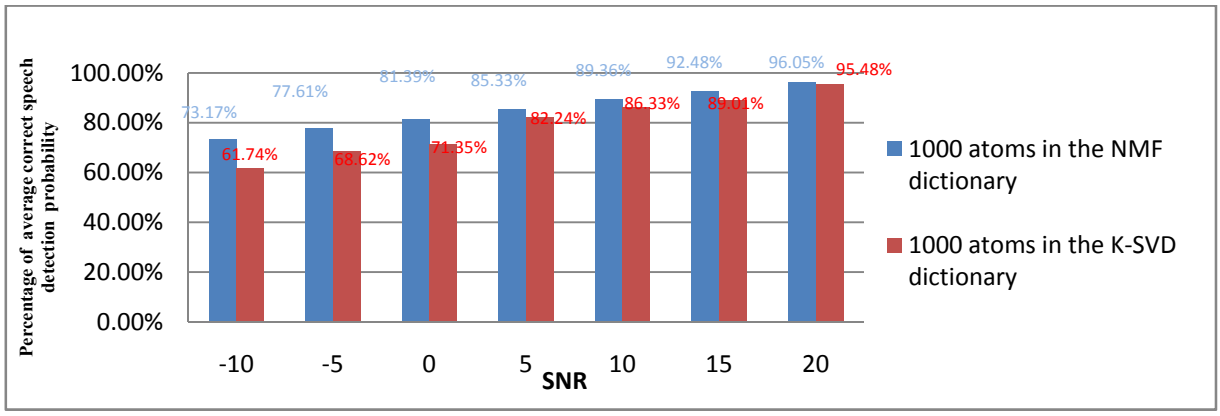

Figure 8. Comparison of the best results of two experiments of proposed VAD in Babble noise.

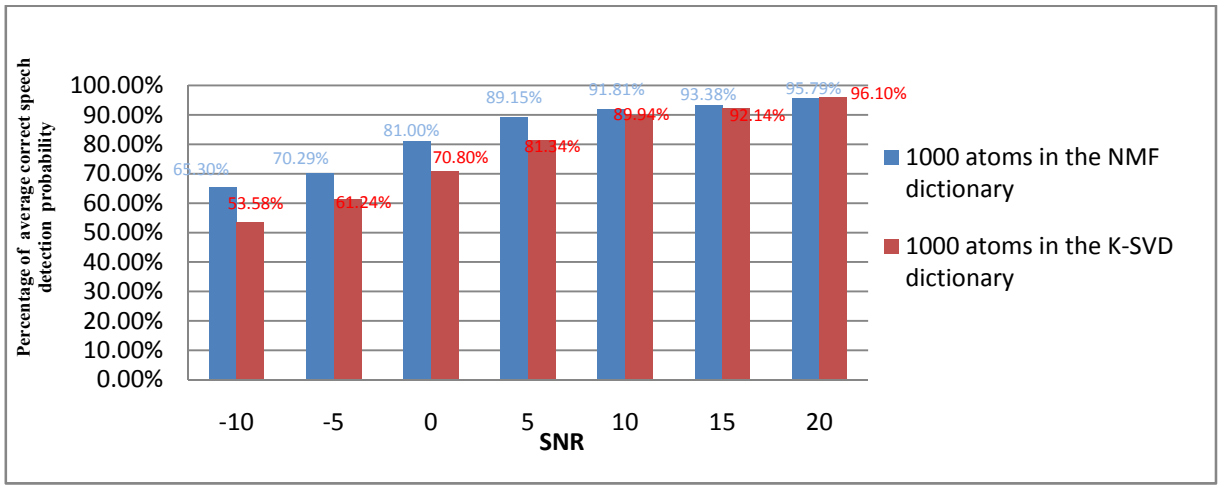

Figure 9. Comparison of the best results of two experiments of proposed VAD in Car noise.

Table 3 indicates the accuracy average percentage of the proposed detector in the three groups of the second simulation that in three different noises and with different SNR. A column study of this table shows the same results as Table 2, which represents that as the number of atoms in the dictionary decreases, the average detector accuracy decreases in all noises. Therefore, reducing the number of atoms in the dictionary increases the computational speed but causes the loss of some main atoms in the dictionary. Using general frequency instead of local frequency, which does not work well in color noise and makes speech and noise separation difficult. But it works well in removing white noise, for example, the fourth line of Table 3 shows the accuracy of this method in separating the speech and noise parts in white noise in SNRs higher than 0dB is more than 84%, which indicates the good performance of this method in removing white noise. Investigations of table 3 shows that in the first category, the first group has better results than the other two groups, so the reduction of the dictionary atoms does not only affect the response speed but also reduces accuracy.

For computational complexity, the number of atoms in a dictionary should be considered. The larger dictionary, the more complicated the calculations. However, this complexity does not cause to slow down the simulation. As the computational complexities of three groups are the same; therefore, the obtained results in comparing these three categories suggest a better performance of each of the categories and groups.

Figure 7, Figure 8, and Figure 9, show the comparison of the best results of both simulations in white, babble and car noises, Comparison of these three figures does not show a significant difference between the NMF method and the KSVD method in dictionary learning, although the first category of first simulation, whose dictionary was trained

by NMF method, has better results than other methods and its results are used to compare with other methods.

4.2 Results Comparison

The proposed algorithm was suggested using sparse representation of two-dimensional auditory spectrogram space exploring different atomic sizes of dictionaries and two dictionary learning methods. It showed good results in different SNRs.

The superior performance of the proposed VAD is illustrated through nonspeech–speech error (NDS) and speech–nonspeech error (MSC) [25]. Noise detected as speech (NDS) is the proportion of nonspeech frames which are misclassified as being speech. Mid-speech clipping (MSC) is the proportion of speech frames erroneously classified as being non-speech.

In comparison, some of up-to-date voice-activity detection methods were compared to the proposed VAD algorithms, which have been proved to be noise robust. They are LTSV [26], Sohn [27], G.729B [28], Mesgerani's VAD [29], Harmfreq [30], LTSD [31], LSFM [32] and LTPD [33].

In Figure 10, the proposed VAD has even less error rate under zero SNRs comparing to other VADs. The proposed speech detector has better results than other methods in the spectral field due to the use of valuable data of the auditory spectrogram and dictionaries updating with the help of sparse representation and re-estimation of speech and noise frames. Of course, although the Mesgerani's VAD uses the four-dimensional components of the STRF space, at low SNRs it considers all noisy speech to be noise, so there is no separation and no NDS error. In contrast, our proposed VAD successfully separates speech and noise parts at low error rates.

Figure 11 represents a comparison of the proposed method with other methods in the MSC error. G729 and LTSD detectors have less error in low SNRs than the proposed method, but a review of Figure 10 shows that these two VADs consider a large part of noisy speech as speech, so they do not make separation between speech and noise. They also have very few MSC errors. In contrast, our VAD successfully separates parts of speech and noise.

Totally, comparing the results of these VADs; it can be observed that the proposed VAD performs better than the other state of the art comparing methods.

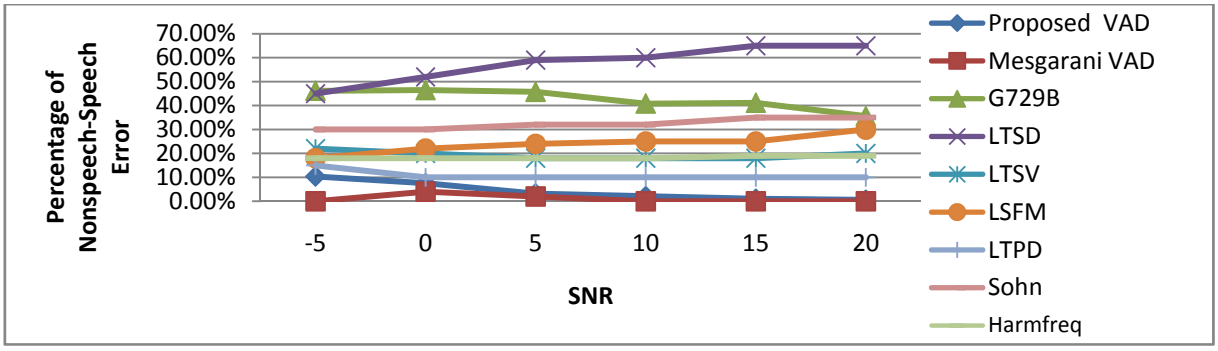

Figure 10.Nonspeech–speech Error(NDS).

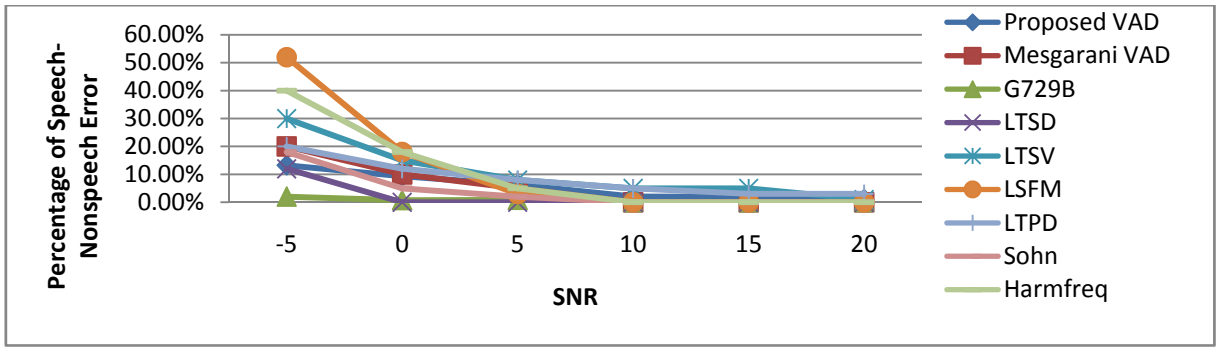

Figure 11. Speech–nonspeech Error(MSC).

5. Conclusion

In this study, speech detector is proposed using a sparse representation in the audio spectrogram space. The results of proposed VAD show that the use of sparse representations and dictionaries updating in two-dimensional spectral-temporal space provides components that put out better and more important information for a more accurate separation of speech and noise. In this algorithm, three groups with different number of atoms in two categories were simulated by different dictionary learning methods. Examination of the results of NDS and MSC evaluation parameters showed good performance of the proposed VAD compared to other methods.

The most important disadvantage of the proposed method is the use of time frames with constant frequency width, which considers the overall frequency instead of the local frequency that is removed by using four-dimensional components of spectral-temporal space. Also to solve the problem of large dimensions of spectral-temporal space, sparse representation can be used. Therefore, using a sparse representation and constructing an appropriate dictionary in spectral-temporal domain may give us hope for good results in separating parts of speech and noise.

References

[1] R. Johny Elton, J. Mohanalin and P. Vasuki,“A novel voice activity detection algorithm using modified global thresholding,” International Journal of Speech Technology, vol. 24, pp. 127–142, 2021.

[2] M. Eshaghi and M.R. Karami Mollaei,“Voice activitydetection based on using wavelet packet,” Digital Signal Processing, vol. 20, pp. 1102-1115, 2010.

[3] C.T. Hsieh, P.Y. Huang, T.W. Chen and Y. Chen,“Speech enhancement based on sparse representation under color noisy environment,” 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS),pp.134 - 138,2015.

[4] G. Martin, A. Abeer, E. Dan and et al.,“All for one: feature combination for highly channel-degraded speech activity detection,”INTERSPEECH, Lyon 2013, pp.709–713, 2013.

[5] M. Kolbæk, Zh. Tan , S. Jensen and J. Jensen,“on Loss Functions for Supervised Monaural Time-Domain Speech Enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 825-838, 2020.

[6] M. Eshaghi,F. Razzazi and A. Behrad,“A New VAD Algorithm using Sparse Representation in Spectro-Temporal Domain,”Journal of Information Systems and Telecominication (JIST),vol. 7, pp.709–713, 2019.

[7] M. Mirbagheri, N. Mesgarani, and Sh. Shamma,“Nonlinear filtering of spectro-temporal modulation in speech enhancement,”2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5478-5481,2010.

[8] N. Mesgarani, S. David, and S.A. Shamma, “Representation of phoneme in primary auditory cortex: how the brain analyzes speech,”2007 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 765-768, 2007.

[9] M. Eshaghi, F. Razzazi and A. Behrad,“A voice activity detection algorithm in spectro-temporal domain using sparse representation,” International Journal of Machine Learning and Cybernetics, vol. 10,pp. 1791–1803, 2019.

[10] W. Li, Y. Zhou, N. Poh, F. Zhou, and Q. Liao,“ Feature Denoising Using Joint Sparse Representation for In-car Speech Recognition,”IEEE Transactions on audio, speech, and language processing, vol. 20, pp. 681-684, 2013.

[11] C. Mart´ınez, J. Goddardb, D. Milone, and H. Rufiner,“sparse spectro-temporal representation of speech forrobust classification,” Computer Speech and Language,vol.26, pp. 336-345,2012.

[12] M. Elad,“Sparse and redundant representations: from theory to applicationsin signal and image processing,”Springer Science & BusinessMedia, 2010.

[13] R. Rubinstein, A. M. Bruckstein and M. Elad,“Dictionaries for sparserepresentation

modeling," Proceedings of the IEEE,vol. 98, pp.1045–1057, 2010.

[14] M. Wei, Zh. Liu, X. Chen and H. Zhao,"Speech enhancement based on sparse representation using joint dictionary,"2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE),vol. 80, pp.500–503, 2018.

[15] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee and T.J. Sejnowski,"Dictionary learning algorithms for sparse representation," Neural Computer,vol. 15, pp.349–396,2003.

[16] P. O. Hoyer,"Non-negative matrix factorization with sparseness con-straints,"The Journal of Machine Learning Research,vol. 5, pp. 1457–1469,2004.

[17] M. Aharon, M. Elad, and A. Bruckstein,"K-svd: A algorithm for designing over complete dictionaries for sparse representation,"IEEE Transactions on Signal Processing,vol.54, pp.4311–4322, 2006.

[18] R. Zdunek, and A. Cichocki,"Non-negative matrix factorization with quadratic programming,"Neural computation,vol. 71, pp. 2309-2320, 2007.

[19] G. H. Mohimani, M. Babaie-Zadeh and Ch. Jutten,"A fast approach for overcomplete sparse decomposition based on smoothed L0 norm," IEEE Transactions on Signal Processing,vol.57, pp.289-301,2009.

[20] M. S. Lewicki and T. J. Sejnowski,"Learning overcomplete represen-tations," Neural computation,vol. 12, pp. 337–365, 2000.

[21] Z. Jiang, G. Zhang, and L. S. Davis,"Submodular dictionary learn-ing for sparse coding,"2012IEEE Conference on Computer Vision andPattern Recognition (CVPR), pp. 3418–3425, 2012.

[22] J.F. Gemmeke, H.V. Hamme, B. Cranen and L. Boves ,"Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition,"IEEE Journal of selected topics in signal processing,vol. 4, pp. 273-82, 2010.

[23] W. M. Fisher, G. R. Doddington, M. Goudie and M. Kathleen,"The DARPA speech recognition research database: specifications and status," Proceedings of DARPA Workshop on Speech Recognition,CD-ROMs, 2005.

[24] A. Varga, H. J. M. Steeneken, M. Tomlinson and D. Jones,"The NOISEX-92 study the effect of additive noise on automatic speech recognition," Documentation included in the NOISEX-92 CD-ROMs, 1992.

[25] J. McLoughlin,"Super-Audible Voice Activity Detection," IEEE Transactions on Speech and Audio Processing,vol.22, pp.1424-1433, 2014.

[26] P.K. Ghosh, A. Tsiartas and S. Narayanan,"Robust voice activity detection using long-term signal variability,"IEEE Transactions on Audio, Speech and Language Processing,vol. 11, pp. 600–613,2011.

[27] J. Sohn, N. S. Kim and W. Sung,"A statistical model-based voice activity detection,"IEEE Signal Process,vol. 6 , pp.1–3,1999.

[28] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin and J. P. Petit,"ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,"IEEE Communications Magazine,vol. 35, pp. 64-73, 1997.

[29] N. Mesgarani and Sh. Shamma,"Denoising in the Domain of Spectro-temporal Modulations," EURASIP Journal on Audio, Speech, and Music Processing,vol. 12, pp. 1-9 ,2007.

[30] L. N. Tan, B. J. Borgstrom, and A. Alwan,"Voice activity detection using harmonic frequency components in likelihood ratio test,"2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4466 - 4469, 2010.

[31] J. Ramirez, J. Segura, C. Benitez, A. Torre and A. Rubio," Voice activity detection with noise reduction and long-term spectral divergence estimation,"2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.271–287, 2004.

[32] M. Yanna and A. Nishihara,"Efficient voice activity detection algorithm using long-term spectral flatness measure,"EURASIP Journal on Audio, Speech, and Music Processing, ,vol. 87, pp. 1-18, 2013.

[33] X.K Yang, L. He, D. Qu and W. Q.Zhang,"Voice activity detection algorithm based on long-term pitch information," EURASIP Journal on Audio, Speech, and Music Processing, vol. 14, pp. 1-9 ,2016.