A Gaussian Clustering Based Voice Activity Detector for Noisy Environments Using Spectro-Temporal Domain

Sara Valipour Valipour_Sara@yahoo.com
Faculty of Engineering Islamic Azad University Ghaemshahr Branch
Ghaemshahr, Iran

Farbod Razzazi razzazi @ srbiau.ac.ir
Faculty of Engineering Islamic Azad University, Science and Research Branch
Tehran, Iran

Azim Fard azimfard@cra.ir
Communications Regulatory Authority
Tehran, Iran

Nafiseh Esfandian Na_Esfandian@yahoo.coo
Faculty of Engineering Islamic Azad University Ghaemshahr Branch
Ghaemshahr, Iran

Abstract

In this paper, a voice activity detector is proposed on the basis of Gaussian modeling of noise in the spectro-temporal space. Spectro-temporal space is obtained from auditory cortical processing. The auditory model that offers a multi-dimensional picture of the sound includes two stages: the initial stage is a model of inner ear and the second stage is the auditory central cortical modeling in the brain. In this paper, the speech noise in this picture has been modeled by a 3-D mono Gaussian cluster. At the start of suggested VAD process, the noise is modeled by a Gaussian shaped cluster. The average noise behavior is obtained in different spectrotemporal space in various points for each frame. In the stage of separation of speech from noise, the criterion is the difference between the average noise behavior and the speech signal amplitude in spectrotemporal domain. This was measured for each frame and was used as the criterion of classification. Using Noisex92, this method is tested in different noise models such as White, exhibition, Street, Office and Train noises. The results are compared to both auditory model and multifeature method. It is observed that the performance of this method in low signal-to-noise ratios (SNRs) conditions is better than other current methods.

Keywords: Voice activity detector, Spectro-temporal Domain, Gaussian modeling, Auditory model.

1. INTRODUCTION

In general, sound signal is composed of two parts, speech and non-speech. The latter is either silence or background noise. Accordingly, detection of speech signal from non-speech signal, known as voice activity detection (VAD), is one of the most important issues in the speech processing systems. In particular, the complexities increases in low SNRs where there are challenging in VAD design. One of the applications of VAD is in speech enhancement systems [1], where VAD is used to estimate noise characteristics from the silence parts of the signal. Robust speech recognition [2], speech coding [3] and echo cancellation are among the other applications of VAD.

The first, but of course the most, usual VAD algorithm has been presented in [4]. There are other VAD algorithms as well. In [5], a VAD has presented on the basis of MFCC features and SVM, as MFCC coefficients provide good information of the speech signal. Sohn in [6] has used a Gaussian statistical model for VAD. As another work, has obtained a VAD based on Taylor series [7]. Chang has performed a class from VAD algorithm using different statistical models. Moreover, he has combined Gaussian model, complex laplacian and gamma probability density equations to analytically characterize statistical properties of noise and speech parts [8]. Another VAD has been obtained based on the generalized Gamma distribution [9], where a distribution of noise spectra and noisy speech spectra has been obtained based on inactive speech intervals. In all these algorithms, the results are not promising in low signal to noise ratios (SNR) and VAD performance in low SNRs has remained as a challenging issue.
In this research, our proposed VAD algorithm is based on spectro-temporal representation of the speech. The idea is based on neuro-physiological and psycho acoustical investigations in various stages of auditory system. This model consists of two main stages. The first one is the auditory system which represents the acoustic signal as an auditory spectrogram. The second stage, which is the central cortical stage, is the stage of analyzing the spectrogram by using a set of 2D filters. The new successful achievements in the spectro-temporal studies reveal a significant improvement of the performance for enhancement systems [1], Speech Recognition [10] and also robust pitch extraction [11].
In this work, a VAD algorithm is proposed on the basis of noise Gaussian model in the spectro-temporal domain. Evaluating the efficiency, it is shown that the spectro-temporal domain is a suitable space for this separation. The rest of the paper is organized as follows. In section 2, the auditory model and spectro-temporal model are briefly reviewed. In section 3, the proposed VAD method is presented and analyzed in the spectro-temporal domain. In section 4, the method is evaluated and compared to other methods. Finally, the paper is briefly concluded in Section 5.

2. SPECTRO-TEMPORAL MODEL

Auditory Model

The auditory model has been obtained on the basis of neuroscience and biology researches. They are achieved based on two different sections of the auditory systems, mammals and, in particular, humans [12]. The model has two main parts [13-14]. In the first part of this model, the acoustic signal is represented by an auditory spectrogram [14]. While in the next part, the auditory spectrogram is analyzed using a set of 2D filters [15].

calculation of auditory spectrogram

In first part of auditory model, the auditory spectrogram of input signal is calculated by passing various through stages. The stages are shown in figure 1 [12-15].
As shown in figure 1, the input signal, enters a hair cell stage after passing through a filter bank which makes a 2D representation of the input signal. This part consists of three stages: a high-pass time domain filter, a nonlinear compression stage, and a low-pass time domain filter. The output of this stage is applied to a lateral inhibitory network which is in fact a first-order frequency domain derivative, followed by a half-wave rectifier and a low pass time domain integrator. At the

final stage, the auditory spectrogram of the signal is calculated. The analytical characterization of sequential stages for the first section of the auditory model is given as follow [1].

$$y_{coch} = s(t) *_t h(t,x) \quad (1)$$
$$y_{AN}(t,x) = g((\partial y_{coch}(t,x)/\partial t) *_t w(t)) \quad (2)$$
$$y_{LIN}(t,x) = Max((\partial y_{AN}(t,x)/\partial x), 0) \quad (3)$$
$$y_{final}(t,x) = y_{LIN}(t,x) *_t \mu(t,x) \quad (4)$$

In the above relations, the operator $*_t$ shows the convolution in time domain.

The central auditory section

In this section, the auditory spectrogram is analyzed to extract the spectro-temporal features[16]. The signal is applied through a bank of 2-D filters. The contents of spectro-temporal modulation of the auditory spectrogram are determined using selected filter banks, centered along tonotopic axis [17]. The spectrotemporal impulse response of these filters is called the spectro-temporal response field (STRF). STRFs are 2-D Gabor wavelets.

$$STRF_{+} = \Re\{H_{rate}(t;\omega,\theta).H_{scale}(f;\Omega,\varphi)\} \quad (5)$$
$$STRF_{-} = \Re\{H^{*}_{rate}(t;\omega,\theta).H_{scale}(f;\Omega,\phi)\} \quad (6)$$

where $\Re$ is the real part,* is complex conjugate, ω is speed and Ω is scale. θ and φ are a phase of asymmetry along time and frequency domain respectively. In addition, H_{rate} andH_{scale} may be analytically extracted from h_{rate} and h_{scale} [1].

$$H_{rate}(t;\omega,\theta) = h_{rate}(t;\omega,\theta) + j\hat{h}_{rate}(t;\omega,\theta) \quad (7)$$
$$H_{scale}(f;\Omega,\phi) = h_{scale}(f;\Omega,\phi) + j\hat{h}_{scale}(f;\Omega,\phi) \quad (8)$$

where ^ shows the Hilbert transformation. h_{rate} and h_{scale} are respectively the temporal and spectral impulse responses [1].

$$h_{rate}(t;\omega,\theta) = h_r(t;\omega)cos\theta + \hat{h}_r(t;\omega)sin\theta \quad (9)$$
$$h_{scale}(f;\Omega,\phi) = h_s(f;\Omega)cos\phi + \hat{h}_s(f;\Omega)sin\phi \quad (10)$$

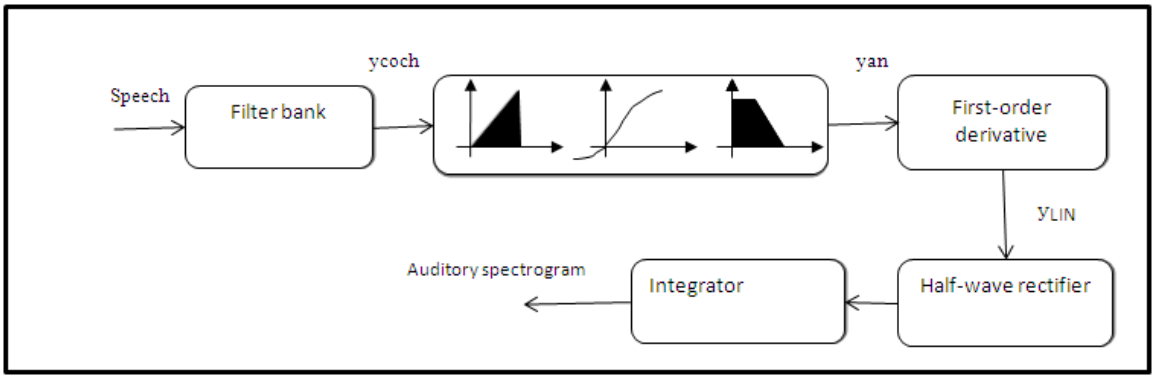

FIGURE 1: Different stages of first part of the auditory model

The impulse responses are obtained as hereunder for various frequencies and times [1]

$$h_r(t;\omega) = \omega h_s(\omega t) \quad (11)$$
$$h_s(f;\Omega) = \Omega h_s(\Omega f) \quad (12)$$

The auditory spectrum, after passing through STRFs is transformed into a 4-D cortical picture. These four dimensions are frequency, time, speed and scale [1].

$$r_+(t, f, \omega, \Omega, \theta, \phi) = y(t, f) *_{t,f} STRF_+(t, f, \omega, \Omega, \theta, \phi) \quad (13)$$

$$r_-(t, f, \omega, \Omega, \theta, \phi) = y(t, f) *_{t,f} STRF_-(t, f, \omega, \Omega, \theta, \phi) \quad (14)$$

where $*_{t,f}$ is the 2D convolution with respect to time and frequency. r_+ and r- are respectively the spectro-temporal responses of downward (+) and upward (-) STRFs. The wavelet transformation is obtained from the filters hrate and hscale as below:

$$h_{rW}(t; \omega) = h_r(t; \omega) + j\tilde{h}_r(t; \omega) \quad (15)$$

$$h_{sW}(f; \Omega) = h_s(f; \Omega) + j\tilde{h}_s(f; \Omega) \quad (16)$$

The complex response of downward and upward selective filters is as follows:

$$Z_+(t, f; \Omega, \omega) = y(t, f) *_{t,f} [h^*_{rW}(t; \omega) h_{sW}(f; \Omega)] \quad (17)$$

$$Z_-(t, f; \Omega, \omega) = y(t, f) *_{t,f} [h_{rW}(t; \omega) h_{sW}(f; \Omega)] \quad (18)$$

Finally, for each speech frame, two 3-D complex valued matrices are obtained for downward and upward filters respectively.

3. VAD METHOD IN SPECTRO-TEMPORAL DOMAIN

In the proposed VAD method, a Gaussian model is applied to model the noise cluster shape in spectro-temporal domain. In this method, it is attempted to estimate the cluster shape of 3-D spectro-temporal representation of noise (silent) using a Gaussian function. The concept originates from the fact that the shape of the noise cluster, created by large amplitude points in spectro-temporal domain is similar to a 3-D Gaussian hyper-surface. Therefore, the parameters of this function should be corresponded to the average of spectro-temporal representation of noise frames. The block diagram of such the noise modeling is shown in figure 2.
As shown in figure 2, the cortical picture of input noise is calculated for each frame with three dimensions of frequency, speed and scale. It is divided into two separate downward and upward representations. In the proposed method, in order to model noise samples in spectro-temporal space, we calculate the parameters of Gaussian model for downward and upward magnitude, separately.

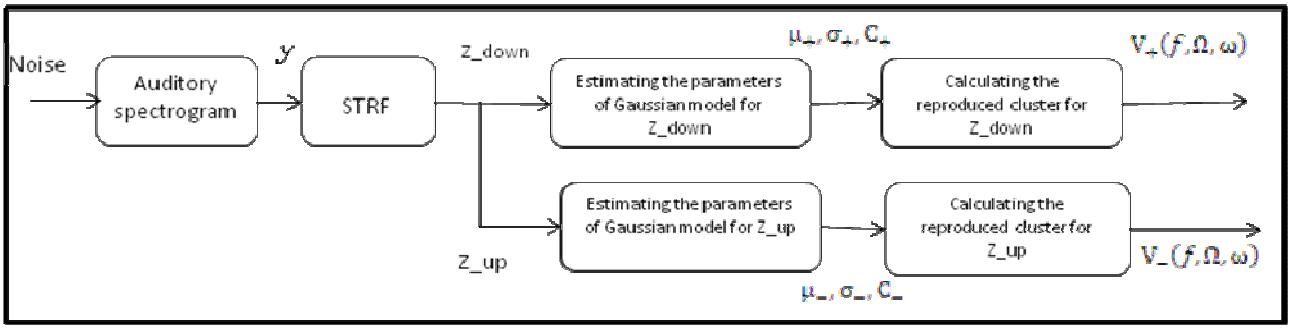

FIGURE 2: Noise Gaussian modeling in the spectro-temporal domain

The parameters include mean, covariance and gain respectively:

$$\mu_{\pm} = \frac{\sum_{t,f,\Omega,\omega} \left| Z_{\pm}(t,f,\Omega,\omega) \right| \begin{pmatrix} f \\ \Omega \\ \omega \end{pmatrix}}{\sum_{t,f,\Omega,\omega} \left| Z_{\pm}(t,f,\Omega,\omega) \right|} \tag{19}$$

$$\sigma_{\pm} = \frac{\sum_{t,f,\Omega,\omega} \left| Z_{\pm}(t,f,\Omega,\omega) \right| \left(\begin{pmatrix} f \\ \Omega \\ \omega \end{pmatrix} - \mu_{\pm} \right) \left(\begin{pmatrix} f \\ \Omega \\ \omega \end{pmatrix} - \mu_{\pm}^{T} \right)}{\sum_{t,f,\Omega,\omega} \left| Z_{\pm}(t,f,\Omega,\omega) \right|} \tag{20}$$

$$C_{\pm} = \frac{\sum_{t,f,\Omega,\omega} \left| Z_{\pm}(t,f,\Omega,\omega) \right|}{T} \tag{21}$$

After estimating the parameters of the Gaussian model for the noise, the reproduced cluster demonstrates the average behavior of the noise in sampling points of the spectro-temporal space for each frame. The reproduced cluster may be formulated as:

$$V_{\pm}(f,\Omega,\omega) = \frac{C_{\pm}}{(2\pi)^{3/2} \left| \sigma_{+} \right|} \cdot e^{-\frac{1}{2} \left(\begin{pmatrix} f \\ \Omega \\ \omega \end{pmatrix} - \mu_{\pm} \right)^{T} \cdot \sigma_{\pm}^{-1} \left(\begin{pmatrix} f \\ \Omega \\ \omega \end{pmatrix} - \mu_{\pm} \right)} \tag{22}$$

The distance of each frame of input signals with this pattern represents the similarity measure of the frame behavior to the noise. Therefore, a distance measure is proposed to calculate the similarity of the modeled cluster and the input frame. For each frame, after spectro-temporal representation, the magnitude of downward and upward representation is calculated:

$$P_{t\pm}(f,\Omega,\omega) = \left| Z_{\pm}(t,f,\Omega,\omega) \right| \tag{23}$$

Our tests have shown that the phase of cortical space is not an acceptable criterion for determining speech and noise sections. Therefore, only the magnitude section of this signal has been used. The distance measure of the input frame and the modeled cluster is proposed as:

$$D_{\pm} = \frac{1}{N} \cdot \frac{\sum_{f,\Omega,\omega} \left(V_{\pm}(f,\Omega,\omega) \cdot \left| P_{t\pm}(f,\Omega,\omega) - V_{\pm}(f,\Omega,\omega) \right| \right)}{\sum_{f,\Omega,\omega} V_{\pm}(f,\Omega,\omega)} \tag{24}$$

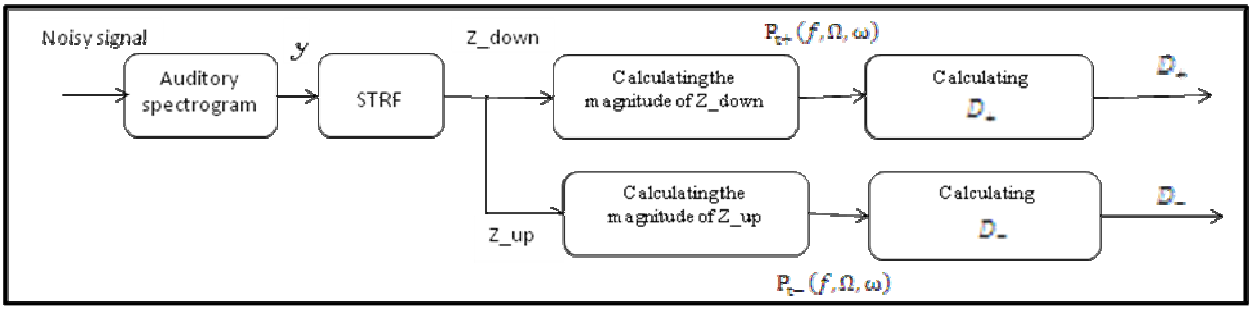

FIGURE 3: Decision making procedure for a frame in the proposed VAD system

In fact, the above relation is the weighted mean of two 3-D hyper-surfaces resulting from present frame and the average statistical behavior of the noise. The weight of this averaging has been determined in such a way that full-energy points would be more effective in this averaging. In figure 3, the block diagram shows the decision making procedure for a frame in the proposed VAD system.
Our VAD method is based on thresholding the resulted difference with an empirically set threshold.

$$VAD = \begin{cases} 0 & D_{\pm} < Th \\ 1 & D_{\pm} > Th \end{cases} \quad (25)$$

Determining the suitable threshold has been performed by testing in various noisy conditions which optimization results are given in section 4.

4. TESTS AND RESULTS

Evaluation framework
In the conducted tests, the speech signals are sampled in 16 KHz sampling frequency, 16 bit resolution. The length of each frame was assumed to be 4 ms. To build noisy signals; we took noises from NOISEX92 database [18] and added them to the clean signal. NOISEX database includes airport, babble, car, exhibition, office, restaurant, train, subway, street and white noises. Exhibition noise as representative of the human uproar, street noise representing open space, office noise representing office environment, Train noise representing industrial environment and white noise as the worst noise were selected. In addition, the noise was added to the clean signal in different SNRs with amounts -15, -10, -5, 0, 5, 10, 15, 20, 30, 40 dbs.
The proposed VAD system accuracy was measured by PS2S and PN2N probabilities. The measured parameters are defined as:

$$P_{S2S} = \frac{N_{S2S}}{N_S} \cdot 100 \quad (26)$$

$$P_{N2N} = \frac{N_{N2N}}{N_N} 100 \quad (27)$$

In the above relation, P_{S2S} is the probability of correct classification of speech frames in percents and P_{N2N} is the probability of correct classification of silence frames in percents. also, NS is the total number of speech frames, N_N is the total number of silence frames, N_{S2S}, is the number of correctly classified speech frames and N_{N2N} is number of correctly classified silence frames.

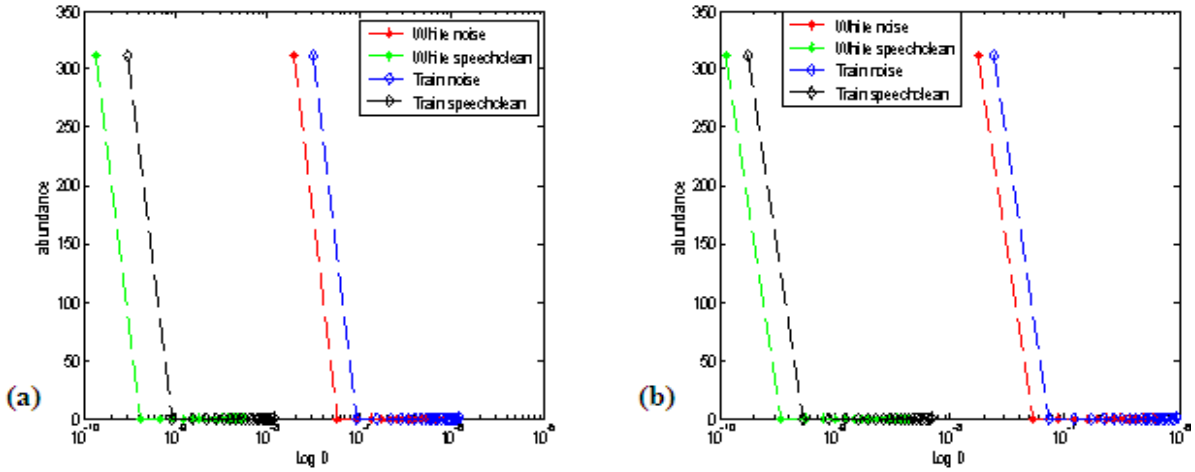

FIGURE 4: The histogram of noise behavior and speech. (a): The histogram of upward magnitude of noise behavior and speech. (b): The histogram of downward magnitude of noise behavior and speech

Histogram test

In figure (4-a) and (4-b), the histogram of upward magnitude of noise behavior and speech, and downward magnitude of noise behavior and speech have been shown in white and train noises respectively. The aim of this test is to show that Gaussian model is suitable for the noise. As seen in figures (4-a) and (4-b), we have shown that noise and speech have completely a separate behavior on the D axis.

Effect of different noises on proposed VAD

In the next experiment, the behavior of proposed VAD system was studied in different noises environments and various SNRs. In tables 1 and 2, the trend of changes in speech and non speech signal classification rates is given for various SNRs and in 5 different types of noise. As it is seen in table 1, street noise had a better behavior comparing to other noises. The system is well behaved in white, street and train noises in zero SNR. In addition, exhibition noise had worse behavior comparing to other 4 noises. In table 2, it may be observed that the classification rate of non-speech signals is equal to 100 percent for all noises and all SNRs.

Also, the figures (5-a) and (5-b) show the effect of various noises in various SNRs on a correct percentage of downward and upward magnitude of speech signal respectively. As it can be observed in both figures, a correct percentage of upward and downward magnitude of speech signal is in the Exhibition noise in zero SNR is around 59 and 69 percent and in SNR -5, is around 0 and 3 percents respectively. This may be explained by the fact that exhibition noise is human uproar and it provided the worst behavior comparing to other noises. As it may be observed in the figure, Street noise provides a better behavior comparing to other noises. Actually, this noise is produced by cars and is typically independent on the speech signal. Therefore, it is easily separable from speech signal in spectro-temporal domain.

SNR	z	PS2S White	PS2S Exhibition	PS2S Street	PS2S Train	PS2S Office
-15	Z_up	24.88	0	51.17	0	28.64
	Z_down	5.16	0	58.68	0	32.86
-10	Z_up	53.52	0	56.81	51.17	38.03
	Z_down	69.01	0	73.71	3.28	38.02
-5	Z_up	59.15	0	78.87	57.74	52.11
	Z_down	92.96	3.28	91.55	82.63	64.79
0	Z_up	92.96	54.93	93.90	91.08	58.69
	Z_down	95.77	62.44	94.37	93.90	91.55
5	Z_up	97.18	68.07	96.24	94.8	92.02
	Z_down	97.18	91.55	95.77	95.31	94.37
10	Z_up	100	92.49	99.53	96.71	95.30
	Z_down	99.53	94.37	97.65	97.18	96.24
15	Z_up	100	95.77	100	100	97.65
	Z_down	100	96.24	99.06	98.59	98.12
20	Z_up	100	98.12	100	100	100
	Z_down	100	98.12	100	100	100
30	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
40	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100

TABLE 1: Speech signal diagnosis correctness percentage in different noises and various SNRs.

SNR	z	PN2N White	PN2N Exhibition	PN2N Street	PN2N Train	PN2N Office
-15	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
-10	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
-5	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
0	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
5	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
10	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
15	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
20	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
30	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100
40	Z_up	100	100	100	100	100
	Z_down	100	100	100	100	100

TABLE 2: Non- Speech signal diagnosis correctness percentage in different noises and various SNRs.

The figures (6-a) and (6-b) show the effect of various noises in different SNRs on a correctly classified non-speech signals using downward and upward magnitude for all 5 noises and all tested SNRs is equal to 100 percents.

Comparison of proposed VAD system behavior with other methods

In this section, the proposed VAD was compared to auditory model [19] and multifeature method [20]. The results of the three systems were shown in figures (7-a) and (7-b).In fact, the obtained results have been reported in white noise on other systems, therefore the systems are compared in these situations.

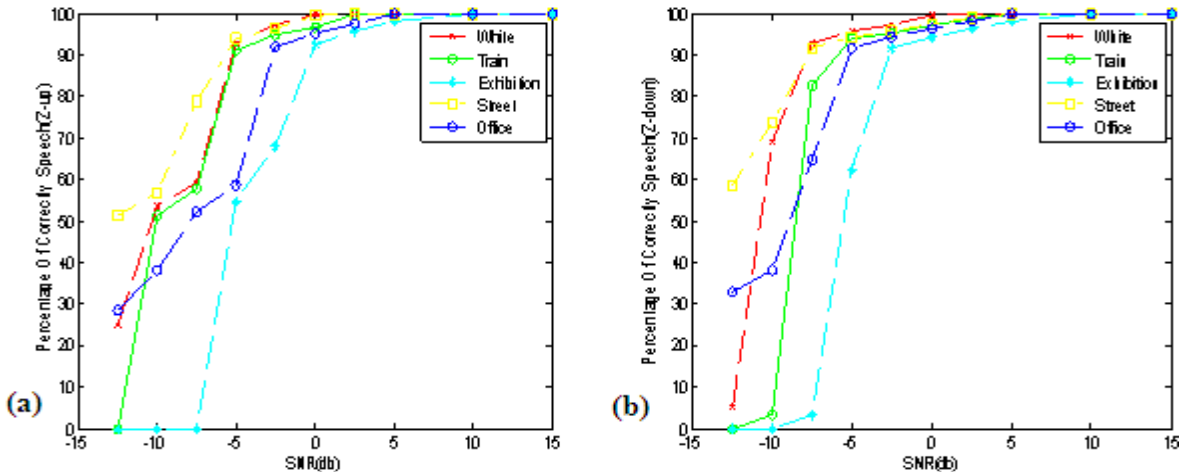

FIGURE 5: Effect of various noises in various SNRs on a correct percentage of magnitude of speech signal. (a): Effect of various noises in various SNRs on a correct percentage of upward magnitude of speech signal. (b): Effect of various noises in various SNRs on a correct percentage of downward magnitude of speech signal.

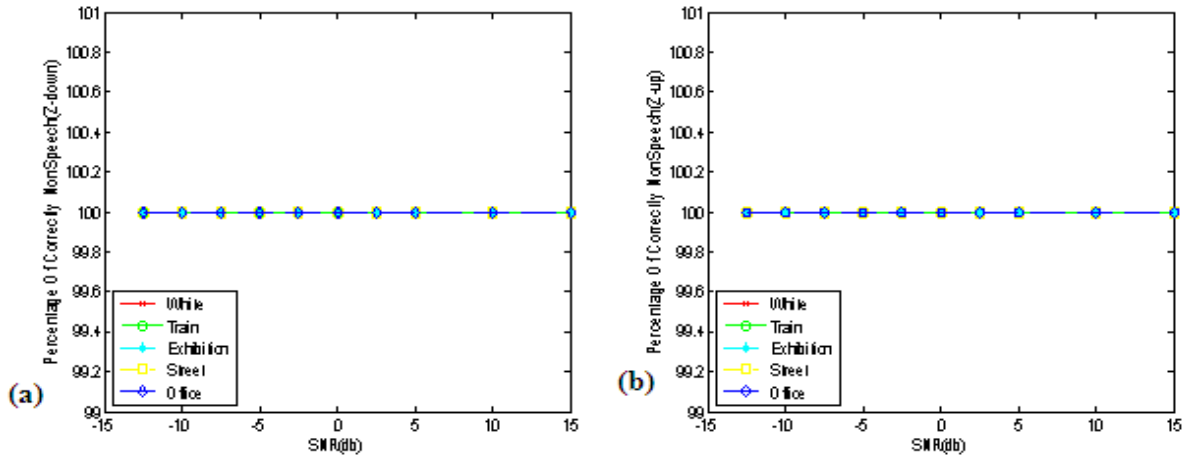

FIGURE 6: Effect of various noises in various SNRs on a correct percentage of magnitude of non-speech signal. (a): Effect of various noises in various SNRs on a correct percentage of upward magnitude of non-speech signal. (b): Effect of various noises in various SNRs on a correct percentage of downward magnitude of non-speech signal.

In figure (7-a) it can be observed that the proposed method had a much better behavior comparing to multi-feature method. However in comparison with auditory model, it is observed that the proposed method performance was better in low SNRs. In addition, figure (7-b) which is the effect of white noise on the correctness percentage of non-speech signal, comparing three systems show that the proposed VAD had a good behavior.

Behavior of correctness change with change in threshold

It is worthy to note that with a very subtle change in threshold, the rate of speech and non-speech signal classification is reduced. In figure (8-a) and (8-b) the classification rate variations in z-up and z-down versus threshold has been analyzed respectively. As seen in the figures (8-a) and (8-b), by increasing the threshold PS2N decreases and PN2S increases.

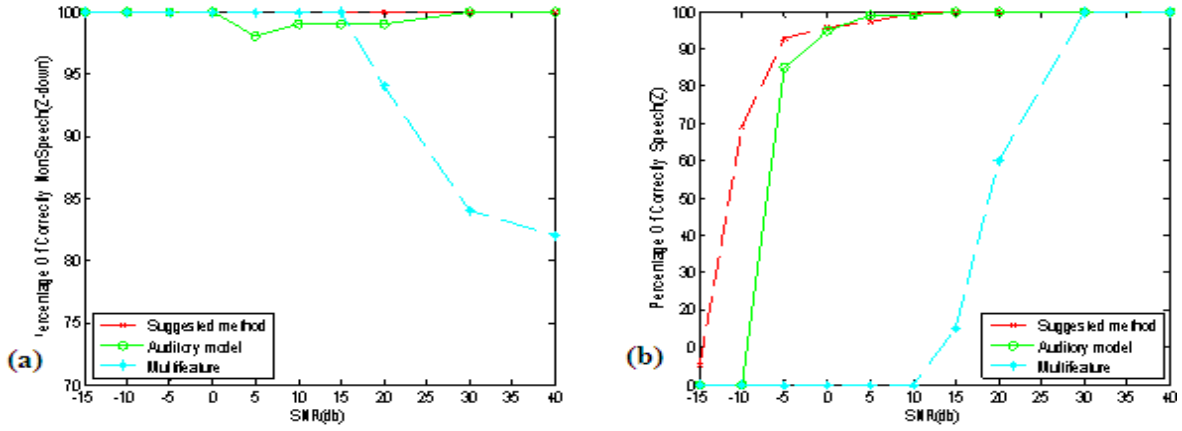

FIGURE 7: Effect of white noise on the correctness percentage for suggested method, auditory model [17] and multifeature method [18]. (a): Effect of white noise on the correctness percentage of speech signal. (b): Effect of white noise on the correctness percentage of non-speech signal

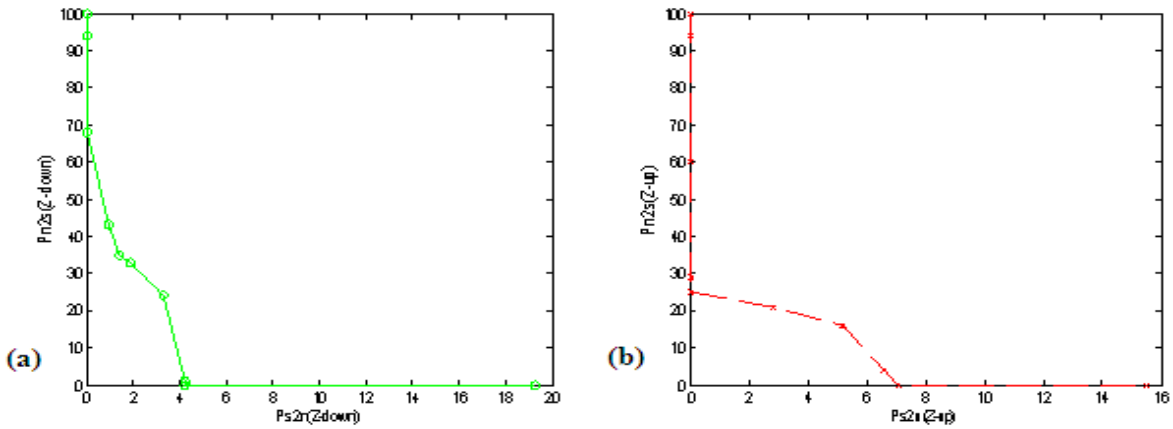

FIGURE 8: The classification rate variations in z-up (a) and z-down (b) versus threshold

5. CONSLUSION & FUTURE WORK

In this paper, a new VAD algorithm was presented on the basis of Gaussian modeling in spectro-temporal domain. The extracted features of this model in 4-D has been used in the proposed VAD. To provide the Gaussian modeling, the noise effectively passes through this space. Then a distance measurement was proposed. and applied to distinguish between noise and speech frames. Finally, the distance was compared to a given threshold for speech-silence classification. In our method, miss-classification rates were used for evaluation purposes. To provide a comparison, it was observed that the proposed method demonstrates better behavior in low SNRs. The proposed VAD algorithm can be applied to speech enhancement systems in spectro-temporal domain.

6. REFERENCES

1. N. Mesgarani, S. A Shamma, “Speech enhancement based on filtering the spectrotemporal modulations”, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Philadelphia, March 2005.
2. N. R. Garner, P. A. Barrett, D. M. Howard, and A. M. Tyrrell, “Robust noise detection for speech detection and enhancement”, Electron. Lett., Vol. 33, no. 4, pp. 270-271, Feb. 1997.
3. J.Sohn, N. S. Kim, and W.Sung, “A statistical model-based voice activity detection”, IEEE Signal Process. Lett., Vol. 6, no. 1, pp. 1-3, Jan 1999.
4. L.F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, “An improved endpoint detector for isolated word recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, pp. 777-758, 1981.
5. T. Kinnunen, E. Chernenko, M.Tuononen, P. Fränti, and H.Li, “Voice activity detection using MFCC features and support vector machine”, Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia, Vol. 2, 556-561, Oct 2007.
6. J.Sohn, W.Sung, “A voice activity detector employing soft decision based noise spectrum adaptation”, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 365-368, 1998.
7. Ángel de la Torre, Javier Ramírez, Carmen Benítez, Jose C.Segura, Luz García, Antonio J.Rubio, “Noise robust model-based voice activity detection”, INTERSPEECH2006, pp. 1954-1957, Pittsburgh, 2006.
8. J. –H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models”, IEEE Trans. Signal Processing, Vol. 56, no. 6, pp. 1965-1976, June, 2006.

9. J.W.Shin, J. -H. Chang, H. S. Yun, and N. S. Kim, “Voice Activity detection based on generalized gamma distribution”, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vol. 1, pp. 781-784, March 2005.
10. B. Meyer and M. Kleinschmidt, “Robust speech recognition based on localized spectro-temporal features”, in Proceedings of the Elektronische Sprach-und Signalverarbeitung (ESSV), Karlsruhe, 2003.
11. C.Shahnaz, W.-P.Zhu and M.O.Ahmad, “Aspectro-temporal algorithm for pitch frequency estimation from noisy observations”, in Proc. 2008 IEEE ISCAS, pp. 1704-1707, May 18-21, 2008, Seattle, USA.
12. T. Chi, P. Ru, and S. A. Shamma, “Multiresolution spectrotemporal analysis of complex sounds”, Journal of the Acoustical Society of America, Vol. 118, no. 2, pp. 887-906, 2005.
13. N. Kowalski, D. A. Depireux, and S. Shamma, “Analysis of dynamic spectra in ferret primary auditory cortex I. Characteristics of signal-unit response to moving ripple spectra”, J.Neurophsiology, Vol. 76, no. 5, pp.3503-3523, 1996.
14. K.Wang and S. A. Shamma, “Spectral shape analysis in the central system”, IEEE Trans. Speech Process. , Vol. 3, no. 5, pp. 382-395, Sep. 1995.
15. K. Wang and S. A. Shamma, ” Self-normalization and noise-robustness in early auditory representations”, IEEE Trans. Speech and Audio Proc, pp: 421–435, 1994.
16. S. A. Shamma, “Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve”, J. Acoust. Soc. Am., pp:1622–1632, 1985
17. S. Shamma, “Methods of neuronal modeling”, in Spatial and Temporal Processing in the Auditory System, pp. 411-460, MIT press, Cambridge, Mass, USA, 2nd edition, 1998.
18. 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.
19. N. Mesgarani, S. Shamma, and M. Slaney, “Speech discrimination based on multiscale spectro-temporal modulations”, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP ’04), Vol. 1, pp. 601-604, Montreal, Canada, May 2004.
20. E. Scheirer, and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator”, in Int. Conf. Acoustic, Speech and Signal Processing, Vol. 2, Munich, Germany, 1997, p. 1331.