A New VAD Algorithm using Sparse Representation in Spectro-Temporal Domain

Mohadese Eshaghi
Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
eshaghi463@yahoo.com
Farbod Razzazi *
Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
razzazi@srbiau.ac.ir
Alireza Behrad
Department of Electrical and Electronic Engineering, Shahed University, Tehran, Iran.
behrad@shahed.ac.ir

Received: 03/Apr/2019 Revised: 25/Aug/2019 Accepted: 16/Sep/2019

Abstract

This paper proposes two algorithms for Voice Activity Detection (VAD) based on sparse representation in spectro-temporal domain. Spectral-temporal components which, in addition to the frequency and time dimensions, have two other dimensions of the scale and rate. Scale means spectral modulation and the rate means temporal modulation. On the other hand, using sparse representation in learning dictionaries of speech and noise, separate the speech and noise segment to be better separated. The first algorithm was made using two-dimensional STRF (Spectro-Temporal Response Field) space based on sparse representation. Dictionaries with different atomic sizes and two dictionary learning methods: NMF (non-negative matrix factorization) and the K-SVD (k-means clustering method), were investigated in this approach. This algorithm revealed good results at high SNRs (signal-to-noise ratio). The second algorithm, whose approach is more complicated, suggests a speech detector using the sparse representation in four-dimensional STRF space. Due to the large volume of STRF's four-dimensional space, this space was divided into cubes, with dictionaries made for each cube separately by NMF (non-negative matrix factorization) learning algorithm. Simulation results were presented to illustrate the effectiveness of our new VAD algorithms. The results revealed that the achieved performance was 90.11% and 91.75% under -5 dB SNR in white and car noise respectively, outperforming most of the state-of-the-art VAD algorithms.

Keywords: Speech Processing; Voice Activity Detector; VAD; Spectro-Temporal Domain Representation; Sparse Representation, NMF, K-SVD.

1. Introduction

Many practical speech processing systems are in use today. Voice activity detection (VAD) unit, which discriminates speech segments from environmental noise segment, is an integral part of a variety of speech communication systems. Speech coding, speech recognition, hands-free telephony, and echo cancellation are some examples of these systems. However, developing a VAD for noisy environments with low signal-to-noise ratios or for any non-stationary noise is still very challenging [1–4].

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

In this model, the primary stage converts the sound waveform into a time-frequency distribution along a logarithmic frequency axis. The cortical stage works as a two-dimensional filter bank on the auditory spectrogram image to investigate efficient clues of different acoustic phenomena. Each filter has a spectro-temporal impulse response (usually called spectro-temporal response field (STRF)) in the form of a Gabor function which is effectively a multi-resolution wavelet filter [7, 8].

STRFs decompose the content of auditory spectrogram into the scale-rate domain. The scale represents the spectral modulation rate with the unit defined as cycles/octave (or cycles/kHz). Also, the rate means temporal modulation

*Corresponding Author

variations with the unit being cycles/second (Hz) [9]. This multi-domain representation model of speech is proven to be useful in estimation of speech intelligibility [10]. Our VAD key features are captured by the sparse representation of the above two stages.

Sparse solutions have recently attracted a great deal of attention because of their potential applications in many different areas. They are used, for example, in compressed sensing, under-determined sparse component analysis (SCA), and source separation based on atomic decomposition on over-complete dictionaries [11, 12]. In this article, we present and assess the two new approaches to VAD systems based on sparse nature of information in spectro-temporal domain. In the first approach, separation of speech and noise regions is performed using auditory spectrogram and sparse decomposition.

In the second approach, we transform the input utterance into spectro-temporal domain. Owing to large dimensions in the space, each temporal frame in the new domain is divided into small 3D sub-cubes. Then, using sparse decomposition of the sub-cubes on pre-trained speech and noise dictionaries for each sub-cube, speech and noisy parts are classified by combining the results of sparse classification of the sub-cubes within a time frame.

This paper is organized as follows: Section 2 describes the architecture of the system: sparse representation model and the proposed algorithm. Section 3 evaluates the performance of the proposed algorithm, and finally section 4 concludes the paper.

2. Proposed Algorithms

2.1 System Architecture

The first algorithms were proposed using two-dimensional STRF space and via sparse representation in dictionaries with different atomic sizes and two learning methods of K-SVD (generalizing the K-means clustering process) and NMF.

The second algorithm, which is more complicated than the first algorithm, suggests a speech detector in the STRF four-dimensional domain. Due to the large volume of STRF's four-dimensional space, this space was divided into cubes, with dictionaries created for each cube separately by NMF learning method. The results indicate a better performance of the proposed method in STRF space.

2.1.1 The First Algorithm using Auditory Model

In this method, first the input speech is converted to two-dimensional components of time-frequency space. Then, the speech and non-speech frames of the input signal are separated by the sparse method as well as speech and noise dictionaries. Fig. 1 displays the block diagram of this method.

The input signal is converted to two-dimensional auditory spectrogram y(t,f). Then, the sliding window method [13] is used for the continuous speech signal owing to the presence of multiple concatenated phonemes. This two-dimensional signal is converted to windows with T_W length. Δ ($1 < \Delta < T_W$) represents the extent of overlap between the windows. Larger step sizes of Δ reduce the computational demand, but can decrease its accuracy.

Y Character vector, y_w is the two-dimensional matrix of each window, w denotes the number index of windows, A_s is the matrix of components of the speech dictionary, and A_n shows the components matrix of the noise dictionary. $\widehat{Y_w}^s$ and $\widehat{y_w}^n$ indicate 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 maximum value is zero.

Fig. 2 illustrates the VAD results of the first proposed algorithm on a sample voice. Fig. 2(a) presents clear speech and the VAD output of the clean speech.

The spectrogram of clean speech is demonstrated in Fig. 2(b) (30 s of silence were added to a clear speech utterance). Fig. 2(c) illustrates the mixed speech with white noise and The VAD results of the proposed algorithm. According to the graph, if the value of the output is 1, it is assumed to be speech; otherwise it is assumed to be noise. The speech and noise signals are mixed to obtain SNR of 5 dB for the signal. Fig. 2(d) exhibits the spectrogram of noisy speech.

2.1.2 The Second Algorithm using STRF domain

The architecture of the proposed VAD algorithm is depicted in Fig. 3. As can be observed, after receiving the input raw speech signal samples by a microphone or other sources, it has been divided into the sequence of time frames. Then, the spectral–temporal features of speech are extracted using the auditory model. These features (Z) include four dimensions of Ω scales (cycles/octave), ω rates (Hz), frequency, and time (frame number). In the next step, the representation space is partitioned into small cubes (named sub-cubes) to manipulate the large volume of data in the resultant four–dimensional space, i.e. each time frame is divided into small 3D sub-cubes.

As a result, in each frame, new feature vectors are extracted with smaller dimensions. In the training phase, speech and noise dictionaries are obtained by dictionary learning algorithms for each sub-cube from the training labeled data. Z_J is a four-dimensional representation of each cube, where J is the cube number index, A_s and A_n represent the speech and noise dictionary matrices, $\widehat{z_j}^s$ and $\widehat{z_j}^n$ denote the speech and noise estimates of each sub-cube using speech and noise dictionaries respectively, and α is the activation coefficient vector of each sub-cube resulting from the sparse representation. In the next step, for each sub-cube the speech and noise frames of speech can be classified for each sub-cube using the proposed algorithm. The final step of this

system is the combining of the classification results of sub-cubes by majority voting among the classification results of the sub-cubes of each frame [14].

Fig. 4 displays the results on a different test sample. Fig. 4(a) presents the raw signal as well as the speech and noise sections of the clean signal as the ground truth. Fig. 4(b,d) represent the spectrogram of clean and noisy speech, which is the same as Fig. 3(b,d). The speech signal in Fig. 4(c) was distorted with babble noise to obtain 5dB SNR and the VAD result of the proposed algorithm.

Comparison of the red diagram of Fig. 3(a) (where the speech and non-speech segments have been manually separated according to the text presented by the "TIMIT" database) with the red diagram of Fig. 3(c) (where the speech and noise segments have been separated by the proposed VAD), as well as red diagram of Fig. 4(a) with the red diagram of Fig. 4(c), reveals the good performance of the second proposed VAD in the separation of speech and non-speech segments. It can be observed that in the absence of information on the structure of the speech and noise, these systems have performed accurately and acceptably. The following subsections capture the functions of the blocks in Fig. 1 and Fig. 3.

2.2 The Proposed Algorithms Based on STRF and Sparse Representation

2.2.1 The First Algorithm using Auditory Model

According to the block diagram of the proposed system in Fig. 1, the spectro–temporal features of speech were extracted using the auditory spectrogram. These features (y) include two dimensions, frequency and time (frame number). Then, the sliding window method [12] is used for the continuous speech signal given the presence of multiple concatenated phonemes. This two-dimensional signal is converted to windows with a T_W length. Δ ($1 < \Delta < T_W$) reflects the extent of overlap between the windows. Larger windows are associated with lighter calculations, while the smaller the window, the more careful the reconstruction will be. Accordingly, 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:

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

$$\widehat{y_w}^s = A^s\alpha^s \tag{2}$$

$$\widehat{y_w}^n = A^n\alpha^n \tag{3}$$

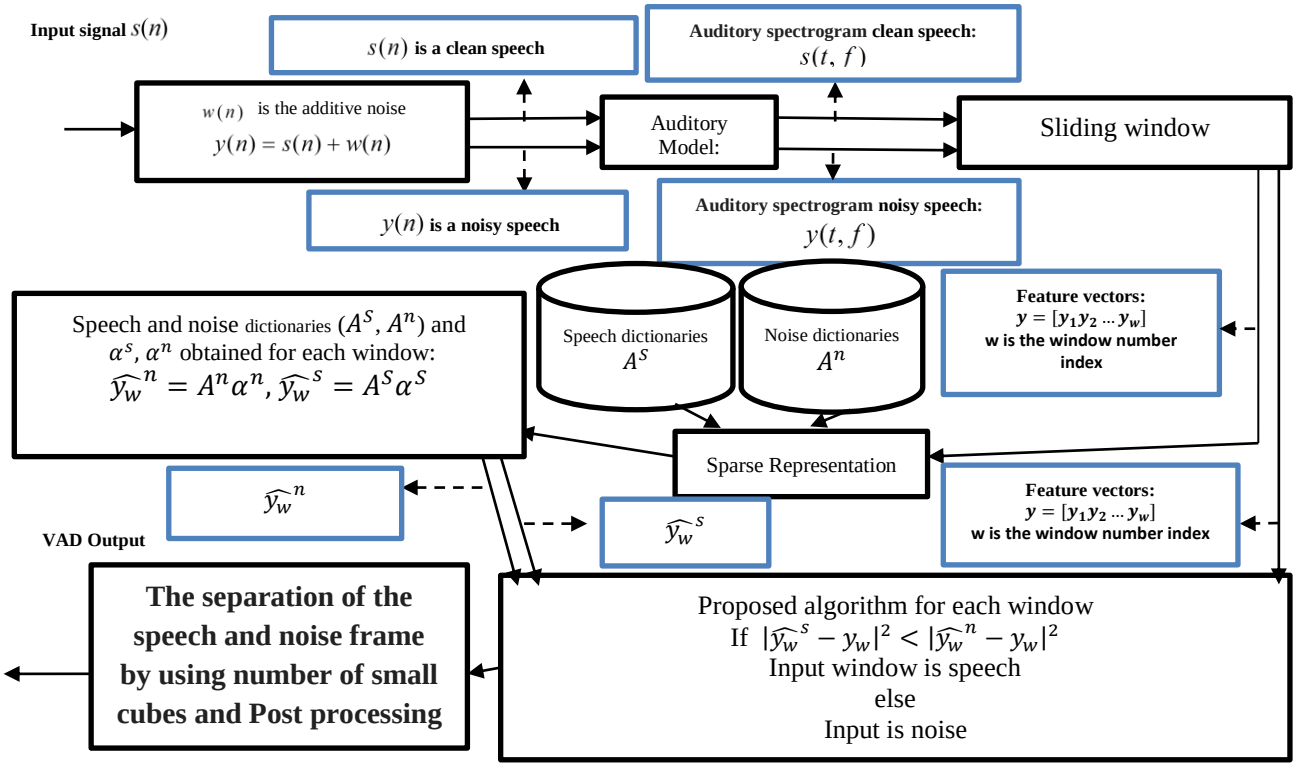

Fig. 1 The block diagram of the proposed first method.

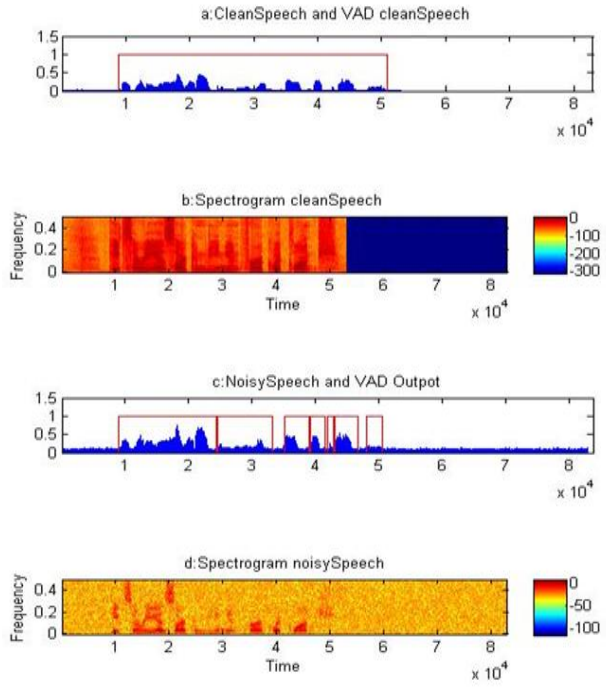

Fig. 2 The VAD results of the proposed first algorithm.

y_w is the two-dimensional matrix of each window, w is the number index of windows, A_s represents the matrix of components of the speech dictionary, and A_n shows the components matrix of the noise dictionary.

Also $\widehat{y_w}^s$ and $\widehat{y_w}^n$ are the speech and noise estimates of each window using speech and noise dictionaries respectively, and α denotes the activation coefficient vector of each window whose most value is 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. $V(y_w)$ refers to the VAD output of each window. Fig. 5 displays the VAD output decision algorithm of each window.

2.2.2 The Second Algorithm using STRF domain

Based on the block diagram of the proposed system in Fig. 3, initially the spectral–temporal features of speech have been extracted using STRF filter bank. These features (Z) included four dimensions of Ω densities or scales (cycles/octave), ω rates or velocities (Hz), frequency, and time (frame number). The auditory cortical model is obtained using spectral-temporal filter banks, with each of these filters operating within the range of different rates and scales. The space is partitioned into small sub-cubes to manipulate the large volume of data in the four–dimensional space, i.e. each time frame is divided into small cubes (Fig. 6). Hence, in each frame, new feature vectors are extracted with smaller dimensions, after which using sparse representation, speech and noise dictionaries are obtained for each cube from the labeled training data.

The sparse formulation for each cubic subset can be summarized as:

$$\alpha = argmin\{\|A\tilde{\alpha} - Z_J\| + \lambda\|\tilde{\alpha}\|1\} \quad \tilde{\alpha} \in IR^N \tag{4}$$

$$\widehat{z_J}^s = A^S\alpha^S \tag{5}$$

$$\widehat{z_J}^n = A^n\alpha^n \tag{6}$$

Z_J is a four-dimensional matrix of clean speech each cube, J refers to the cube number index, A_s and A_n represent the speech and noise dictionary matrices, $\widehat{z_J}^s$ denotes the speech estimation of each cube using speech dictionary, $\widehat{z_J}^n$ is the noise estimation of each cube using noise dictionary and α shows the activation vector of the cube . Most of the elements of this vector are zero. The speech and noise frames of speech can be recognized in each cube using the minimum Euclidean distance of the sparse reconstructed signal on two dictionaries. $V(Z_J)$ is the VAD output of each sub-cube. Fig. 7 indicates the VAD output decision algorithm of each sub-cube.

In the last step, the decision on the cubes is fused together by majority voting. If the number of speech cubes is greater than the number of noisy cubes within a time frame, the frame is regarded as a speech frame; otherwise it is noise.

2.3 Dictionary learning

Most dictionary learning algorithms use two-step iterative techniques to solve the problem. In the first step, they use a sparse representation algorithm to determine the sparse coefficients given by the dictionary. In the second step, they update the dictionary based on some criteria such as maximizing a likelihood probability or minimizing a cost function [15].

Here, we use two dictionary learning algorithms called NMF [16, 17, 18] and KSVD [19]. Two separate dictionaries are created for signal and noise signals.

2.4 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. So, the class of small duration portions of speech and noise, surrounded by the opposing class segments, is 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 either between different speech frames.

A 32ms single frame noise cannot lie between two speech frames. Again, a single 32ms frame of speech cannot be placed between two noise frames either.

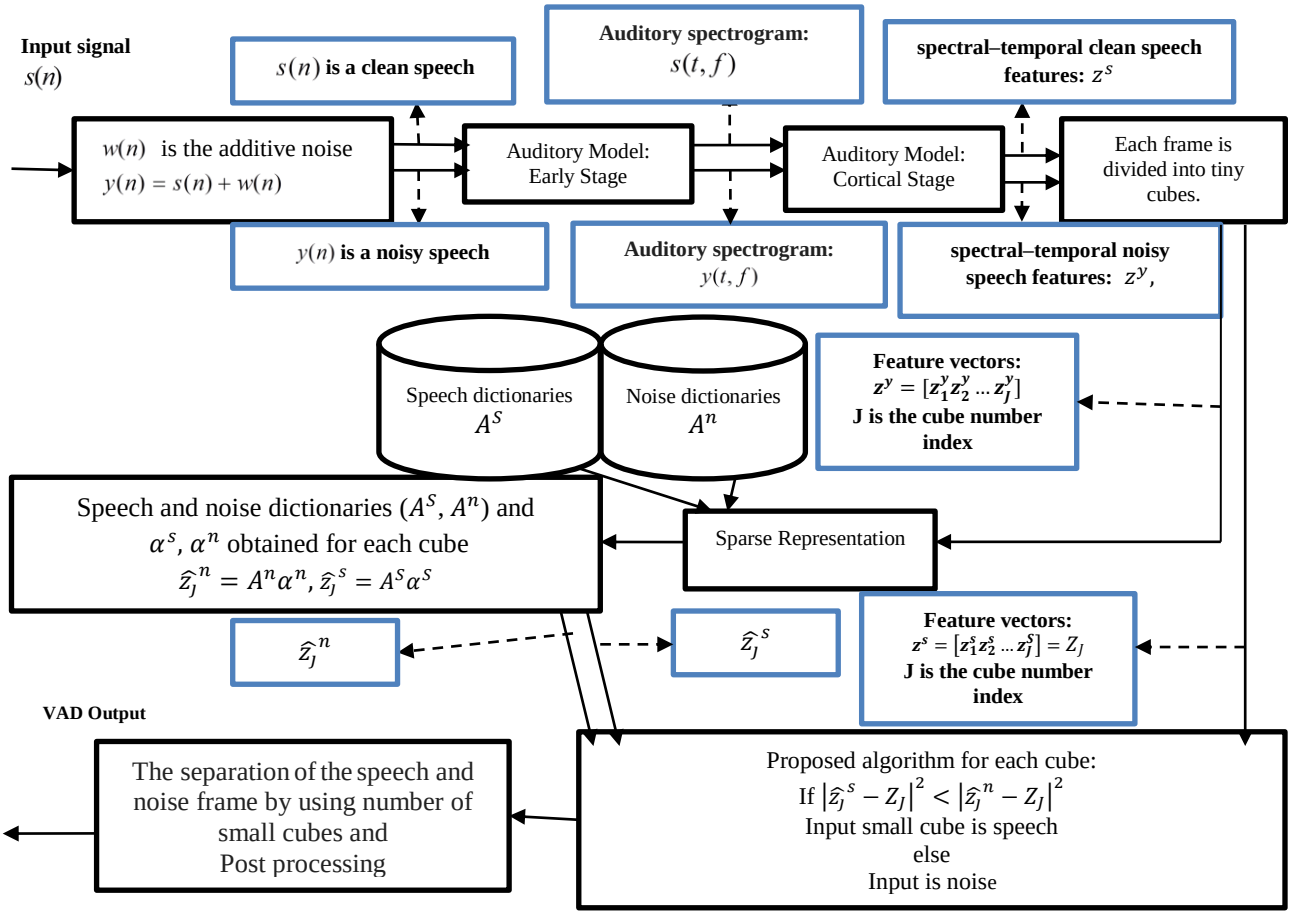

Fig. 3 The block diagram of the proposed second system.

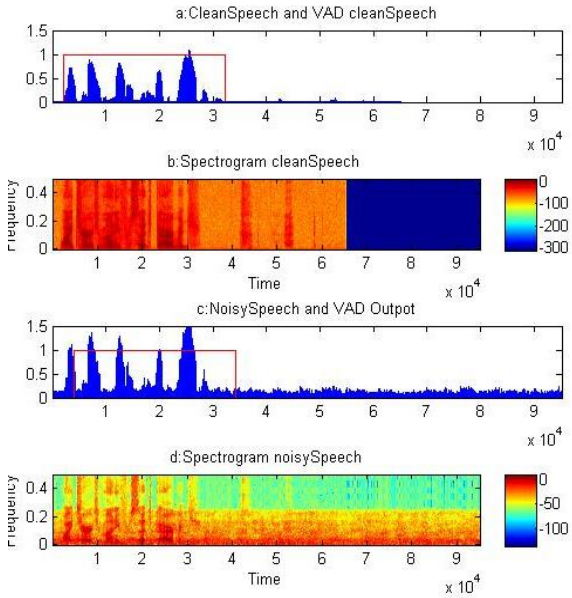

Fig. 4 The VAD results of the proposed second algorithm.

If $|\widehat{y_w}^s - y_w|^2 < |\widehat{y_w}^n - y_w|^2$
$V(y_w) = 1$ //1 means speech window
Otherwise
$V(y_w) = 0$ //0 means non-speech window
end

Fig. 5 The VAD output decision algorithm of each window.

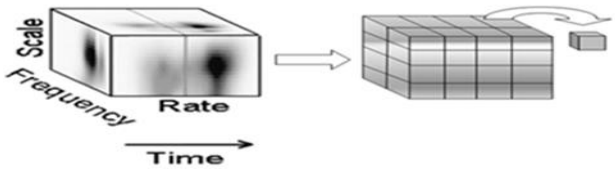

Fig. 6 Each frame 3D spectro-temporal space is divided into tiny cubes.

If 1 denotes a speech frame and 0 denotes a non-speech frame, and $V(X_t)$ is the VAD output for any time frame, the post-processing algorithm is applied as presented in Fig. 8.

If $|\hat{z}_j^s - Z_J|^2 < |\hat{z}_j^n - Z_J|^2$
$V(Z_J) = 1$ //1 means speech cube
Otherwise
$V(Z_J) = 0$ //0 means non-speech cube
end

Fig. 7 The VAD output decision algorithm of each cube.

for t=3 to T-3
If $V(X_t) = 0$ //0 means non-speech frame

$$V(X_t)_{out} = (1 - V(X_t)) \prod_{i=-3}^{3} V(X_{t+i})$$

If $V(X_t) = 1$ //1 means speech frame

$$V(X_t)_{out} = 1 - \left\{ V(X_t) \prod_{i=-3}^{3} (1 - V(X_{t+i})) \right\}$$

end for

Fig. 8 The VAD post-processing algorithm for any time frame.

3. Simulation Results

This section provides the simulation results of these proposed algorithms. All training 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 quality recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes 16-bit, 16 kHz speech waveform files for each utterance. [20].

NOISEX database includes airport, babble, car, exhibition, office, restaurant, train, subway, street, and white noises [21].

We applied the proper level of background noise to set the SNR to the desired values.

3.1 Evaluation of the First Algorithm

Table 1 shows the simulation parameters. The most important point in the sparse method is generation of a proper dictionary. The more complete the dictionary, the better the performance of this method will be. Thus, two types of dictionaries are created simultaneously, with the first created from the clean speech of the speech training set and second derived from noise in the noise training set by NMF and KSVD methods. 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.

Accordingly, there are two categories for the research: i) those whose dictionary was prepared using the NMF method; and ii) the categories whose dictionaries were trained using the K-SVD method. In each category, there are three groups that have different dictionary atoms.

Tables 2 and 3 report the average percent accuracy of the proposed detector in the first two simulation groups at four different noises and different SNRs, respectively.

For computational complexity, the number of atoms in a dictionary should be considered. The larger the dictionary, the more complicated the calculations will be. However, this complexity does not slow down the simulation. As the computational complexities of three groups are the same, therefore in comparing these three categories the obtained results suggest a better performance of each of the categories and groups. According to Table 2, in the first category, the first group has better results than the other two groups. Thus, the reduction of the dictionary atoms does not only affect the response speed but also reduces accuracy.

Table 3 also suggests that in the first category, the first group has better results compared to the other two groups. Hence, the reduction of the dictionary atoms does not only affect the response speed but also reduces accuracy.

Study of the results indicates that the first group in the second category has better results compared to two other groups.

by compare the best results of both simulations in white, babble, car and factory noises suggests that although the first group of second simulation has better results in white noise, but according to the accuracy criterion, the first category of first simulation, whose dictionary was trained by NMF method, showed better results than the other methods, whose results have been used to be compared with other methods.

Thus, NMF method operates better than K-SVD method in separation. The major difference between these two methods is eliminating negative segments in the non-negative matrix factorization method, which results in better separation of speech and noise segments.

Table 1: Simulation Test bench Parameters.

Parameters	Description
Sampling	16 KHz. 16 bit
frame size	8ms
frame window	40ms
Δ	4
Test utterances	100 TIMIT sentences.
Types of noise	White, babble , car, factory
Noise levels	-10dB to 20dB

3.2 Evaluation of the Second Algorithm

Tables 4 and 5 outline the simulation conditions as well as the characteristics and parameters of simulated cases respectively.

In STRF space, the components of the representation are driven from a three-dimensional space in each frame. The sparse representation in this space is almost impossible due to high dimensionality of the space and lack of sufficient training samples. Therefore, as previously indicated, due to the huge data volume in STRF space, each temporal frame is partitioned into some sub-cubes.

The main issue in this simulation is analysis of the conditions of the cubes: overlapping, non-overlapping or removing the cubes, enlarging and downsizing the cubes, and increasing and decreasing the number of cubes influence the obtained results.

Therefore, three experiments are designed according to the above parameters.

Mesgarani has applied 1-32 as the range for rates and 0.5-8 for the scale [6]. In order to assess the impacts of changes related to rate and scale values in simulations, the mentioned ranges of values were evaluated. It can be said that the ranges of scale and rate affect the number of samples in a cube as well as the number of cubes within a time frame. The number of dictionaries is equal to the number of cubes within a time frame.

After feature extraction, the dictionary is built using Non-Negative Matrix Factorization approach (NMF) [22]

Partitioning the space into cubes leads to emergence of the following questions:

- How to select the appropriate size of the cubes? What is the effect of this size on the performance?
- Should the cubes be overlapped? How the results alter with overlapping cubes?
- Can we ignore some of the cubes? Can we employ a portion of the set of cubes rather than all of them?

The size of cubes was selected according to the selected range for rate and scale. as Also, in the first and second simulations, the size of cubes with the same or different segregation sizes on each dimension should be selected such that a time frame could be formed out of the sum of these cubes.

Therefore, the following tests are conducted to achieve the best performance of the system by addressing the answers of the above questions. In each test, the average percentage of correctly detected speech was considered as the measure of performance.

3.2.1 First experiment: Finding the proper size of the cubes

In this section, each time frame was divided into cubes of three different sizes.

In the first and second simulations, the scale and rate change from 0.25-4 cycles/octave and 2-16 Hz, respectively. Nevertheless, the size of the partition of the three dimensions of scale, rate, and frequency in the first case is different, while in the second case it is same.

Finally, in the third case, the rate and scale changes range from 0.5 to 8 cycles/octave and from 1 to 32 Hz, respectively.

The sizes of the partitions are different on the three dimensions.

In the first category, each cube has 168 samples and each time frame is divided into 96 cubes. Thus, 96 dictionaries with 1000 atoms for speech and noise are created for each time frame separately with NMF method. In the second category, each time frame consists of 32 cubes with each cube constituting 504 components.

Thus, 32 dictionaries of 1000 atoms are trained for speech and noise in each frame using the NMF method. In the third category, each cube consists of 792 samples with each time frame composed of 32 cubes. Therefore, 32 dictionaries with 1000 atoms are created for speech and noise at any time using NMF method.

Table 6 summarizes the average percentage of speech/noise detection rate in three cases of the first experiment over 100 utterances.

Studies show that the results of these three groups are not significantly different in the first experiment, though the second category has had better results than the two other groups. It suggests that the speech and noise boundaries are better separated by reducing the range of the scale and rate as well as increasing the size of the cubes, though the same size partition is also effective for the three dimensions. The cubes were partitioned in this space to simplify the calculations of four-dimensional -time spectral space. The larger the cube sizes, the higher the number of iterations of nested loops will be, thereby increasing computational complexity.

The STRF domain is a four-dimensional domain. Due to the complexity of four-dimensional computations, the domain was divided into cubes within a time frame. By increasing the number of samples within each cube, the number of nested loops in the program grows, lengthening the simulation.

Therefore, the time index was assumed fixed in every cube while the other three parameters of frequency, scale, and rate were altered.

If we consider F, R and S as the parameters for frequency, rate and scale respectively. NFq as the magnitude of samples in each cube:

$$NFq = R \times S \times F \quad (7)$$

Thus, a bigger cube will demand more calculations. If we consider the number of cubes as Q within a time frame, NFt is the value of iteration within a time frame, calculated in the following way:

$$NFt = \sum_{i=1}^{Q} NFq_i \quad (8)$$

Thus, more cubes within a time frame will require a longer time for computations. If we consider T as the time frame in a signal, NF is the number of iterations in a signal:

$$NF = \sum_{i=1}^{T} NFt_i \quad (9)$$

Hence, the size of cubes affects the complexity of the computations.

The computational complexity of the second and third cases per sample takes about 1.6 times and 3 times longer on average compared to the first category.

The complexities of the second and third cases are proportional to the number of dictionaries. The second case was less complex because of fewer samples in each cube. If the accuracy is the main concern, the overall accuracy of the second case has been slightly better than that of the two other categories, representing a trade-off between speed and accuracy.

3.2.2 Second Experiment: Cubes Overlapping

We conducted the experiments on three new cases using the second cube size of the previous experiment. In these three cases, the cubes overlap with each other within each frame.

Table 2: Performance of the proposed first VAD under four types of noise and different SNR values with NMF dictionary.

case	Number of atoms in dictionary	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE)							
		SNR / Noise	20	15	10	5	0	-5	-10
first	1000	**White**	**92.17%**	**92.10%**	**92.40%**	**90.42%**	**86.58%**	**81.06%**	**63.07%**
		Babble	**95.55%**	**90.29%**	**85.10%**	**81.58%**	**75.88%**	**69.64%**	**60.17%**
		Car	**95.79%**	**93.38%**	**91.81%**	**89.15%**	**81.00%**	**70.29%**	**65.3%**
		Factory	**93.72%**	**91.45%**	**87.31%**	**85.27%**	**80.62%**	**73.55%**	**64.93%**
second	500	**White**	92.11%	91.95%	92.15%	90.91%	85.22%	75.45%	65.96%
		Babble	90.23%	86.40%	84.89%	79.78%	72.62%	65.67%	57.00%
		Car	91.44%	91.50%	90.34%	85.15%	79.38%	65.43%	56.20%
		Factory	89.95%	85.22%	82.36%	80.68%	78.15%	67.41%	60.34%
third	100	**White**	91.43%	92.87%	93.33%	86.76%	84.07%	73.17%	67.21%
		Babble	91.58%	89.29%	86.85%	78.26%	72.64%	66.33%	62.21%
		Car	96.67%	89.92%	83.79%	78.08%	69.19%	61.63%	55.02%
		Factory	90.18%	87.38%	84.65%	79.43%	75.50%	68.98%	65.77%

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

case	Number of atoms in dictionary	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE)							
		SNR / Noise	20	15	10	5	0	-5	-10
first	1000	**White**	**95.82%**	**94.43%**	**91.26%**	**90.72%**	**83.58%**	**72.29%**	**60.45%**
		Babble	**94.25%**	**87.54%**	**81.60%**	**78.46%**	**68.18%**	**66.13%**	**54.68%**
		Car	**96.10%**	**92.14%**	**89.94%**	**81.34%**	**70.80%**	**61.24%**	**53.58%**
		Factory	**92.35%**	**89.23%**	**85.47%**	**80.72%**	**77.15%**	**69.59%**	**57.36%**
second	500	**White**	96.27%	95.88%	92.78%	90.16%	84.21%	74.96%	63.58%
		Babble	92.44%	85.67%	81.09%	76.05%	67.15%	58.74%	53.54%
		Car	95.30%	91.54%	88.37%	79.54%	68.61%	59.21%	52.32%
		Factory	90.69%	83.48%	80.56%	78.15%	71.37%	65.28%	58.92%
third	100	**White**	94.65%	93.98%	90.84%	86.94%	81.64%	71.94%	61.77%
		Babble	93.63%	86.36%	82.49%	76.70%	67.28%	60.51%	53.49%
		Car	95.54%	92.59%	89.48%	79.91%	70.42%	64.56%	57.61%
		Factory	90.15%	85.44%	84.32%	80.57%	73.68%	65.04%	58.92%

Table 4: Simulation Test bench Parameters.

Parameters	Description
Sampling	16 KHz. 16 bit
frame size	32 ms
Test utterances	100 TIMIT sentences.
Types of noise	White, babble , car, factory
Noise levels	-10dB to 20Db

Table 5: The characteristics and parameters of simulated cases.

Experiment	cubes partitioning (in a time frame)	Case	Number of samples of each cube	Number of small cubes of each time frame	Dimensions of speech and noise dictionaries	Range of scale (Cycle/Octave)	Range of rate(HZ)
1	**Considering all of cubes**	First	168	96	168×1000	0.25 - 4.00	2 – 16
		second	504	32	504×1000	0.25 - 4.00	2 – 16
		Third	792	32	792×1000	0.5 - 8.00	1 – 32
2	**Overlapping cubes**	First	280	96	280×1000	0.25 - 4.00	2 – 16
		second	504	62	504×1000	0.25 - 4.00	2 – 16
		Third	504	80	504×1000	0.25 - 4.00	2 – 16
3	**removing some cubes**	First	343	36	343×1000	0.5 - 8.00	1 – 32

Table 6: Performance of the proposed VAD under four types of noise and seven specific SNR values without overlapping.

case	Number of samples of each cube	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE)							
		SNR / Noise	20	15	10	5	0	-5	-10
first	**168**	**White**	91.61%	91.35%	91.49%	90.91%	87.77%	82.09%	82.92%
		Babble	91.37%	91.76%	90.96%	89.77%	83.73%	78.58%	69.46%
		Car	90.68%	91.29%	90.47%	89.44%	86.67%	77.48%	64.13%
		Factory	90.15%	88.41%	85.27%	85.97%	82.61%	80.02%	73.21%
second	**504**	**White**	**93.09%**	**92.45%**	**92.94%**	**89.34%**	**90.48%**	**90.93%**	**86.93%**
		Babble	**94.04%**	**93.37%**	**93.39%**	**95.29%**	**92.78%**	**84.57%**	**75.31%**
		Car	**93.45%**	**94.06%**	**93.85%**	**95.36%**	**96.11%**	**91.87%**	**82.50%**
		Factory	**92.54%**	**91.48%**	**91.62%**	**94.10%**	**93.31%**	**89.25%**	**80.96%**
third	**792**	**White**	92.20%	92.34%	93.27%	88.65%	85.00%	87.84%	86.20%
		Babble	93.34%	93.33%	93.14%	93.55%	90.43%	87.64%	81.20%
		Car	92.90%	92.54%	92.04%	89.50%	92.91%	89.83%	80.69%
		Factory	90.64%	91.01%	90.28%	92.82%	91.51%	88.34%	78.74%

In the first case, the overlapping occurs in three cubes. A total of 96 dictionaries were formed separately for speech and noise. The rate and scale ranges have been as same as the previous second case. The results are presented in Table 7.

The second case overlaps in four cubes. In the case, 62 dictionaries were made separately for speech and noise with 504 attributes for each atom, with Table 7 tabulating the results.

In the third case, the overlap is in five cubes, with each cube composed of 504 samples and each time frame covering 80 cubes. Therefore, 80 dictionaries were made for speech and noise with 1000 atoms.

The segregation sizes of the three dimensions of the scale as well as the rate and the frequency have been different in the first and third case. However, they have been the same in the second case.

Studies suggest that the results of the second case have been better than those of the other two cases. In addition, the increase in the number of overlapping cubes worsens the detection rate. Although overlapping the cubes at SNRs above zero results in better rates in babble noise, increasing

the number of overlapping cubes showed no significant effect overall.

On the other hand, if N is the number of overlapping cubes, as with Eq. 7, NF_{qov} is defined as the magnitude of samples in each cube with overlaps:

$$NF_{qov} = N \times R \times S \times F \quad (10)$$

$$NF_{tov} = \sum_{i=1}^{Q} NF_{qov_i} \quad (11)$$

$$NF_{ov} = \sum_{i=1}^{T} NF_{tov_i} \quad (12)$$

Therefore, similar to previous section, the size of cubes highly affects the complexity of the computations. The complexity of the algorithm is 25% and 50% greater than that of the first case for the second and third cases respectively.

3.2.3 Third Experiment: Removing some cubes

So far, when using small cubes overlapping, we have considered the whole data with redundancy in the time frame. We now consider only a number of cubes in making the final decision. However, in STRF space, the low frequency, low rate, and low scale portion of the space are considered to be more important [23]. Therefore, we considered low frequencies, low rates, and low scales cubes (36 cubes) with the size 7*7*7. Accordingly, 36 dictionaries were made for speech and noise with 343*1000 dimensions. In this test, the scale and rate ranged from 0.5 to 8.00 cycles/octave and 1 to 32 Hz, respectively. The results are reported in Table 8.

The results of the second case of the first experiment which has considered all cubes within one time frame, in addition to the first overlapping case of the second experiment demonstrate its superiority. Also, the first case of the third experiment which considers the low portion of the space has been compared in this assessment. According to the comparisons of best result of three experiments, we can conclude that the second case of the first experiment and the first case of the third experiment at beyond 0dB have a better performance with close outputs in relation to the first overlapping case of the second experiment.

The result is in contrast with SNRs below 0dB. Overall, the first case of the third experiment slightly outperforms other cases and will be compared with other competing VADs in the next section. Accordingly, better results were achieved by removing cubes away from the source and keeping the three-dimensional information of scale, rate, and frequency near the source.

3.3 Results Comparison

We proposed two algorithms for the speech detector. The first algorithm was proposed using sparse representation of two-dimensional auditory spectrogram space exploring different atomic sizes of dictionaries and two dictionary learning methods. It presented good results at high SNRs. On the other hand, the second algorithm, which has been more complicated than the first one, suggests a speech detector employing the sparse representation of the STRF four-dimensional space. Due to the large volume of STRF's four-dimensional space, this space was divided into sub-cubes for each time frame, with dictionaries trained based on the conditions of these cubes by NMF training algorithm.

Figs. 9, 10, 11 and 12 compare the best results of both algorithms in white, babble, and car noises, respectively.

Investigation of these three diagrams suggests that although the four-dimensional space of STRF has huge computational complexity, unlike the two-dimensional space which considers the overall frequency behavior, it has used local frequency behaviors, which yielded better characteristics and results.

Therefore, the first case of the third experiment in the second algorithm slightly outperforms other cases and will be compared with other competing VADs.

The superior performance of the proposed VAD is illustrated through nonspeech–speech error (NDS) and speech–nonspeech error (MSC) [24]. Noise detected as speech (NDS) is the proportion of nonspeech frames 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 against the proposed VAD algorithms, which have proved to be noise robust. They are LTSV [2], Sohn [3], G.729B [5], Mesgerani's VAD [6], Harmfreq [25], LTSD [26], LSFM [27], and LTPD [28].

In Fig. 13, the proposed VAD has even a lower error rate under zero SNRs compared to other VADs. Specifically, it should be noted that Mesgerani's VAD considers all noisy speeches as noise at low SNRs (i.e. practically it did not perform any classification). In contrast, our proposed STRF-sparse VAD successfully classified the noise and speech with a low error rate at even low SNRs.

Fig. 14 represents the comparison of our VAD against state-of-the-art VADs. Mesgerani and G729B degrade at low SNRs. At high SNRs, the proposed method error is also the same as Mesgerani's methods.

Totally, comparing the results of these VADs, it can be observed that the proposed VAD outperforms the other state-of-the-art methods compared here.

Because of the similarity of structures of babble noise and speech, at some SNRs, the proposed VAD NDS performance degrades in our method in contrast to MSC performance.

4. Conclusion

In this paper, two new VAD algorithms were proposed based on sparse representation in spectro-temporal domain.

The simulation results indicated that the results of considering total cubes or removing some of them within a time frame are similar and at SNRs below zero, overlapping cubes perform better. However, if we consider the computational complexity, in general, removing some cubes is a better tradeoff than the other two experiments.

The simulation results suggested that our second proposed VAD algorithm is effective in low SNR situations. Our future work focuses on developing a low calculation complexity version of the algorithm to be suitable for real-time processing.

Table 7: Performance of the proposed VAD under four types of noise and seven specific SNR values with overlapping.

case	Number of samples of each cube	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE)							
		SNR / noise	20	15	10	5	0	-5	-10
first	280	White	91.94%	89.26%	88.38%	86.91%	87.80%	85.78%	84.56%
		Babble	88.78%	87.33%	87.16%	86.84%	85.55%	81.23%	77.34%
		Car	91.68%	91.06%	90.74%	89.32%	88.83%	83.47%	80.62%
		Factory	90.27%	88.41%	86.98%	87.55%	86.39%	83.61%	79.01%
second	504	White	**91.76%**	**91.80%**	**91.86%**	**89.15%**	**86.26%**	**86.42%**	**86.55%**
		Babble	**91.47%**	**91.27%**	**91.27%**	**90.48%**	**89.85%**	**80.72%**	**75.47%**
		Car	**93.30%**	**93.70%**	**93.96%**	**90.50%**	**84.61%**	**83.72%**	**81.84%**
		Factory	**91.55%**	**91.64%**	**91.71%**	**91.07%**	**88.35%**	**82.15%**	**78.92%**
third	504	White	86.88%	87.42%	87.64%	89.71%	91.89%	88.92%	86.79%
		Babble	90.64%	90.73%	90.89%	87.24%	80.68%	78.69%	76.05%
		Car	87.43%	85.63%	84.92%	84.17%	83.32%	81.52%	80.19%
		Factory	89.33%	90.12%	90.26%	88.41%	82.09%	79.85%	75.76%

Table 8: Performance of the proposed First case VAD under four types of noise and seven specific SNR values after removing some cubes.

case	Number of samples of each cube	MEAN VAD ACCURACY (PERCENTAGE TRUE POSITIVE)							
		SNR / noise	20	15	10	5	0	-5	-10
first	343	White	**94.59%**	**94.82%**	**95.53%**	**94.19%**	**91.42%**	**90.11%**	**87.35%**
		Babble	**94.44%**	**93.84%**	**93.12%**	**89.86%**	**89.20%**	**88.43%**	**76.84%**
		Car	**95.56%**	**96.04%**	**94.41%**	**94.96%**	**95.24%**	**91.75%**	**82.23%**
		Factory	**94.98%**	**94.87%**	**94.32%**	**93.51%**	**92.11%**	**91.24%**	**88.69%**

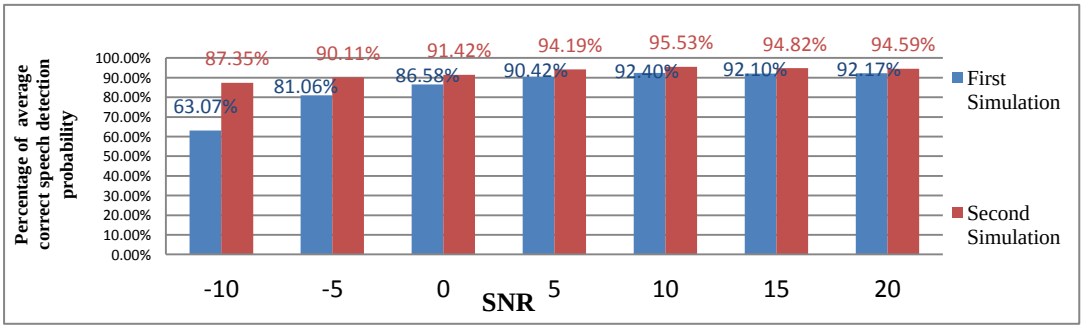

Fig. 9 Comparison of the best results of both simulations in white noise.

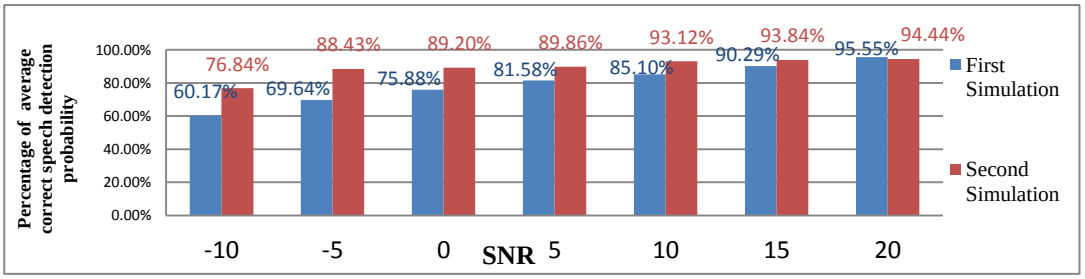

Fig. 10 Comparison of the best results of both simulations in babble noise.

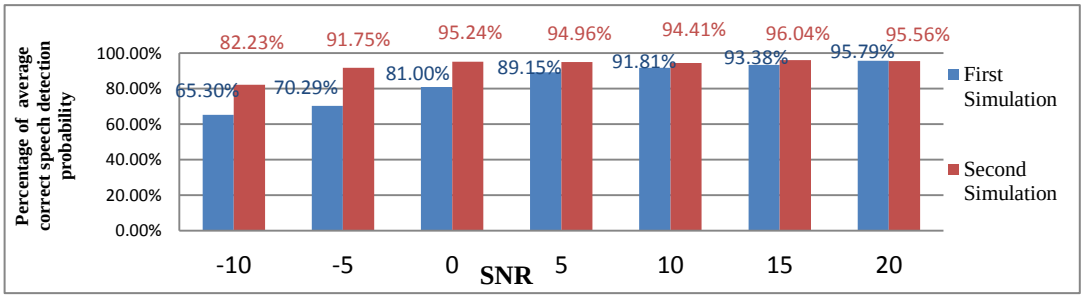

Fig. 11 Comparison of the best results of both simulations in car noise.

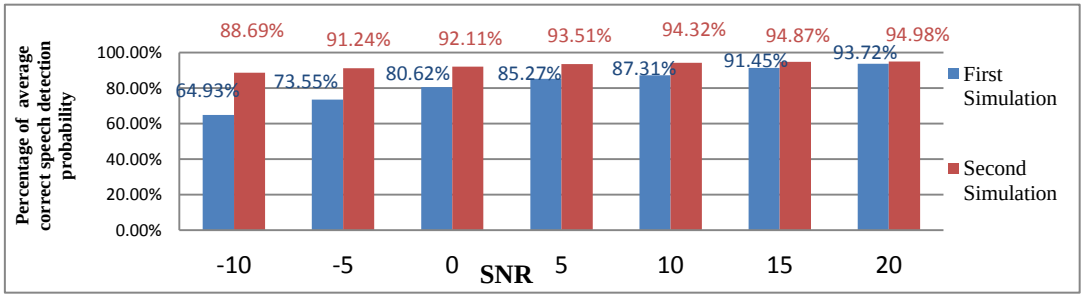

Fig. 12 Comparison of the best results of both simulations in factory noise.

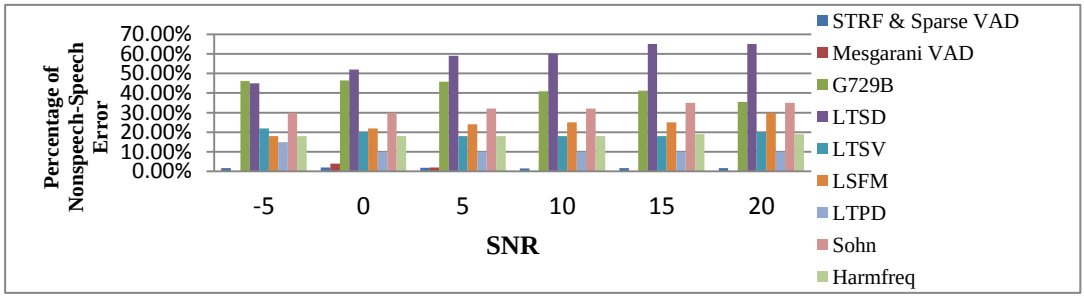

Fig. 13 Nonspeech–speech Error(NDS).

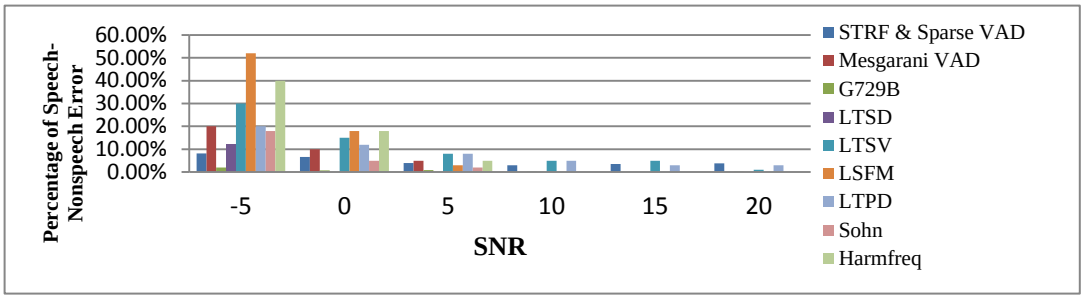

Fig. 14. Speech–nonspeech Error(MSC).

References

[1] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. L. Hansen, A. Janin, B. S. Lee, Y. Lei, V. Mitra, N. Morgan, S.O. Sadjadi, T.J. Tsai, N. Scheffer, L.N. Tan and B. Williams, "All for one: feature combination for highly channel-degraded speech activity detection," *in ISCA Interspeech*, pp.709–713, 2013.

[2] M. Eshaghi, and M. R. Karami Mollaei, "Voice activity detection based on using wavelet packet," *in Digital Signal Processing*, vol. 20, No. 4, pp. 1102-1115, 2010.

[3] Y. Datao, H. Jiqing, Z. Guibin and Z. Tieran, "Sparse power spectrum based robust voice activity detector," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, pp. 289-292, 2012

[٤] W Hongzhi, X Yuchao and L Meijing, "Study on the MFCC similarity-based voice activity detection algorithm," in *International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, Dengleng, 2011.

[5] S.G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.11, No.7, pp. 674-693,1989.

[6] Nima Mesgarani, and Shihab Shamma, "Denoising in the Domain of Spectro-temporal Modulations", *in EURASIP Journal on Audio, Speech, and Music Processing* , ID. 42357, 8 pages, doi:10.1155/2007/42357 ,2007.

[7] Weifeng Li, Yicong Zhou, Norman Poh, Fei Zhou, and Qingmin Liao, "Feature Denoising Using Joint Sparse Representation for In-car Speech Recognition", *in IEEE Transactions on audio, speech, and language processing*, vol.20, No.7, pp. 681-684, 2013.

[8] N. Mesgarani, S. David, and S.A. Shamma, "Representation of phoneme in primary auditory cortex: how the brain analyzes speech," *in IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Hawaii, April, 2007.

[9] Majid Mirbagheri, Nima Mesgarani, and Shihab Shamma, "Nonlinear filtering of spectrotemporal modulation in speech enhancement," *in IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 5478-81, 2010.

[10] C. Kim, K. Kumar and R. M. Stern, "Binaural sound source seperation motivated by auditory processing," *in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP2011*, Prague, 2011.

[11] C. Mart´ınez, J. Goddardb, D. Milone, and H. Rufiner, " Bio inspired sparse spectro-temporal representation of speech for robust classification,"*in Computer Speech and Language*, vol.26, No.5, pp. 336-345, 2012.

[12] Jort Florent Gemmeke, Hugo Van Hamme, Bert Cranen, and Lou Boves, "Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition", *in IEEE Journal of selected topics in signal processing*, vol.4,No.2, pp. 272-287, 2010.

[13] B. K. Natarajan, "Sparse approximate solutions to linear systems," *in Society for Industrial and Applied Mathematics (SIAM J). Computer*, vol.24,No.2, pp.227–234, 1995.

[14] Mohadese Eshaghi, Farbod Razzazi, and Alireza Behrad, "A voice activity detection algorithm in spectro-temporal domain using sparse representation," *in International Journal of Machine Learning and Cybernetics*, 2018, DOI: 10.1007/s13042-018-0856-z.

[15] G. Hosein Mohimani, Massoud Babaie-Zadeh, and Christian Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed L0 norm," *in IEEE Transactions on Signal Processing*, vol.57, No.1, pp.289-301, 2009.

[16] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee, and T.J. Sejnowski, "Dictionary learning algorithms for sparse representation," *in Neural Computer*, vol.15, No.2, pp.349–396,2003.

[17] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *in Journal of Machine Learning Research*, vol.5, No. 9, pp.1457–1469, 2004.

[18] R. Zdunek, and A. Cichocki, "Non-negative matrix factorization with quadratic programming," *in Neurocomputing*, vol.71, No.10-12, pp. 2309-2320, 2007

[19] M. Aharon, M. Elad, A. Bruckstein, "K-SVD: An algorithm for designing over complete dictionaries for sparse representation," *in IEEE Transactions on Signal Processing*, vol.54, No.11, pp.4311–4322, 2006.

[20] W. M. Fisher, G. R. Doddington, M. Goudie and M. Kathleen, "The DARPA speech recognition research database: specifications and status", *in DARPA Workshop on Speech Recognition*, 1986.

[21] 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.

[22] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh, "Non-negative matrix factorization based compensation of music for

automatic speech recognition," *International Conference on Speech and Language Processing*, 2010.

[23] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," *in IEEE International Conference on Acoustics, Speech and Signal Processing*, vol4, No.1, pp.601–604, 2004.

[24] Ian Vince McLoughlin, "Super-Audible Voice Activity Detection," *in IEEE Transactions on Speech and Audio Processing*, vol. 22, No.9, pp.1424-1433, 2014.

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

[26] J. Ramirez, J.C. Segura, C. Benitez, A. de la Torre and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *in Speech Communication*, vol.42, No.3-4, pp.271–287, 2004

[27] M. Yanna and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *in EURASIP Journal on Audio, Speech and Music Processing*, 2013, DOI: 10.1186/1687-4722-2013-21.

[28] Xu-Kui Yang, Liang He, Dan Qu1 and Wei-Qiang Zhang, "Voice activity detection algorithm based on long-term pitch information," *in EURASIP Journal on Audio, Speech, and Music Processing*, 2016, DOI: 10.1186/s13636-016-0092-y.

Mohadese Eshaghi received her B.Sc. and M.Sc. in Electrical Engineering from Mazandaran University in 2005 and 2008, respectively. She received her Ph.D. from Department of Electrical and Computer Engineering, Islamic Azad University, Science and Research Branch in 2019 in Electrical Engineering. Her research interests include Speech enhancement, Image processing, Pattern recognition methods and Data mining.

Farbod Razzazi received his B.Sc. and M.Sc. in Electrical Engineering from Sharif University of Technology in 1994 and 1996, respectively. He received his Ph.D. from Amirkabir University of Technology (Tehran Polytechnic) in 2003 in Electrical Engineering. He is currently an Associate Professor in Department of Electrical and Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran. His current research interests are signal forensics and anti-forensics, pattern recognition methods and their applications in statistical signal processing systems.

Alireza Behrad received the B.Sc. degree in Electrical Engineering from Tabriz University, Tabriz, Iran, in 1995. In 1998, he received the M.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran. He received the Ph.D. degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 2004. Currently, he is an associate professor of Engineering Faculty, Shahed University, Tehran, Iran. His research fields are image and video processing, machine vision and digital multimedia authentication.