A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform

Amard Afzalian a.afzalian@gmail.com
Islamic Azad University, Science
and Research,Tehran, IRAN.

Mohammad Reze Karami Mollaei mkarami@nit.ac.ir
Faculty of Electrical and Computer Engineering,
Babol Noshirvani University of Technology
Babol, P.O. Box 47135-484, IRAN

Jamal Ghasemi jghasemi@stu.nit.ac.ir
Faculty of Electrical and Computer Engineering,
Signal Processing Laboratory
Babol Noshirvani University of Technology,
Babol, P.O. Box 47135-484, IRAN

Abstract

Voice activity detector (VAD) is used to separate the speech data included parts from silence parts of the signal. In this paper a new VAD algorithm is represented on the basis of singular value decomposition. There are two sections to perform the feature vector extraction. In first section voiced frames are separated from unvoiced and silence frames. In second section unvoiced frames are silence frames. To perform the above sections, first, windowing the noisy signal then Hankel's matrix is formed for each frame. The basis of statistical feature extraction of purposed system is slope of singular value curve related to each frame by using linear regression. It is shown that the slope of singular values curve per different SNRs in voiced frames is more than the other types and this property can be to achieve the goal the first part can be used. High similarity between feature vector of unvoiced and silence frame caused to approach for separation of the two categories above cannot be used. So in the second part, the frequency characteristics for identification of unvoiced frames from silent frames have been used. Simulation results show that high speed and accuracy are the advantages of the proposed system.

Keywords: Speech, Voice Activity Detector, Singular Value.

1. INTRODUCTION

Voice activity detection is an important step in some speech processing systems, such as speech recognition, speech enhancement, noise estimation, speech compression ... etc. In speech recognition when a word or utterance begins or ends (the end points) must be specified [1]. Also VAD is used to disable speech recognition for silence segments. Some speech transition systems transmits active segments in high rate of bits and transmits silence in low rate of bits, by this method they improve the band-width [2]. In some speech enhancement algorithm for example spectral subtraction method, a VAD method is required to know when to update the noise reference [3,20,21]. Conversational speech is a sequence of consecutive segments of silence and speech. In noisy signal silence regions are noisy. Voice sound contain more energy than unvoiced sound, while unvoiced sounds are more noise-like, so in noisy condition activity detection is harder In unvoiced regions. Feature extraction is the most important section in VAD system that elicits required parameters from desired frame. To achieve an accurate algorithm, the system parameters must be selected until by them can be able to separate from each other areas. After proper feature election, threshold is applied to the extracted parameters and the decisions are made. To achieving good detection level threshold can be adapted to the change of the noise conditions. Many of the algorithms assume that the first frames are silence [5, 6, and 7], so we can initialize noise reference from these frames. Common features used in VAD's are short term Fourier transform and zero crossing rate [4, 6, 11, and 12]. Another important and widely used parameter in this regard is Cepstral Coefficient [7, 9]. In this method the Cepstral coefficients are calculated within frames and then by calculating the difference between this vector and the value assigned to the noise signal and then comparing the result with the basic threshold value, the frame identity could be determined. LPC method is also another major applicable method for VAD implementation [13]. Generally in LPC based algorithms a series of mean coefficients are experimentally considered for voice, unvoiced and silent modes. In the next step the LPC coefficients of suspicious frame and their relative difference with mean indices are calculated and the frame identity is recognized based on these values. The other parameters for implementing VAD in combined algorithms are LTSE (long term Spectral Estimation)[5], wavelet coefficient [8,10], the ratio of signal to noise in sub-band [14], LSPE(Least Square Periodicity Estimator)[11] and AMDF(Average Magnitude Difference Function)[15]. One of most important cases in VAD system is speed of system performance beside proper accuracy. In this paper is to present a new algorithm of VAD based on single value decomposition and frequency features, specifications accuracy and speed simultaneously be fulfilled. Based on this, paper organization as follows that in Section 2 single values decomposition (SVD) will be explained. In Section 3 the proposed method with the system block diagram is given. In Section 4 simulation results in terms of quantitative and qualitative criteria is evaluated. Finally, the article concludes with Section 5 ends.

2. Singular Value Decomposition

The singular value composition is one of the main tools in digital signal processing and statistical data. By doing SVD on a matrix with dimensions of M×N, we have:

$$X = U \Sigma V^T \tag{1}$$

On above relation U and V matrixes are singular vectors matrix with dimensions of M×M and N×N, respectively. Also, $\sum$ with r order of a diagonal matrix M×N is included singular values so that components on the main dial gauge are not zero and other components are zero. The elements on main dial gauge areas $\sigma_{11} > \sigma_{22} > ... > \sigma_{rr} > 0$ And are the values of X matrix. For exercising SVD to one dimensional matrix, the vector of signal samples must map to subspace with more dimensions, on the other hand must be changed to a matrix in certain way. Different ways have indicated for one dimensional signal transformation to a matrix that in this article (here) have used Hankel’s way [16,19].

3. Purposed algorithm

The main question on sound discovery is the classification of listening signal characteristic to diagnosis sound parts. Thus, listening signal are classified to sound classes: silence, voiced, and unvoiced. For classifying, suitable characteristics must elicit from the speech signal parts (frame). Before studying the details of purposed system, general block diagram are showed in figure 1.

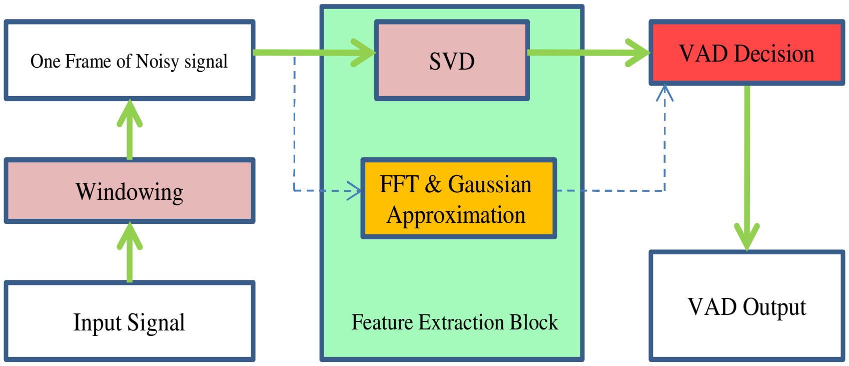


Figure (1) block diagram of propositional method for indicator system of voice activity.

As shown on figure 1, feature vector extraction done in two parts. In first part, voiced frames are separated from unvoiced and silence frame as for statistical characteristics of singular values matrix. In second part, unvoiced frames are separated from silence frames that this separation is based on its frequency spectrum and Gaussian rate in each frame. At the end, one value accrues to voiced and unvoiced frames that including voice information, and zeros one to silence frames.

3.1. Voiced frames separation from the other parts

In suggested system to separate the voiced parts from two parts of unvoiced and silence ones, it is used on slop of singular value curve in related part. For doing atop stages, first noisy signal divide to 16ms frames. According to relation (2), Hankel matrix makes for every frame.

$$X_k = [x_0, x_1, ..., x_{n-1}] \rightarrow H_k = \begin{pmatrix} x_0 & x_1 & \cdots & x_{M-1} \\ x_1 & x_2 & \cdots & x_M \\ \vdots & \vdots & \vdots & \vdots \\ x_{L-1} & x_L & \cdots & x_{N-1} \end{pmatrix} \quad (2)$$

Where X_k is the vector of exist samples in K frame in input signal and H_k is the isomorphic Hankel matrix of L×M dimensions with X_k . The percent of sample overlapping in matrix and the conditions of H_k dimensions are brought on (3) and (4) relations.

$$\%overlapping = \frac{L-1}{L} \times 100 \quad (3)$$

$$M + L = N + 1, L \geq M \quad (4)$$

For gaining full primitive pattern of X_k frame, the number of added zeros must lessen in M column of H_k matrix that its results have brought on (5) relation.

$$ZeroPadding = L - \mathrm{mod}(N, L) \Rightarrow L = [\frac{N}{2}] + 1, M = N - [\frac{N}{2}] \quad (5)$$

Gained values of L, M in (5) relation get H_k semi rectangular matrix with maximum sample overlapping. The singular values of each frame are got by Hankel matrix of existed frame and using SVD map on related singular values.

$$H_k = U_k \Sigma_k V_k^T \quad (6)$$

In (6) relation U_k and V_k the singular vectors of diagonal matrix and also Σ_k are isomorphic singular values matrix with H_k (part 2). Figure (2), singular values vectors show the every voiced frames in 16 millisecond length and SNR=10db with white Gaussian noise.

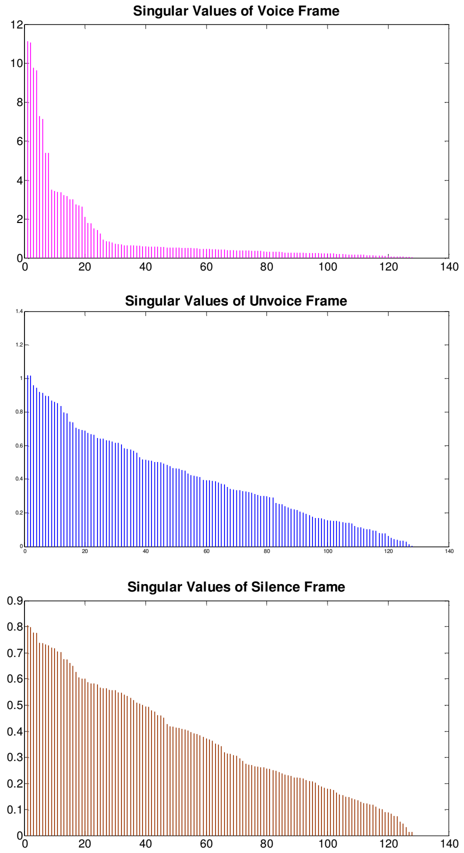


Figure (2) singular values vectors of voiced, unvoiced and silence frame in SNR=10db (voice signal from housewives_16.wav on TIMIT database)

According as seeing on figure (2) the slop of curve in singular values between voiced frames are different from the other parts. In fact, singular values of voiced frames have more slop than singular values of unvoiced and silence frames. The base of statistical feature extraction for separating voiced frames is the slop of singular values curve related to each frame by using linear regression. Table 1 shows the values of this slop in different SNRs on three certain frames that have chosen from voiced, unvoiced and silence parts.

Table (1):
slope of singular values curve in linear regression related to species of frame on different SNRs
(voice signal from housewives_16.wav on TIMIT database)

SNR	Mean amount of singular values curve slope on 10 times repeat for each SNR		
	Voiced frame	Unvoiced frame	Silence frame
0db	0.1498	0.0987	0.0949
5db	0.1232	0.0566	0.0528
10db	0.1170	0.0338	0.0305
15db	0.1143	0.0232	0.0176
20db	0.1139	0.0175	0.0098

Results of table (1) are support on this thesis that the slop of singular values curve on different SNRs in voiced frames are more than the others and by using this trait we can achieve the goal of first section that was the feature vector extraction in voiced frames from related singular values matrix.

2.3. Separating the voiced and silence frames

By studying figure (2) and table (1) are deduced that according to approximation through slop of singular values curves related to voiced and silence frames, it can't divide these two type frames from each other. Because of this in purposed system have used the other trait to separate these two parts. In this part, frequency trait has used for recognition unvoiced frames from silence ones. The base of comparison is the curve of Gaussian function, (7) relation.

$$f(x;\gamma,c) = e^{\frac{-(x-c)^2}{2\gamma^2}} \qquad (7)$$

Atop relation C is mean and γ is variance of curve. By doing study, we see the discrete Fourier transform of unvoiced frames in meddle frequency are similar to the curve of Gaussian functions to some extent. Figure (3) and (4) show the smooth frequency spectrum related to unvoiced and silence frames and their comparison with the curve of Gaussian function. (Voice signal from TIMIT database with SNR=15db)

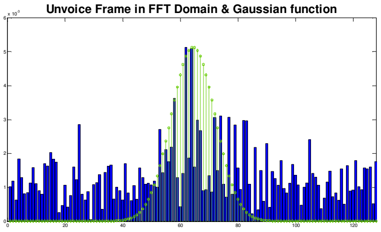


Figure (3) frequency spectrum of Unvoice frame in SNR=15db and the curve of Gaussian function ($\gamma = 0.6$)

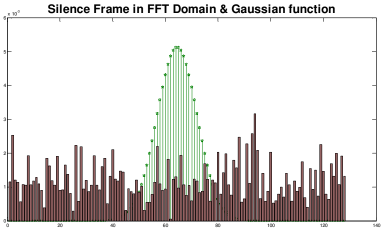
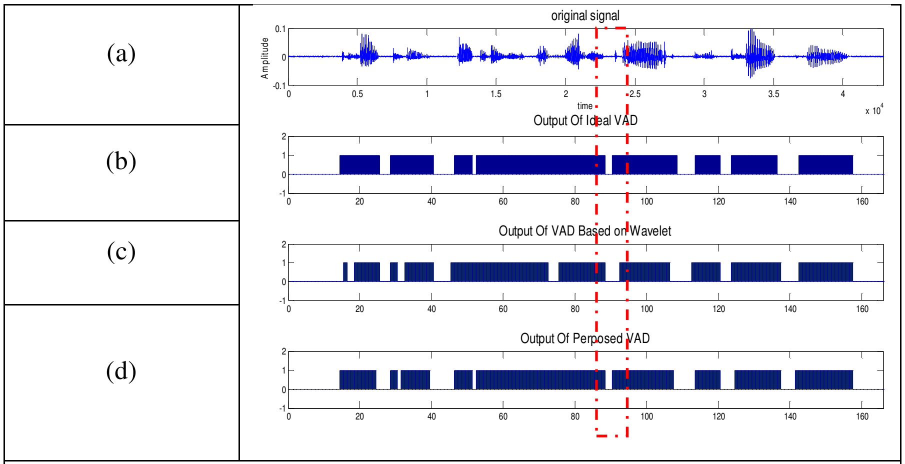


Figure (4) frequency spectrum of Silence frame in SNR=15db and the curve of Gaussian function ($\gamma = 0.6$)

According as atop figures are seen, frequency spectrum of voiced frame is more similar to Gaussian curve than silence frame noise in meddle frequency. Also by different examinations, the optimal values of γ parameter in varying SNRs are as 0.2 with a view of minimizing the difference between Gaussian pattern and frequency spectrum of silence frame pattern.

4. SIMULATION RESULTS

In this section, the operation of purposed system study and compare considered as accuracy and speed. Sx114.wav voice signal from TIMIT database including 166 of 16 millisecond frames with sampling rate is 16 kHz that each frame has 256 samples. Figure (5) shows the noisy signal with SNR=10db and the indicator systems output of voice activity.



Figure(5). Simulation Results: (a)TIMIT Signal, (b) Output of ideal VAD, (c) Output of VAD based on Wavelet, (d) Output of Purposed VAD

According as seeing, suggested method has more efficiency to keep the parts of signals that the activity signal is weak. In figure (5) and dashed line-drawn area, one unvoiced letter has omitted in VAD system as wavelet (figure 5-c) that it has kept on VAD output propositional system (figure 5-d). In table (2) the accuracy rate of VAD propositional system is shown as wavelet transform [17] about voice signal sx114.wav for different SNRs.

Table (2): percent of VAD system error comparison by using purposed algorithm and wavelet based algorithm in different SNRs

SNR	Purposed algorithm	Wavelet based algorithm
0db	25%	36%
5db	18%	20%
10db	15%	14%
15db	13%	12%
20db	8%	10%

One of the strength points of the purposed algorithm is its speed. In table (3), speed of two algorithms has compared with each other in processing the sx114.wav sound file (CPU Intel Core2Duo 2.5 GHz, 4 M Cache 733 MB RAM)

Table (3):
Speed comparison of VAD system by using purposed algorithm and wavelet based algorithm

Consumed time in purposed algorithm	Consumed time in wavelet based algorithm
4 second	12 second

In this part by using two VAD systems as preprocessing block has brought the results of hearing test for a speech signal rich-making system that accuracy of VAD operation system be proved in keeping unvoiced areas. In table 4 the standards has come that is used in evaluation of speech with hearing factor.

Table (4)
Five-Point adjectival scales for quality and impairment, and associated scores

Score	Impairment
5 (Excellent)	Imperceptible
4 (Good)	(Just) Perceptible but not Annoying
3 (Fair)	(Perceptible and) Slightly Annoying
2 (Poor)	Annoying (but not Objectionable)
1 (Bad)	Very Annoying (Objectionable)

In table 5, results of using two said VAD algorithms have shown as preprocessing block for enhancement method of multi band spectral subtraction. Specifications of this test are in [3].

Table (5)
Results of MOS test;
17 clean speech signal from TIMIT database; Noise Type: White Gaussian Noise.

Used Algorithm	Input SNR		
	0db	5db	10db
Wavelet Based	1.6	2.3	2.8
Purposed	1.8	2.7	3.3

Studying the results of table (5) are shown the reform efficiency of speech enhancement by using of propositional VAD algorithm to wavelet transform way.

5. CONCLUSION

In this paper a new method for Voice activity detector based on Singular Value Decomposition and discrete Fourier transform was proposed. The proposed method is evaluated by using various criteria. By using the mentioned criteria, it is presented that this method can compete with other methods. Also, the aim of indicator systems of voice activity is control of destruction unvoiced signal sites in rich-making operation that observantly to results of hearing test, the propositional algorithm have proper power in compare with wavelet transform way. The propositional system has manifold speed than usual method that is one of the obvious characters in practical usage and hardware implementation.

6. REFERENCES

[1] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, A. J. Rubio, “A new Kullback-Leibler VAD for speech recognition in noise”, IEEE Signal Processing Letters, 11(2) :266– 269. 2004.

[2] A. Kondoz, “Digital speech: coding for low bit rate communication systems”, J. Wiley, New York, 1994.

[3] Y. Ghanbari, M. R. Karami Mollaei, “A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets”, Speech Communication 48 (2006) 927–940.

[4] B. V. Harsha, “A Noise Robust Activity Detection Algorithm”, proc. Of int. symposium of intelligent multimedia, video and speech processing, pp. 322-325, Oct. 2004, Hong Kon.

[5] J. Ramırez, J. C. Segura, C. Benıtez, A. de la Torre, A. Rubio, “A New Adaptive Long-Term Spectral Estimation Voice Activity Detector,” EUROSPEECH, pp. 3041-3044, 2003, Geneva.

[6] J. Faneuff, “ Spatial, Spectral, and Perceptual Nonlinear Noise Reduction for Hands-free Microphones in a Car,” Master Thesis Electrical and Computer Engineering July 2002.

[7] S. Skorik, F. Berthommier, “On a Cepstrum-Based Speech Detector Robust To White Noise,” Accepted for Specom 2000, St. Petersbourg.

[8] J. Stegmann, G. Schroeder, “Robust Voice Activity Detection Based on the Wavelet Transform", Proc. IEEE Workshop on Speech Coding, Sep. 1997, pp. 99-100, Pocono Manor, Pennsylvania, USA.

[9] J.A. Haigh and J.S. Mason, “Robust Voice Activity Detection Using Cepstral Features,” In Proc. of IEEE TENCON'93, vol. 3, pp. 321-324, 1993, Beijng.

[10] J. Shaojun, G. Hitato, Y. Fuliang, “A New Algorithm For Voice Activity Detection Based On Wavelet Transform,” proc. Of int. symposium of intelligent multimedia, video and speech processing, pp. 222-225, Oct. 2004, Hong Kong.

[11] Tanyer S G, Ozer H, “ Voice activity detection in nonstationary guassian noise,” proceeding of ICSP’98 pp.1620-1623.

[12] Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R.,Prasad, R. V., Gaurav, V., “VAD Techniques for Real-Time Speech Transmission on the Internet”, 5th IEEE International Conference on High-Speed Networks and Multimedia Communications, pp. 46–50, 2002.

[13] L. R. Rabiner, M. R. Sambur, “ Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem,” IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, No. 4, pp. 338-343, August 1977.

[14] A. Vahatalo, I. Johansson, “Voice Activity Detection For GSM Adaptive Multi-Rate Codec,” IEEE 1999, pp. 55-57.

[15] M.Orlandi, a. santarelli, D. Falavigna, “Maximum Likelihood endpoint Detection with time-domain features,”eurospeech 2003, Geneva, pp.1757-1760.

[16] S. H. Jensen, P. C. Hansen, S. D. Hansen, “Reduction of Broad-Band Noise in Speech by Truncated QSVD,” IEEE, Trans on speech & Audio Processing, Vol.3, No.6, November 1995.

[17] Y. Ghanbari, M. Karami, “Spectral subtraction in the wavelet domain for speech enhancement,” International Conference on Information Knowledge Technology (IkT2004), CD ROM, 2004.

[18] H. Sameti, H. Sheikhzadeh, Li Deng, R. L.Brennan, “HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise”, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 5, September 1998.

[19] S. Sayeed, N. S. Kamel, R. Besar, “A Sensor-Based Approach for Dynamic Signature Verification using Data Glove”. Signal Processing: An International Journal (SPIJ), 2(1)1:10, 2008.

[20] J.Ghasemi, M. R. Karami Mollaei, "A New Approach for Speech Enhancement Based On Eigenvalue Spectral Subtraction" Signal Processing: An International Journal, (SPIJ) 3(4), 34:41, 2009.

[21] A. Afzalian, M.R. Karami Mollaei, J. Ghasemi، " A New Approach for Speech Enhancement Based On Singular Value Decomposition and Wavelet Transform", AJBAS In Press(2010).