Australian Journal of Basic and Applied Sciences, 4(8): 3602-3612, 2010
ISSN 1991-8178

A New Approach for Speech Enhancement Based On Singular Value Decomposition and Wavelet Transform

[1]1Amard Afzalian, [2]M.R. Karami Mollaei, [1]Massoud Dousti and [2]Jamal Ghasemi

[1]Islamic Azad University, Science and Research, Tehran, Iran.
[2]Signal Processing Laboratory, Babol Noshirvani University of Technology, Faculty of Electrical and Computer Engineering, P.O. Box 47135-484, Babol, Iran

Abstract: In this paper a new approach for speech enhancement is presented. The proposed algorithm is based on singular value decomposition (SVD) and wavelet transform. A model of contaminant noise is estimated by using SVD in the recommended method and then, using of noise estimation determines thresholding value. Needlessness of silence frame in order to estimate the noise model is an advantage of suggested method. Singular value tools give us a robust method of feature extraction from speech signal frames as well. The qualitative and quantitative evaluation shows that our proposed method highly improves the performance of speech enhancement based on wavelet thresholding.

Key words: Singular Value Decomposition,Thresholding On Wavelet Coefficients, Feature Extraction

INTRODUCTION

Speech enhancement methods tries separating the noise mixed with the speech signal and as much as possible, reducing its effects. Reducing the noise effect is very important in enhancing comprehensibility and reliability of human speech signal. Also, vast domain of sciences such as speech processing sciences like speech recognition, speech signal coding and …., it is used. For this purpose, many methods presented for reducing the noise effect from the speech signal, but none of them is completely optimized and needs to be improved or new methods should be introduced. Some of these methods are based on spectral subtraction (Berouti *et al.,* 1979; Ghanbari *et al.,* 2004; Boll, 1979; Lee *et al.,* 1997), adaptive filter (Whitehead *et al.,* 2003), Wiener filter (Tierney, 1980), LPC analysis (Sambur *et al.,* 1976; Ing Yann Soon, 1997), singular value decomposition (Khalifa et al., 2008; Ghasemi and Karami, 2009), and wavelet transform (Sheikhzadeh and Abutalebi, 2001; Donoho and Johnston, 1994; Johnston and Silverman, 1997; Seok and Bae, 1997; Ang et al., 1997).). However, the spectral subtraction method is so simple and efficient, but it produces a new noise that is called musical noise. To reduce the effect of this noise, spectral subtraction method is uses by using of over subtraction and spectral floor introduced by M.Berouti (Ghanbari *et al.,* 2004). Thereafter, Kamath and P.Loizoa developed multi band spectral subtraction. In this method, the noisy speech signal is divided into several frequency bands first, and then spectral subtraction method is uses on each band (Berouti *et al.,* 1979). Linear predictive coding method (LPC) is on the basis of this idea that speech signal is predictable due to the linear combination of previous samples of input and output. Therefore, enhanced signal is recreated by using of this assumption. Iterative Wiener filtering constructs an optimal linear filter using estimates of both the underlying speech and underlying noise spectra (Ang *et al*, 1997). The noise spectrum is estimated from silence frames as in spectral subtraction, while the speech spectrum in each frame is estimated iteratively, beginning with the noisy signal spectrum and using the Wiener filter output to get an improved estimate (Ghanbari and Karami, 2006; Whitehead *et al.,* 2003). However, such methods significantly depend on suitable estimation of signal variation of speaker's voice. In these methods a group of filters is used to filter the frequencies between the main frequency and its harmonics. The basis of wiener filter is the estimation of an optimized filter from input noisy speech that results from minimizing the mean square error (MSE) between the estimated signal and desired signal. In many noisy signal enhancement methods, there should be an estimation of noise model. Therefore, the success of such methods highly depends on the noise specification. Silence frame of signal usually is used to identify noise characteristics and this activity will not be enough effective without a voice activity detector system (VAD). But methods such as hidden Markov model (HMM) (loizou, 2007; Fujimoto and Nakamura, 2005), minimum mean square error (Srinivasan et al., 2007; Martin, 2001), MCRA method on

Corresponding Author: Amard Afzalian, Islamic Azad University, Science and Research, Tehran, Iran.
E-mail: a.afzalian@gmail.com

the basis of Short Fourier Transform (Stouten *et al.,* 2006; Ningping Fan *et al.*, 2006), or a combination of them is represented for noise estimation. On wavelet based noise decreasing methods, noise elimination is usually carried out by thresholding on wavelet coefficients (Donoho and Johnston, 1994; Johnson *et al.,* 2003). Donoho introduced a new algorithm based on wavelet thresholding for denoising the signals corrupted by white Gaussian noise (WGN) (Donoho, 1995). After that, employing this new method in speech enhancement was widely studied (Ghanbari and Karami, 2006; Ing Yann Soon, 1997; Sheikhzadeh and Abutalebi, 2001; Seok and Bae, 1997). In the most techniques which use the wavelet thresholding for speech enhancement. Choosing thresholding values are the basic parameters in such methods and depend on the noise model. In this paper an algorithm based on singular value decomposition is used to estimate the noise model, then the estimated noise model is used for determine the thresholding value in Donoho algorithm. Paper organized as follows, in section 2 wavelet, speech enhancement based wavelet thresholding and SVD are introduced. In section 3 the proposed algorithm is presented. Section 4 is including the simulation results. Finally the paper will be concluded in section 5.

2- Literature of Issue:

In this section, the used implements that are consisting of wavelet factor thresholding method description and introduction of singular value decomposition are introduced.

2-1- Wavelet Transform:

Wavelet analysis is totally a windowing technique with various locations in their dimension and make feasible to use longer of time periods in location that more accurate information of low frequency is needed and shorter of time period in locations that more frequency information is needed. It can be intuitionally seen that the signals with high variations can be better analyzed with erratic wavelet and local characteristics are more explainable by wavelet as well. Continues wavelet transform (CWT) as a summation on all the times is defined from signal multiplication by shifted and scaled version, given of wavelet function φ

$$C(scale, position) = \int_{-\infty}^{+\infty} f(t)\psi(scale, position, t)dt \quad (1)$$

CWT result of wavelet coefficients is C that is a function of location and scale. Multiplication of each coefficient by shifted and correct scaled wavelet gives the wavelets of basic signal constitutive. Figure (1) shows the applying of wavelet transform on a noisy sinusoidal wave.

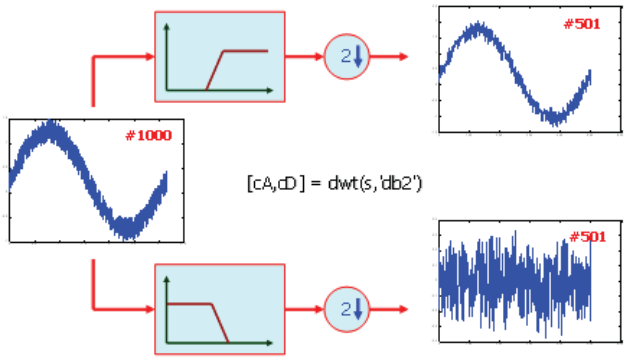


Fig. 1: Discrete wavelet transform applying method on signal

The results of discrete wavelet transform, as it was shown on figure (1), are two group of approximation (cA) and detail (cD) coefficients. Detail coefficients explain trivial details of high frequency of input wave.

2-2- Thresholding On Wavelet Coefficients:

Since noise mostly has random characteristic, variation of noise is high, detail coefficients are considered as noise. Eliminating of noise constructive components by thresholding on wavelet coefficients is on the basis of this fact that in many signals such as speech, energy is often aggregated in a few number of wavelet dimension (Kamath and Loizou, 2002). The coefficients of this dimension are approximately great in comparison with wavelet coefficients of other dimensions or noise coefficients (that their energy is propagated

in many numbers of coefficients), too. Therefore, by thresholding on wavelet coefficients in order to convert small coefficients to zero, noise components can be separated from signal components (Johnson *et al.,* 2003). To represent the basic algorithm of thresholding on wavelet coefficients beneath assumptions are supposed:

1. Noise is added to signal.
2. Speech signal and noise signal are uncorrelated.
3. There's only one channel to reach the signal.

Basic Algorithm:

First, let's assume that the input speech signal ***y*** in time domain is corrupted by additive noise ***n*** which is uncorrelated with clean speech ***s*** by following equation:

$$y = s + n \tag{2}$$

Here, there is no consideration of the channel distortion for convenience. If W is wavelet transforming matrix then the equation above is rewritten as follow:

$$Y = S + N; Y = W.y; S = W.s; N = W.n \tag{3}$$

Assuming above, wavelet coefficients of speech signal can be calculated by thresholding on wavelet coefficients of noisy signal:

$$\hat{S} = THR(Y,T) \tag{4}$$

That THR (…) is thresholding function and T is thresholding value.

Two well known thresholding functions that are used for decreasing the noise of signals by wavelet are include of soft and hard thresholding functions (Johnson *et al.,* 2003). These functions are specified by beneath equations:

$$THR_H(Y,T) = \begin{cases} Y & |Y| \geq T \\ 0 & |Y| < T \end{cases} \tag{5}$$

$$THR_S(Y,T) = \begin{cases} sign(Y).(|Y| - T) & |Y| \geq T \\ 0 & |Y| < T \end{cases} \tag{6}$$

One of the basic parts in basic noise eliminating algorithm on the basis of wavelet transform is noise estimation stage. Figure (2) shows the location of this part on Donoho's block diagram of recommended basic algorithm (Johnson *et al.,* 2003).

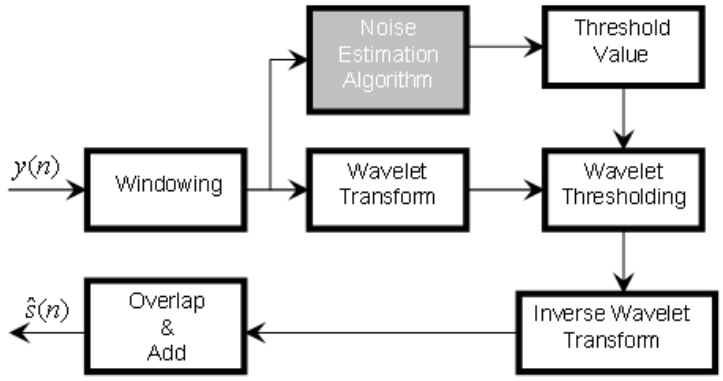


Fig. 2: Noise eliminating on the basis of thresholding on wavelet coefficients

As it can be seen on figure (2), after signal windowing on time domain, 1-D wavelet transform is applied on that and approximation and detail coefficients are extracted. Represented equation by Donoho in order to determine thresholding is:

$$T = \frac{median(|C|)}{0.6745}\sqrt{2Ln(length\ of\ noisy\ signal)} \tag{7}$$

In equation above, C is the wavelet coefficient sequence related to noise that an estimation of noise is required to determine it. In Donoho method, detail coefficients are used as estimated noise.

2-3- Singular Value Decomposition:

Singular value decomposition or SVD is an important tool to process digital signals and statistic (Sayeed et al., 2008; Khalifa et al., 2008; Ghasemi and Karami, 2009). By applying SVD on an M×N matrix, X, we have:

$$X = U\Sigma V^T \tag{8}$$

Where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $\boldsymbol{M}\times\boldsymbol{M}$ and $\boldsymbol{K}\times\boldsymbol{K}$ matrices, respectively. The columns of both $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices are orthogonal bases which span the row space and column space of the $\boldsymbol{X}$ matrix, respectively. Actually, $\boldsymbol{U}$ is a set of the eigenvectors of XX^T and $\boldsymbol{V}$ is a set of the eigenvectors of X^T X. On the other hand, $\sum$ is $\boldsymbol{M}\times\boldsymbol{K}$ diagonal matrix whose diagonal entries are known as the singular values of $\boldsymbol{X}$. Also it is the alternative of Eigen values corresponding to U and V. These singular values can be thought of as the weights of each basis vectors. In general they are sorted as ascending order of their values. In order to apply SVD on a 1-D signal, samples of signal should be mapped to a higher dimensional subspace. In other words, it should be converted to a matrix with special method (Ghanbari and Karami, 2006; Sayeed et al., 2008).

4- Proposed Algorithm:

Before investigating suggested method we showed total block diagram of purposed system in Figure (3).

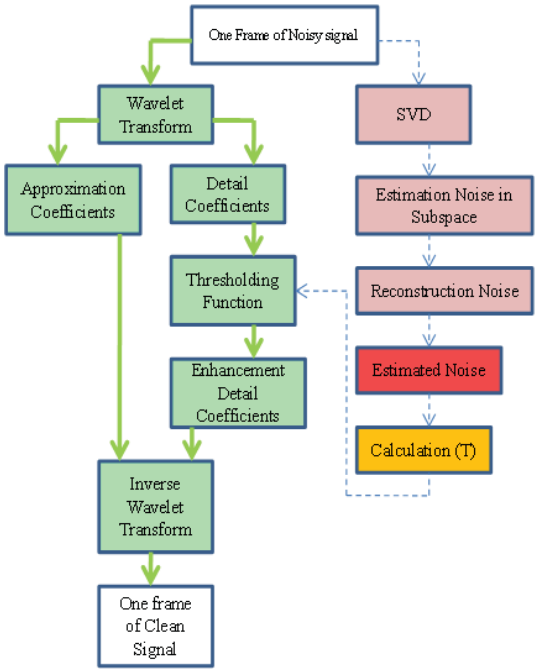


Fig. 3: Block diagram of recommended system

As it can be seen above, recommended noise reduction algorithm is separately applied on each signal frame. Recommended system has two steps on the basis of wavelet transform and SVD decomposition. It means, first, wavelet transform is applied to related frame and its approximation and detail coefficients are acquired. Then, thresholding value of detail coefficients is determined by applying noise estimation algorithm based on SVD. Finally, related frame is reconstructed by using of primary approximation coefficients and optimized detail coefficients.

3-1- Noise estimation by using SVD:

Assuming a noisy signal vector with the length of an instance N, we have:

$$x = [x_0, x_1, \ldots, x_{(N\text{-}1)}]^T \tag{9}$$

Considering additive and uncorrelation characteristics of noise with clean signal we have:

$$x = \overline{x} + n \tag{10}$$

$\overline{x}$ and n show the clean signal and noise respectively. There are many methods to convert a signal vector to a matrix. Constructing an L×N Henkel matrix in the following way is one of the most credible methods (Johnson *et al.,* 2003):

$$H = \begin{pmatrix} x_0 & x_1 & \ldots & x_{M-1} \\ x_1 & x_2 & \ldots & x_M \\ \vdots & \vdots & \vdots & \vdots \\ x_{L-1} & x_L & \ldots & x_{N-1} \end{pmatrix} \tag{11}$$

M and N are row and column numbers of matrix respectively, and L is amount of overlapping. With following limitation:

$$M + L = N + 1, L \geq M \tag{12}$$

Henkel matrix of noisy signal can be broken down as (Jensen *et al.,* 1995):

$$H = \overline{H} + N \tag{13}$$

and ***N*** are Henkel matrixes that are associated with clean signal and noise respectively. The observation matrix ***H*** is decomposed by SVD as follows:

$$H = U \sum V^T, H = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \sum_1 & 0 \\ 0 & \sum_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \tag{14}$$

$$\sum = \begin{pmatrix} \sigma_0 & 0 & \ldots & 0 \\ 0 & \sigma_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \end{pmatrix}, \sigma_0 > \sigma_1 > \ldots > \sigma_r \tag{15}$$

$\sigma_0 > \sigma_1 > \ldots > \sigma_r$ are singular values of matrix H with the rank of *r*. SVD, and N can be separated from each other by assuming $H \in R^{L \times M}, U \in R^{L \times M}, V \in R^{M \times M}$ (equations 16, 17):

$$\overline{H} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \overline{\Sigma}_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \tag{16}$$

$$\bar{N} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \bar{\Sigma}_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \tag{17}$$

To transmit samples of data to a higher dimension subspace, data set which are more correlated to each other are placed in one direction. It means that their related singular values are larger (Jensen et al., 1995; Sayeed et al., 2008). Corresponding Singular values of a noisy and clean frame are represented to show this subject in the following figures.

Comparing figures (4) and (5) shows that, if clean frame became noisy then the noise related singular values are added to smaller values of main diagonal of matrix $\sum$. By an inverse process of generating Henkel matrix that was explained in equation (11), acquired estimated noise matrix in equation (17) is decomposed to one dimensional space and then by taking into consideration of equation (7) is assumed as estimated noise.

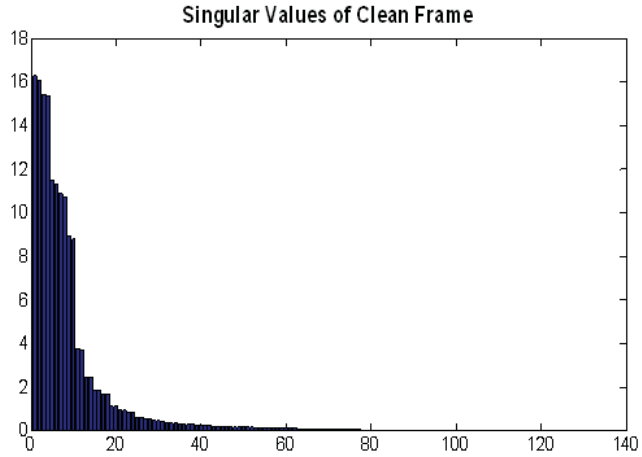


Fig. 4: Singular values of a clean speech frame

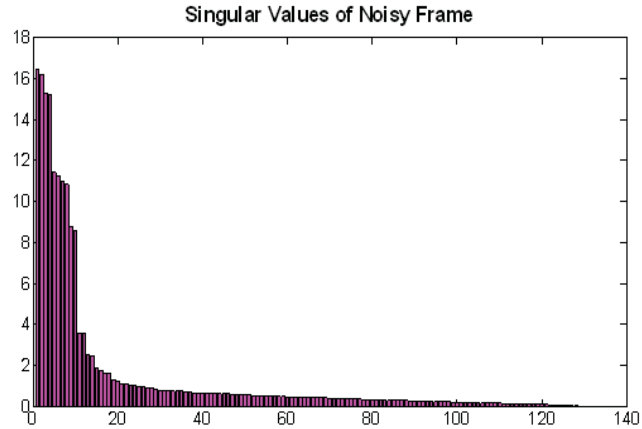


Fig. 5: Singular values of a noisy speech frame

3-2- Determining Thresholding Value:

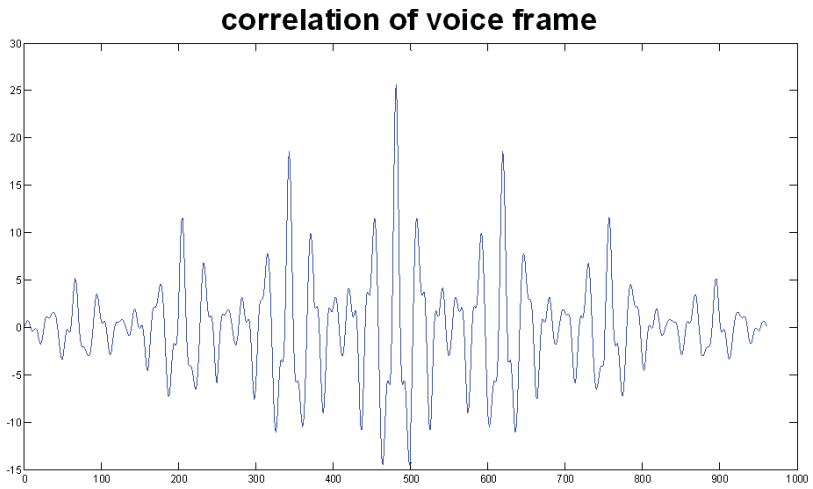
Determining a suitable value of thresholding is related to a suitable estimation of noise. According to equations (14-17) it can be seen that suitable separation of $\sum_1$ and $\sum_2$ has significant importance on suitable estimation of noise. Speech signal frames are divided into 3 separated pattern as Voiced, unvoiced and silence. Unvoiced letters has smaller amplitude than the voiced letters and usually has noisy essence, whereas voiced letters are naturally periodic that this fact increases the self correlation of these frames in comparison with unvoiced types. To show that, self correlation function of the 3 mentioned pattern is illustrated on figure (6):

Figure (6) shows the periodic nature of voiced frame and noisy nature of unvoiced frame. More self correlation by itself makes the singular value of voiced frames bigger than unvoiced frames. Figure (7) shows the singular values of voiced and unvoiced frames.

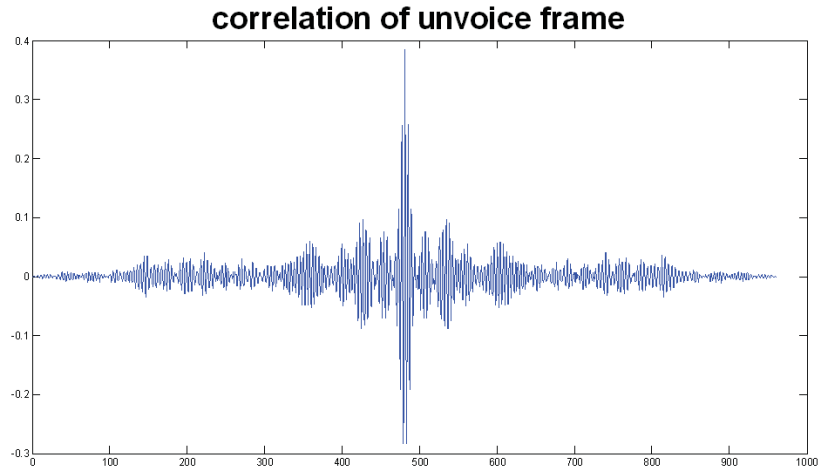
Figure (7) shows that decreasing tendency of singular values associated with voiced and unvoiced frames is obviously different. Therefore, to separate the desired singular values and signal from matrix $\sum$, type of the frame must be considered more. As a matter of fact, VAD is used to recognize the frame type (Kamath and Loizou, 2002). By using VAD and for determine noise value, following instruction is applied:

In silence frames, all the obtained singular values imply the noise. Therefore, all these values take part in reconstructing of noise sequence. The following equation is used for voiced and unvoiced frames.

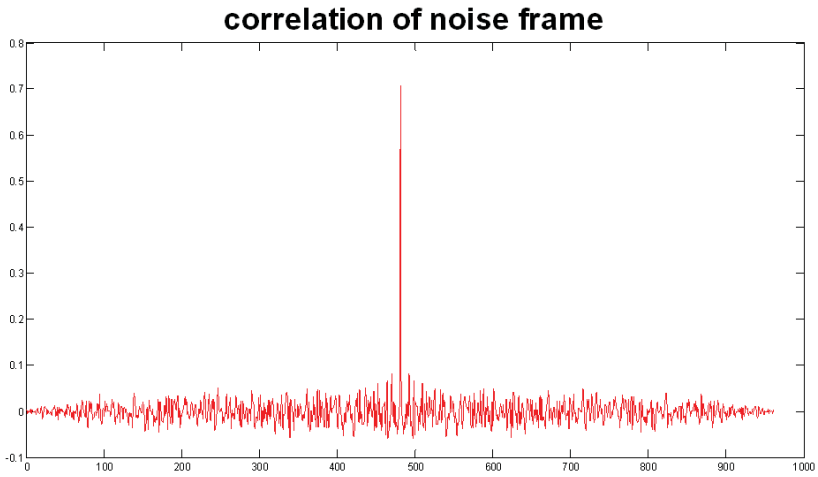
$$\alpha \sum_{i=1}^{n} \sigma_i^2 = \sum_{j=m}^{n} \sigma_j^2 \tag{18}$$



(a)



(b)



(c)

Fig. 6: Correlation function of the 3 pattern: (a)voiced, (b)unvoiced and (c)silence

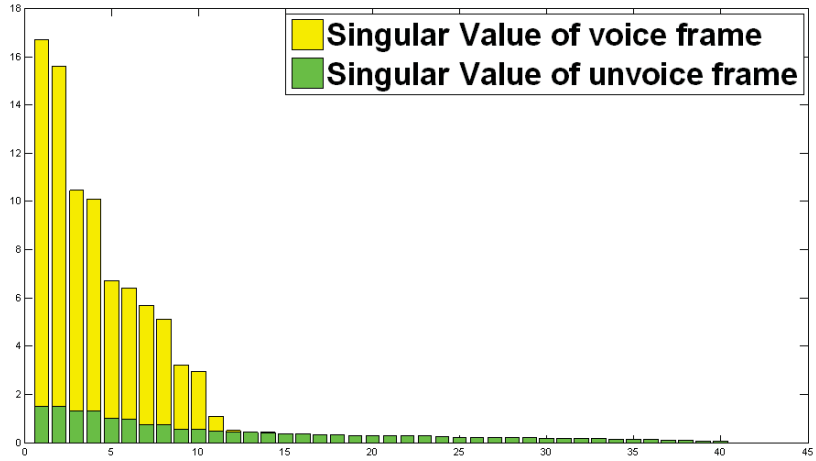


Fig. 7: Singular values comparing of voiced and unvoiced frames

In equation (18), σ shows the singular value and α is a parameter that depends on type of voiced and unvoiced frame and SNR of input frame as well. Optimized values of α has been shown in table (1):

Table 1: Optimized values of α for frame type and SNR value

Frame type	SNR=-5dB	SNR=0dB	SNR=5dB	SNR=10dB	SNR=15dB
Voiced	12%	10%	7%	4%	3%
Unvoiced	27%	22%	20%	17%	15%

It can be seen in table (1) that the value of α increases by increasing SNR. Increasing of α results in increasing of more number of singular values during the creation of matrix. Value of α with identical SNR is bigger for unvoiced frames than voiced. It shows that the system must be more sensitive when it encounters with unvoiced signals because, by uncontrolled increasing of α, some information of signal may be lost together with noise. It is due to the noisy nature of unvoiced frame.

4- Simulation Results:

In simulation results, frames have 30 msec (480 samples) lengths. Purposed method had been compared

with Donoho's method (thresholding function is a soft type in both methods.) and in order to recognize voiced and unvoiced frames from each other, recommended VAD system on (Kamath and Loizou, 2002) had been used. In first section, the results of purposed method applied on one of TIMIT standard signals that had been noisy with a 10 db white Gaussian noise has been shown on time domain in figure (8).

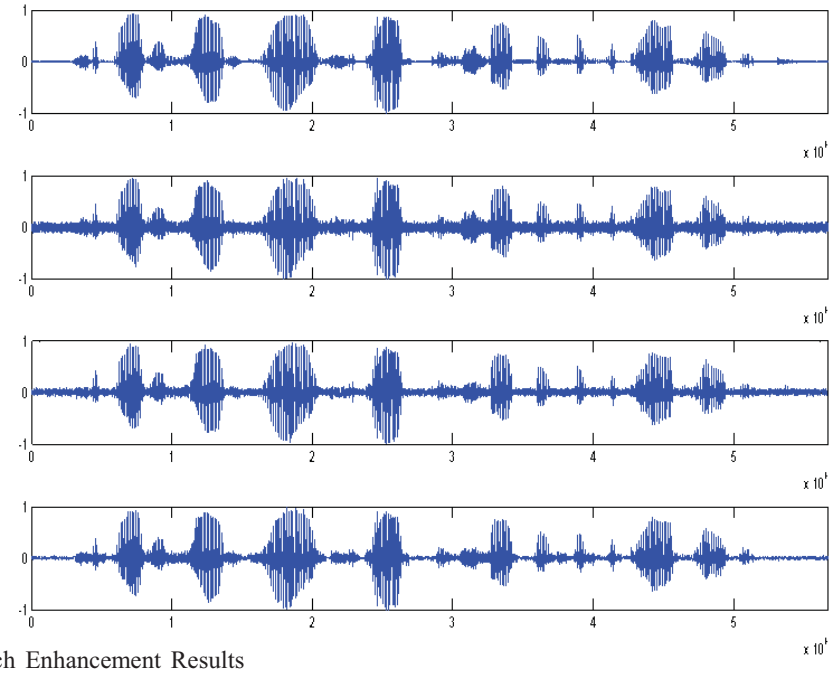


Fig. 8: Speech Enhancement Results
(a) Clean Speech
(b) Noisy Speech (SNR=10dB)
(c) Enhanced Speech By Wavelet method (SNR=11.7dB)
(d) Enhanced Speech By SVD-Wavelet Method(SNR=13.1dB)

Figure (8) shows that purposed method in voiced and silence frames, has more capability than Donoho's method. In other word, it had been caused less distortion. In other section of simulation Donoho and purposed method applied on 17 signals from TIMIT. SNR calculating progress for each signal is repeated 10 times. Figure (9) shows the simulation results.

Table 2: PESQ test results for seventy speech signal degraded by white noise

Method	InputPESQ=1.45, InputSNR=0dB	InputPESQ=1.84, InputSNR=5dB	InputPESQ=2.2, InputSNR=10dB	InputPESQ=2.55, InputSNR=15dB
Wavelet	1.46	1.85	2.21	2.56
SVD-Wavelet	1.48	1.88	2.25	2.7
Wavelet Packet	1.6	1.95	2.41	2.83
SVD-Wavelet Packet	1.76	2.21	2.6	3

By comparing recommended method and Donoho's method (with soft thresholding) and in figure (9), it is obvious that recommended method on all signals to noise levels is better than Donoho's method. Despite it is hard to separate the noisy and silence parts in low values of signal to noise, SNR optimization in this level is more important. Figure (9) also verifies that recommended method well corrected Wavelet Packet algorithm described in (Ghanbari and Karami, 2006). The perceptual evaluation of speech quality (PESQ) measures was used for objective evaluation of the proposed estimators (Loizou, 2007). In Table (2) improvement of PESQ test can be easily seen.

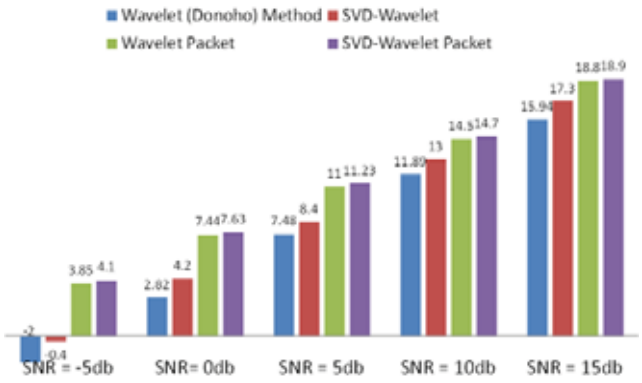


Fig. 9: The average performance of four algorithms for seventy noisy signals by WGN.

Conclusion:

Simulation results show the admissible capability of singular values decomposition in noise characteristic recognition. This characteristic had been used in this research on thresholding of wavelet factors in order to optimize the speech signal. Despite recommended method estimates noise regardless of silence frame, it excels than Donoho's method and Wavelet Packet algorithm. Simulation results verified the performance of the method on SNR optimization, PESQ test and preserving time characteristics.

REFERENCES

Adrian Ang Ee-Luang Ang A.B.Premkumar, A.S Madhukumar, 1997. "Time-frequency plane Wiener filtering for robust processing of speech signals", TENCON '97. IEEE Conf., 1: 35-38.

Berouti, M., R. Schwartz and J. Makhoul, 1979. enhancement of speech Corrupted by acoustic noise, *proc. IEEE ICASSP,* Washington DC, 208-211.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. on Acoust. Speech & signal processing*, 27: 113-120.

Donoho, D.L., I.M. Johnston, 1994. "Ideal Spatial adaptive via wavelet shrinkage", Biometrika, 81: 425-455.

Donoho, D.L., 1995. "De-noising by soft-thresholding, "*IEEE Transactions on Information Theory,* 41(3): 613-627.

Ghanbari, Y., M.R. Karami, 2006. "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", Speech Communication, 48: 927–940.

Ghanbari, Y., M.R. Karami, B. Amelifard, 2004. "improved multiband and spectral subtraction method for speech enhancement", *Proceedings of the 6th ISTED International conference signal and image processing*, pp: 225-230.

Ing Yann Soon, 1997. Soo Ngee Koh, Chai Liat Yeo, Wavelet for Speech Denoising, TENCON 97, Brisbane, Australia, pp: 479-482.

Jensen, S.H., P.C. Hansen, S.D. Hansen, 1995. "Reduction of Broad-Band Noise in Speech by Truncated QSVD*," IEEE, Trans on speech & Audio Processing,* 3(6).

Johnson, M.T., A.C. Lindgren, R.J. Povinelli, X.Yuan, 2003. "Performance of nonlinear speech enhancement using phase space reconstructions," IEEE, International Conference on Volume 1, Issue, 6-10 April 2003 Page(s): I-920 - I-923.

Johnston, I.M., B.W. Silverman, 1997. "Wavelet threshold estimators for data with correlated noise", J. Roy. Static. Soc., 59: 319-351.

Kamath, S. and P. Loizou, 2002. A Multi-band spectral subtraction method for Enhancing speech corrupted by colored noise, proceedings of ICASSP-Orlando, FL.

Lee, K.Y., B.G. Lee, S. Ann, 1997. Adaptive filtering for speech enhancement in colored noise, IEEE Trans. On Signal Processing Letters, 4: 277-279.

Loizou, P., 2007. "Speech enhancement: Theory and Practice", CRC Press.

Sambur, M.r. and N.s. Jayant, 1976. LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise, IEEE Trans.Acoust. Speech and signal process, 24(6): 488-494.

Seok, J., K. Bae, 1997. "Speech enhancement with reduction of noise components in wavelet domain, Proceeding of the 1997 IEEE International Conference on Acoustics', Speech, and Signal Processing (ICASSP'97), 2(21-24): 1323-1326.

Sheikhzadeh, H., H. R. Abutalebi, 2001. "An Improved Wavelet-Based Speech Enhancement System", in proc. 7th European Conference on Speech Communication and Technology (Euro Speech), Aalborg, Denmark, Sep.

Tierney, J., A study, 1980 of LPC analysis of speech in additive noise, IEEE trans. Acoust. Speech and signal process, 28(4): 389-379.

Whitehead, P.S., D.V. Anderson, M.A. Clements, 2003. Adaptive acoustic noise suppression for speech enhancement, IEEE International Conference on Multimedia & Expo.

Ing Yann Soon, 1997. Soo Ngee Koh, Chai Liat Yeo, Wavelet for Speech Denoising, TENCON 97, Brisbane, Australia, pp: 479-482.