A New Approach for Speech Enhancement Based On Eigenvalue Spectral Subtraction

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

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

Abstract

In this paper, a phase space reconstruction-based method is proposed for speech enhancement. The method embeds the noisy signal into a high dimensional reconstructed phase space and uses Spectral Subtraction idea. The advantages of the proposed method are fast performance, high SNR and good MOS. In order to evaluate the proposed method, ten signals of TIMIT database mixed with the white additive Gaussian noise and then the method was implemented. The efficiency of the proposed method was evaluated by using qualitative and quantitative criteria.

Keywords: Eigenvalues, singular values decomposition, Spectral Subtraction, Speech enhancement.

1. INTRODUCTION

Speech enhancement aims to improve the performance of speech communication systems in noisy environments. Speech enhancement may be applied, for example, to a mobile radio communication system, a speech to text system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired [6, 16 17]. Existing approaches to this task include traditional methods such as spectral subtraction [1, 16], Wiener filtering [16, 17], and Ephraim Malah filtering [2]. Wavelet-based techniques using coefficient thresholding approaches have also been applied for speech enhancement [3, 4, 13]. As alternative to these traditional techniques is studying speech as a nonlinear dynamical system [9, 10]. In [15] two nonlinear methods for speech enhancement based on Singular Value Decomposition are studied. In [11] chaoslike features have been proposed for speech enhancement. Generally, the approaches can be classified into two major categories of single-channel and multi-channel methods. Single channel speech enhancement is a more difficult task than multiple channel enhancements, since there is no independent source of information with which to help separating the speech and noise signals. In these applications, the spectral subtraction is one of the most popular methods in which noise is usually estimated during speech pauses [1, 7, 8, 14] In this research, a new speech enhancement method is presented by using Singular Value Decomposition (SVD) regarding to spectral subtraction idea. The efficiency of the proposed method is evaluated by using the qualitative and quantitative criteria. The organization of this paper is as follows: In Section 2, Spectral Subtraction (SS) method, phase space

reconstruction and Singular Value Decomposition (SVD) are discussed. In Section 3 and 4, suggested Eigenvalues Spectral Subtraction (ESS) algorithm and simulation results are presented respectively. Finally the paper will be concluded in section 5.

2. BACKGROUND

2.1 Spectral Subtraction Process

As a classic speech enhancement technique, Spectral subtraction (SS) works well when the noise is stationary. In this method noise spectra is estimated by using the silence segment and subtracted from the noisy signal spectra. For applying this method three conditions must be assumed [16]:

a. Noise must be additive.
b. Signal and noise must be uncorrelated.
c. One canal must be accessible.

There are many methods that work based on Spectral Subtraction and the original of them is Power Spectral Subtraction (PSS).

- **Power Spectral Subtraction (PSS)**

Assuming the noise is additive, we can model the corrupted speech signal by following equation:

$$y(n) = s(n) + d(n) \tag{1}$$

Where $s(n)$ and $d(n)$ is clean speech signal and noise respectively. According to the second assumption, the signal and noise are uncorrelated, so we can write:

$$r_d(\eta) = D_0 \delta(\eta) \tag{2}$$

Where r_d is autocorrelation function of noise signal and D_0 is a constant [16]. According to the equation 2 and by supposing that $s(n)$ and $d(n)$ signals are stationary, we can show:

$$\Gamma_x(\omega) = \Gamma_s(\omega) + \Gamma_d(\omega) \tag{3}$$

Where Γ is the power spectral density (PSD). So, if we can estimate $\Gamma_d(\omega)$ we will be able to estimate $\Gamma_s(\omega)$ as equation 4.

$$\hat{\Gamma}_x(\omega) = \Gamma_s(\omega) + \hat{\Gamma}_d(\omega) \tag{4}$$

Noise is estimated from silence frames. PSD is related to Discrete-Time Fourier transform (DFT) as:

$$\Gamma_y(\omega) = \frac{Y(\omega) Y^*(\omega)}{N^2} = \frac{|Y(\omega)|^2}{N^2} \tag{5}$$

We can conclude from equation 4 and 5.

$$\left|\hat{S}(\omega)\right|^2 = |Y(\omega)|^2 - \left|\hat{D}(\omega)\right|^2 \tag{6}$$

As mentioned above when $s(n)$ and $d(n)$ are stationary, equation 3 and 4 will be correct. Since the clean speech signals are locally stationary in short-time frames and also the assumption that noise is stationary is more acceptable in short time intervals, windowing is applied to the corrupted speech signal. Then the spectral subtraction is applied to each frame. To estimate the speech signal frames, the other necessary factor is $\varphi_s(\omega)$ as the estimated phase spectrum of speech frame. Boll has shown [1] that in practical applications, it is sufficient to use the noisy phase spectrum as an estimation of clean speech phase spectrum.

$$\hat{\varphi}_s(\omega) = \varphi_y(\omega) \tag{7}$$

Therefore from equation 6 and 7, we can obtain the estimated speech frames as shown in equation 8.

$$\hat{S}(\omega) = \left|\hat{S}(\omega)\right| e^{j\hat{\varphi}_s(\omega)} = \left[\left|Y(\omega)\right|^2 - \left|\hat{D}(\omega)\right|^2\right]^{1/2} e^{j\varphi_y(\omega)} \tag{8}$$

The PSS algorithm is shown in Fig. 1.

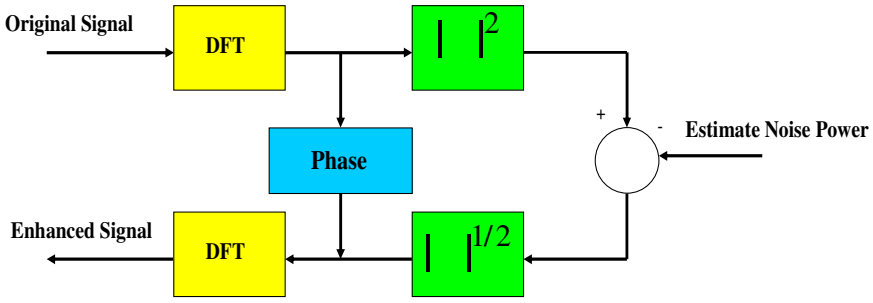

Figure 1. Spectral Subtraction Method

2.2 Phase Space Reconstruction

The dynamics of a system can be studied in a phase space, also called state space. Nonlinear time series methods perform analysis and processing in a reconstructed phase space, a time domain vector space whose dimensions are time lagged versions of the original time series [18, 19]. Takens time-delay-embedding method is probably the most common attractor reconstruction method in the literature [20]. Takens showed that if the embedding dimension is large enough, the reconstructed phase spaces have been shown to be topologically equivalent to the original system. Specifically, a scalar time series can be disclose in a multidimensional phase space using time delay coordinates. A brief describe of The Takens's method is as follows:
Given the time series $\{ s_n, n = 1, 2, ..., N \}$, the reconstructed attractor consists of the m vector

$$S_n = (s_n, s_{n+\tau}, s_{n+2\tau}, ..., s_{n+(m-1)\tau}) \tag{9}$$

Where τ and m are the time delay and embedding dimension respectively. A reconstructed phase space matrix S of dimension m and lag τ is called a trajectory matrix and defined by:

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_M \end{bmatrix} = \begin{bmatrix} s_1 & s_{1+\tau} & \cdots & s_{1+(m-1)\tau} \\ s_2 & s_{2+\tau} & \cdots & s_{2+(m-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ s_M & s_{M+\tau} & \cdots & s_{M+(m-1)\tau} \end{bmatrix} \quad (10)$$

Where in each row S_i , $i = 1, 2, \ldots, M$ represent individual points in the reconstructed phase space. The number of the points is $N = M + (m-1)\tau$.

2.3 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a very important tool in the problems of digital signal processing and data statistical analysis. The aim of SVD is to reduce the dimensions of a dataset as the reduced dataset still contains the variability features presented in the original data. The SVD theorem states that every real $m \times n$ ($m > n$) matrix X can be decomposed into a product of three matrices, as:

$$X = U \Sigma V^T \quad (11)$$

Where $U \in R^{m \times m}$ and $V \in R^{n \times n}$ are orthogonal matrices, i.e. $U^T U = I \in R^{m \times m}$ and $V^T V = I \in R^{n \times n}$ (with I identity matrix) and

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m \end{bmatrix} \quad (12)$$

Σ is a diagonal matrix with singular values $\sigma_1 > \sigma_2 > \ldots > \sigma_m > 0$. The singular values are the non-negative square roots of the eigenvalues of the covariance matrix $X^T X$ [5].

3. SUGGESTED ALGORITHM: ESS METHOD

Since noise is a random phenomenon, if we build a trajectory matrix from a noise segment and apply SVD on it, related $\sigma_i s$ steadily reduce. Comparing the clean and noisy signals shows that the eigenvalues corresponding to the noisy signals are some different from the clean signals. This difference depends on noise amount added to clean signal and also is related to eigenvalues corresponding to the added noise. The eigenvalues contain some information about the signal energy so it is reasonable to perform a spectral analysis. According to the above discussion similar to SS algorithm, our proposed algorithm called ESS is shown in Fig. 2.

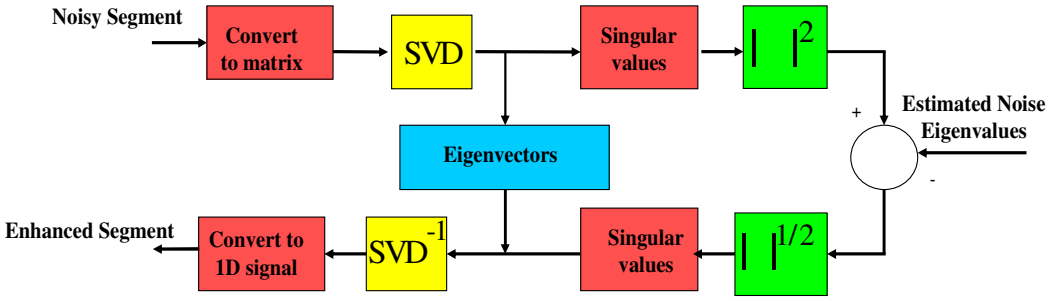

Fig.2: Eigenvalue Spectral Subtraction

In ESS, initial silence is used to estimate eigenvalues of the noise. The signal is segmented according to the silence segment and each segment is transformed to a matrix and SVD is applied (see section 2.2 and 2.3). As shown in Fig.2 the eigenvalues of the estimated noise are subtracted from eigenvalues of the noisy signal. The new obtained eigenvalues are used to reconstruct the corresponding segment. Simulation shows that, it is sufficient to use the noisy eigenvectors as an estimation of clean speech eigenvectors in each segment. It is noticeable that the eigenvectors are also saved to reconstruct the enhanced signal. The estimation and segmentation are shown in Fig. 3.

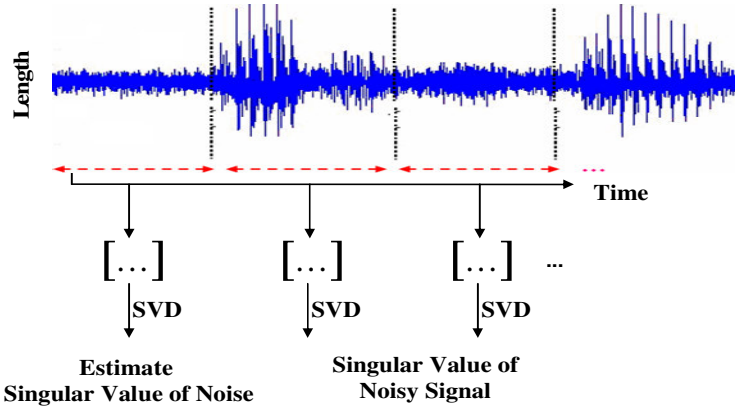

Fig. 3. Signal segmentation

4. SIMULATION RESULTS

Ten signals of TIMIT set were used to evaluate the algorithm efficiency. These signals are contaminated with additive white Gaussian noise. The new techniques were applied ten times to ten signals and the corresponding results were averaged. We arranged the simulation results as qualitative and quantitative results.

4.1 Quantitative Results

The signal to noise ratio (SNR) was used as quantitative criterion. Fig. 4 shows our algorithm results in comparison with results obtained by using the wavelet based method discussed in [4].

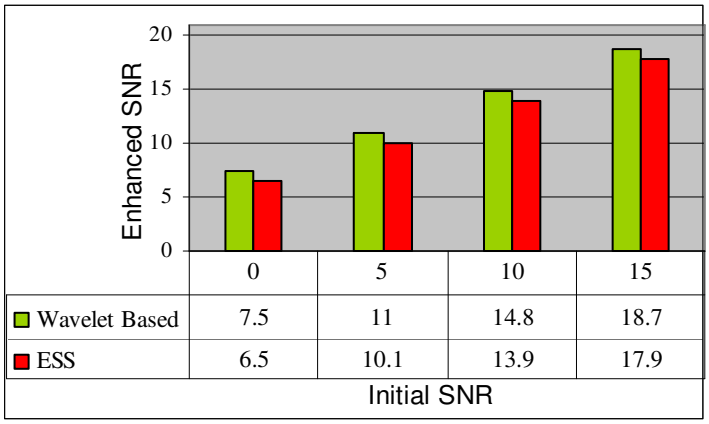

Fig 4. The comparison of the proposed method with wavelet based method

According to fig. 4 it is obvious that both algorithms perform much similarly to each other in aspect of SNR. It is noticeable that, since the proposed algorithm works in time domain, it is much faster than the last one which works in frequency domain (both algorithms implemented in Intel(R) Core(TM) 2Duo CPU).

4.2 Qualitative Results

For qualitative evaluation, we have shown the temporal results of clean, noisy and enhanced speech signal in Fig. 5.

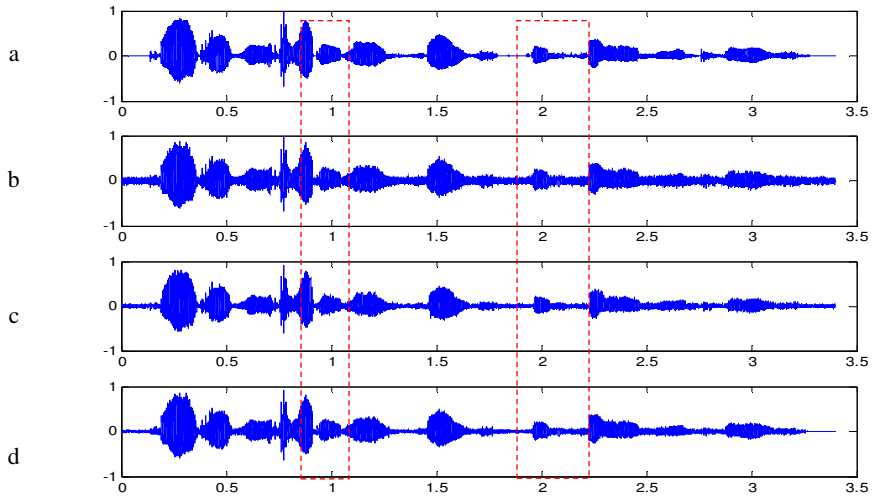

Figure 5: signals in time domain (ms)
a: Clean signal
b: Noisy signal
c: Wavelet based enhancement
d: ESS based enhancement

In Fig. 5, for simple comparison, two parts of the signals have been specified by using dashed lines. The signal enhanced by ESS (Fig. 5.d) is more similar to clean signal (Fig. 5.a) in comparison with wavelet-based enhanced signal (Fig. 5.c). For more investigation, audio experiments and the result of its implementation are presented.

- **Audio Experiment**

In this experiment, 6 persons (three women and three men) gave a mark to signals from 1 to 5. Ten speech signals with various SNRs (0, 5, 10 dB) and also their enhanced signals were used [12].The mean opinion scores (MOS) corresponding ESS and wavelet-based method are illustrated in Fig. 6.

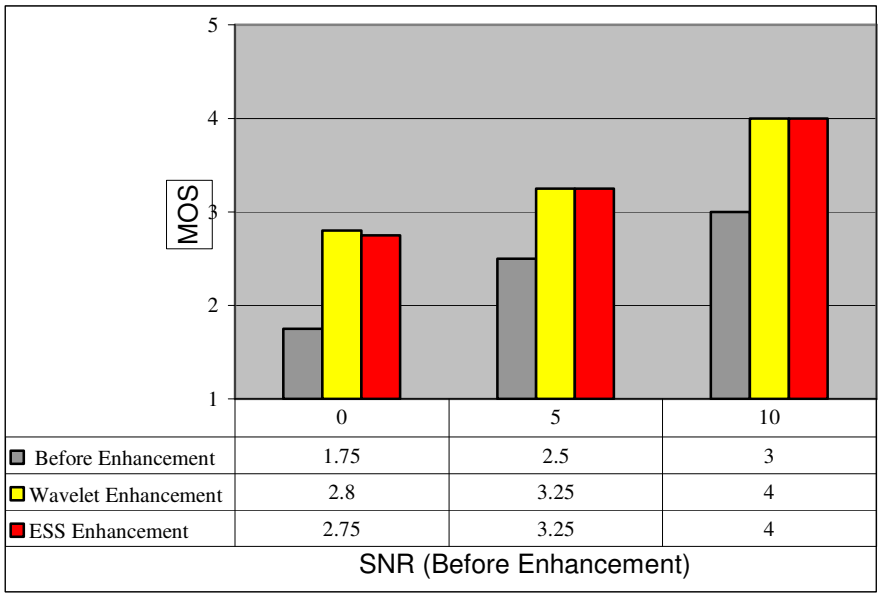

Fig. 6 MOS results for audio experiment

The MOS results show that both ESS and wavelet-based method have similar performance. So the hearing system of human beings is intelligence, it may neglect the noise undesirable and as a result, the audio test in both methods lead to the same results.

5. CONCLUSION

In this paper a new method (ESS) for speech enhancement was proposed. ESS is evaluated by using various criteria of quality and quantity. By using the mentioned criteria, it is presented that this method can compete with other speech enhancement methods. As seen in Fig.4 and 5 the proposed method provides proper performance in comparison with wavelet based method in terms of SNR. Mean opinion score also verifies the efficiency of the proposed method. Since ESS works in time domain it has faster than the frequency based methods. Another advantage of the proposed algorithm is that it does not require any voiced/unvoiced detection process by which the performance of the system is highly decreased. All of these mentioned advantages make ESS suitable for real time applications.

6. REFERENCES

1. S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”. IEEE Transactions Acoustics Speech Signal Process. 27, 113:120, 1979.
2. Y. Ephraim, D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator”. IEEE Transactions Acoust. Speech Signal Process. ASSP. 32 (6), 1109:1121. 1984.
3. D.L. Donoho, “Denoising by soft thresholding”. IEEE Transactions Information Theory, 41(3), 613:627, 1995.
4. Y. Ghanbari, M. R. Karami-M, “A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets”, Speech Communication 48, 927:940, 2006
5. S. Sayeed, N. S. Kamel and R. Besar. “A Sensor-Based Approach for Dynamic Signature Verification using Data Glove”. Signal Processing: An International Journal (SPIJ), 2(1)1:10,2008

6. O. O. Khalifa, Z. H. Ahmad, A-H A. Hashim and T. S. Gunawan “SMaTalk: Standard Malay Text to Speech Talk System”. 2(5) 1:16, 2008
7. H. Gustafsson, S. E. Nordholm and I. Claesson, “Spectral Subtraction Using Reduced Delay Convolution and Adaptive Averaging”. IEEE Transactions on Speech and Audio Processing, 9(8),799:807, 2001.
8. D. E. Tsoukalas, J. N. Mourjopoulos and G. Kokkinakis, “Speech Enhancement Based on Audible Noise Suppression”. IEEE Transactions on Speech and Audio Processing, 5(6), 497:514, 1997.
9. M. Banbrook, S. McLaughlin, and I. Mann, “Speech characterization and synthesis by nonlinear methods”, IEEE Transactions on Speech and Audio Processing, 7, 1:17, 1999
10. A. Kumar, S. K. Mullick, “Nonlinear Dynamical Analysis of Speech”, Journal of the Acoustical society of America,100, 615:629, 1966.
11. R. Hegger, H. Kantz, and L. Matassini, “Denoising Human Speech Signals Using Chaoslike Features”, Physical Review Letters, 84(14), 3197:31200, 2000.
12. H. Sameti, H. Sheichzadeh, Li Deng, R. L. Brennan, “HMM Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise”, IEEE Transactions on Speech and Audio processing, 6(5), 1988.
13. D. Guo, W. Zhu, Z. Gao, J. Zhang, “A study of wavelet thresholding denoising”. Paper presented at the International Conference on Signal Processing, Beijing, PR China,2000.
14. R. Martin, “Spectral Subtraction Based on Minimum Statistics”. Paper presented at the Europe Signal Processing Conference, Edinburgh, Scotland, 1994
15. M. T. Johnson , A. C. Lindgren, R. J. Povinelli, X. Yuan, “Performance of Nonlinear Speech Enhancement Using Phase Space Recognition Struction”, ICASSP 2003
16. J. R. Deller, J. H. L. Hansen, J. G. Proakis, “Discrete Time Processing of Speech Signals”, second ed. IEEE Press, New York (2000)
17. S. Haykin. “Adaptive Filter Theory”, third ed. Prentice Hall, Upper Saddle River, New Jersey (1996)
18. H. D. I. Abarbanel, “Analysis of Observed Chaotic Data”. Springer, New York (1996)
19. H. Kantz, T. Schreiber, “Nonlinear Time Series Analysis”, Cambridge University Press, Cambridge, England (1997)
20. F. Takens. “In Dynamical Systems and Turbulence”, edited by D. A. Rand and L.-S. Young, Lecture Notes, in Mathematics Vol. 898, Springer-Verlag, New York (1981)