SPEECH ENHANCEMENT BASED ON SPARSE THEORY UNDER NOISY ENVIRONMENT

Ching-Tang Hsieh,
Tamkang University, Taiwan
hsieh@ee.tku.edu.tw

Yan-heng Chen,
Tamkang University, Taiwan
akermh@hotmail.com

Ting-Wen Chen
Tamkang University, Taiwan
alucardoom@hotmail.com

Li-Ming Chen
Tamkang University, Taiwan
ms0071799@hotmail.com

ABSTRACT

Recently, the sparse algorithm for sparse enhancement is more and more popular issues. In this paper, we classify the process of the sparse theory to enhance speech signal into two parts, one is for dictionary training part and the other is signal reconstruction part. We focus on the White Gaussian Noise. Clean speech dictionary D is trained by K-SVD algorithm. The orthogonal matching pursuit(OMP) algorithm is used to obtain the sparse coefficients X of clean speech dictionary D. Denoising performance of the experiments shows that our proposed method is superior than other methods in SNR, LLR, SNRseg and PESQ.

Keywords - Speech enhancement, sparse representations, K-SVD, discrete cosine transform (DCT), orthogonal matching pursuit (OMP).

1. INTRODUCTION

Speech is the most important tool of expression and it is quite natural for the people who communicate with each other. If we need to use machine for communicate, we need speech processing for help. Speech processing has been a very popular research issues. Such as used for speech recognition system. When the speech in noisy environment, noise will make speech recognition rate decreased. Then, speech enhancement processing is necessary. Speech enhancement research methods in time domain, frequency domain and wavelet domain such as spectral subtraction [1], Wiener filter [1], Kalman filter [2] have been proposed. Those methods have a certain effect on speech enhancement. Recently, more and more people concerned about the

sparse representations issue. The main thing is dictionary learning. Prior to dictionary learning was based on probabilistic data [3]. After that, Michal Aharon, Michael Elad and Alfred Bruckstein proposed the K-SVD method [4]. The method of dictionary update step, we can update the dictionary and its coefficients at the same time. Recently, learning dictionary by sparse representation based on L1-minimization [5]. Some methods have been proposed based on sparse representation. Such as, Image processing based on sparse representations has been successfully used for Image denoising [6]. Zhiminand and Yuantao [7] use sparse prior information for speech enhancement. In this paper, each input signal will get their learning dictionary and a set of coefficients. After that, we use the trained dictionary and reconstructed coefficients to estimate the clean speech signal. This part is described in Section II. In section III, We compare some experimental results with other methods. We discuss future research works and conclusion in Section IV.

2. Sparse representation of speech denoising

In this section, we classify the process of the sparse theory to enhance speech signal into two parts, one is for dictionary training part and the other is signal reconstruction part.

A. Dictionary training part

In dictionary training part, first, we slide a window to divide the sequence of noisy speech signal into N frames, the window length is K, and then stored in an matrix $Y = [y^1, y^2, y^3, y^4 \cdots, y^N]$ with K×N dimension, where y_k is subtracted with each individual mean to exclude the noise. Second, DCT coefficients of unit matrix are calculated to set up the initialized dictionary ,D, with K×M dimension. M is random number, where M≤N. The DCT is just such a suitable choice and also the number of iterations is less than others coefficients [6]. Then, we optimize these two matrixes Y and D via OMP algorithm to obtain the sparse coefficients representations of matrix ,X, with M×N dimension. We finally use K-SVD algorithm to update the dictionary ,D. These functions are given by (1)-(2). Where $j \in \{1,2,\ldots,N\}$ and T_0 is preset number. We rewrite equation (1) as (3), where $n \in \{1,2,\ldots,N\}$. The total error matrix representation E_n is given as (4).

$$\|Y - DX\|_2^2 = \sum_{j=1}^{N} \|y_j - Dx_j\|_2^2 \qquad (1)$$

$$\min_{x_j} \left\{ \|y_j - Dx_j\|_2^2 \right\}, \ \|x_j\|_0 \le T_0 \qquad (2)$$

$$\|Y - DX\|_2^2 = \left\| Y - \sum_{j=1}^{N} d_j x_j \right\|_2^2$$

$$= \left\| \left(\mathbf{Y} - \sum_{\mathbf{j} \neq \mathbf{n}} \mathbf{d_j x_j} \right) - \mathbf{d_n x_n} \right\|_{\mathbf{2}}^{\mathbf{2}}$$

$$= \|E_n - d_n x_n\|_2^2 \qquad (3)$$

$$E_n = Y - \sum_{j \neq n} d_j x_j \qquad (4)$$

According to the singular value decomposition (SVD) method, we update a column of dictionary ,D, and a column of coefficients ,X, for each iteration. We decompose the total error matrixE_n via SVD method as $E_n = U\Delta V^T$. We take the first column of U to modify the column of dictionary d_n. First column of V is multiplied with $\Delta(1,1)$ to replace the x_n coefficients. We repeat the process several times to obtain a new updated dictionary ,D', without noise. The training process is given in Fig.1.

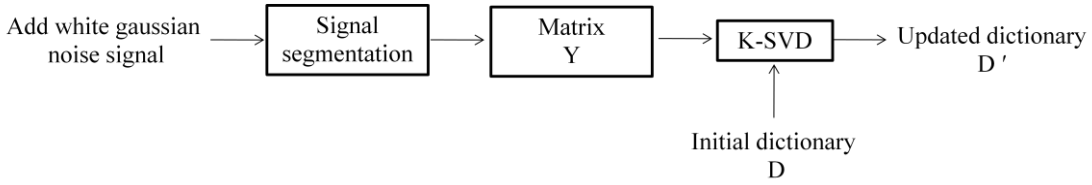

Fig.1 The training process of updated dictionary D'.

B. Signal reconstruction part

In signal reconstruction part, we use matrix ,Y, without the DC value of each frame and updated dictionary ,D', via OMP algorithm to obtain the sparse coefficients representations of matrix ,X, that belonging to dictionary ,D'. Then, we multiply these two matrixes D' and X to reconstruct the clean speech signal. The reconstruction process of clean speech signal is given in Fig.2.

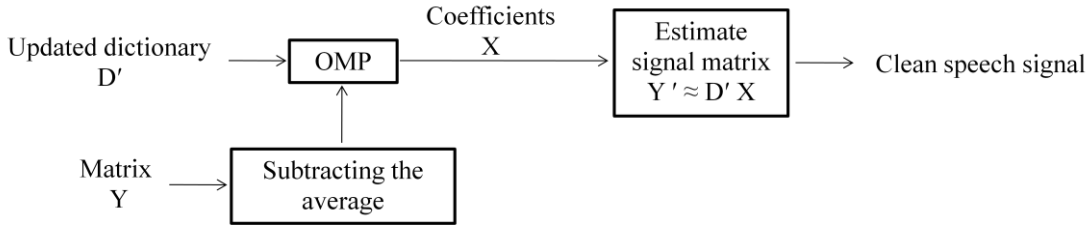

Fig.2 The reconstruction process of clean speech signal.

3. Simulation results

In this section, we try to evaluate the effect of denoising signal by using four kind of objective quality measures [8]-[9] such as the SNR, Log-Likelihood Ratio (LLR), segmental SNR (SNRseg) and Perceptual Evaluation of Speech Quality (PESQ). We also compare the proposed method with other methods such as Spectral subtraction [1], Wavelet coefficients thresholding [10], Noise absolute mean subtraction [11], Wiener filter [1] and Adaptive Wiener filter [1]. The clean utterances are taken from CHIME data [12] which includes 600 utterances by 34 speakers reading 6 sequences of the command-color-preposition-letter-number-adverb. All data have a 16kHz sampling rate. Input signal will be limited to the amplitude range between -1 to 1. The speech signal will pass a high-pass filter to eliminate the effect of vocal cords and lips during phonation. It can also amplify the high-frequency formants. The test signals are added with white gaussian noise at SNR levels of -10, -5, 0, 5 and 10 dB. The

"s17_sbwt1a.wav" clean sample signal taken from CHIME database is shown in Fig. 3. Clean signal is added with white gaussian noise at SNR 5 dB is shown in Fig. 4.We compared with five others denoising method and the results of four kind of quality measures assess are shown in Fig. 5-10. The results of denoising signal and time-varying spectrogram are also given in Fig. 5-10.The compared results with five other enhancement methods under four objective quality measures at SNR 5dB level is tabulated in Table.1. From Table.1, we conclude that our method is superior than other methods at SNR 5dB level.Fig.11 to 14 show four assessment results, respectively, which is compared with five other enhancement methods at SNR levels of -10, -5, 0, 5 and 10 dB. From Fig.11 to 13 show that our proposed method is better than five other enhancement methods at SNR levels of -10, -5, 0, 5 and 10 dB. In PESQ assessment measure, our method is still better than other methods over SNR -5dB level, but is worse than other under SNR -5dB level. Because $Dx_j \approx y_j$ as bounded error limits, $\left\|Dx_j - y_j\right\|_2 \leq \epsilon$ .The threshold $\epsilon = 1.1\sigma$. In our method, the noisy speech under SNR -5 dB level, where the threshold ϵ is obtained withlargerσ value, will lead to worse sparse coefficients X and also the incomplete reconstructed speech signal. In this situation, the incomplete speech signal will drop the PESQ assessment. There is no such case in other assessment measures, because they consider the global effect and but it emphasis the local effect in PESQ assessment.

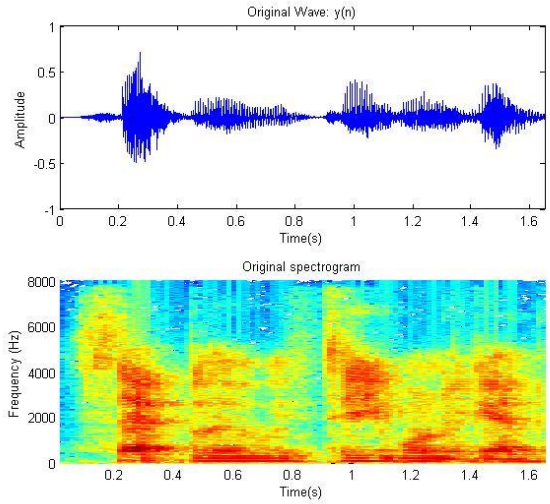

Fig.3 Clean signal waveform and spectrogram.

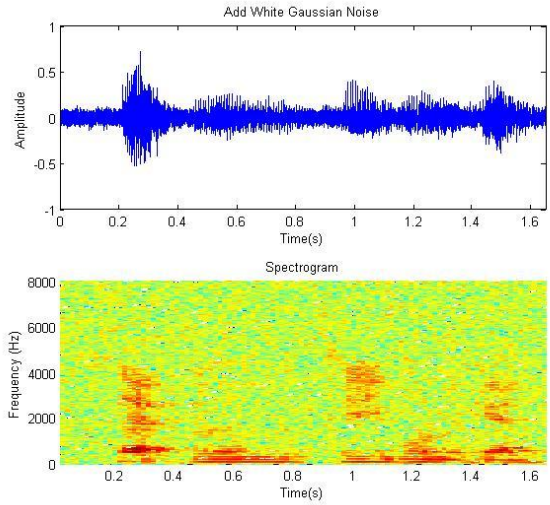

Fig.4 Signal with white Gaussian noise at SNR 5 dB.

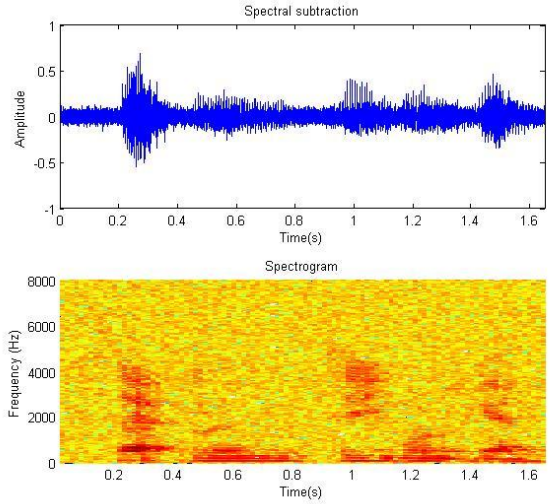

Fig.5 SNR = 5.1594 dB, LLR = 2.4579, SNRseg = 1.0221 dB and PESQ = 2.3328with spectral subtraction denoising method.

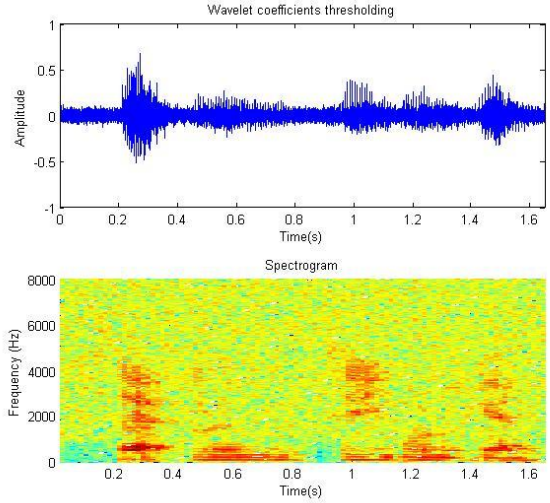

Fig.6 SNR = 5.4057 dB, LLR = 2.7444, SNRseg = 1.1861 dB and PESQ = 2.2804with wavelet coefficients threshold denoising method.

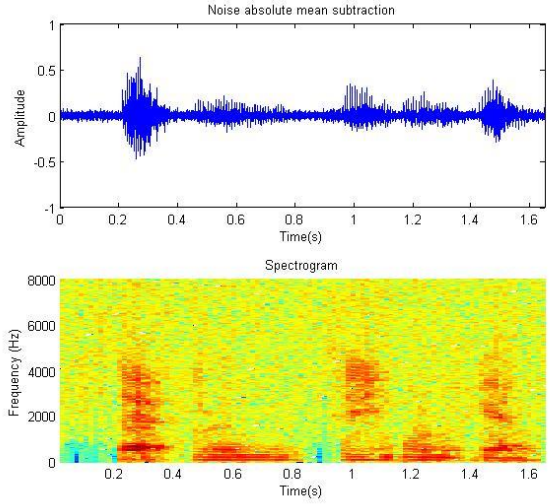

Fig.7 SNR = 7.0433 dB, LLR = 2.6153, SNRseg = 2.8779 dB and PESQ = 2.2457with noise absolute mean subtraction denoising method.

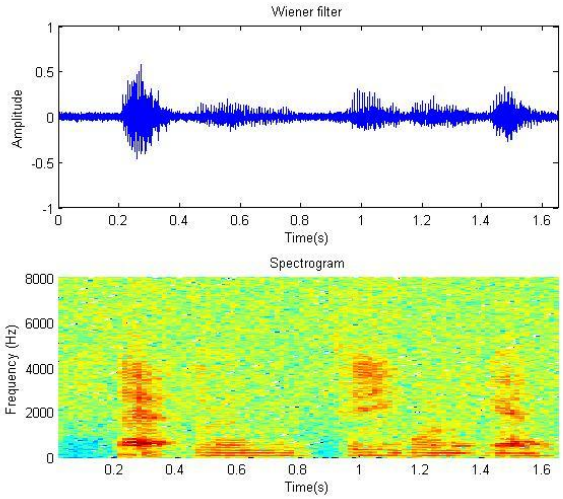

Fig.8 SNR = 7.6251 dB, LLR = 2.5349, SNRseg = 3.4256 dB andPESQ = 2.3865with Wiener filtering denoising method.

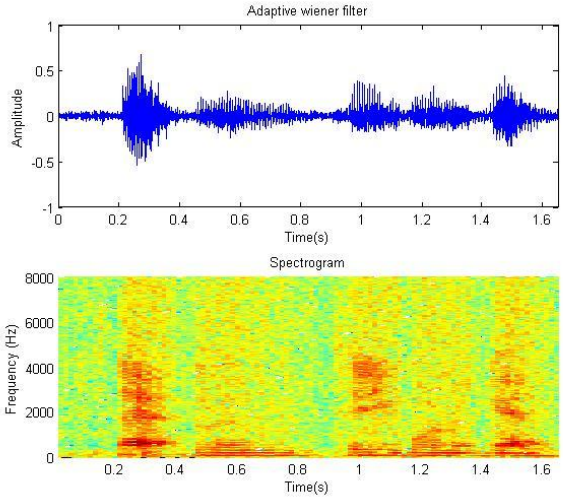

Fig.9 SNR = 8.0673 dB, LLR = 2.3177, SNRseg = 3.9122 dB, PESQ = 2.3170with adaptive Wiener filtering denoising method.

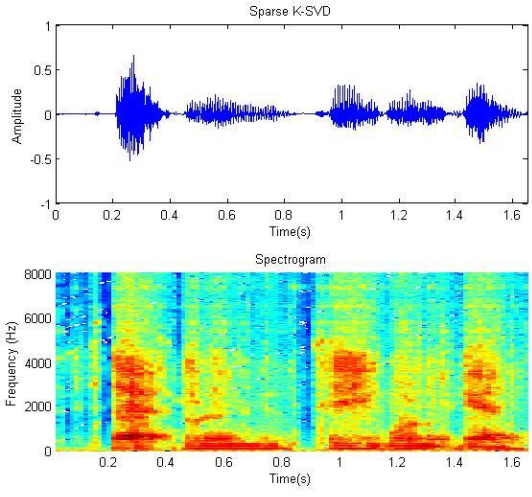

Fig.10 SNR = 11.1045 dB, LLR = 1.3848, SNRseg = 7.2316 dB and PESQ =

2.5473with sparse K-SVD denoising method.

	Noise signal	Spectral subtraction	wavelet coefficients thresholding	noise absolute mean subtraction	Wiener filtering	Adaptive Wiener filtering	Sparse K-SVD
SNR	5.1271	5.1594	5.4057	7.0433	7.6251	8.0673	**11.1045**
LLR	2.4579	2.4579	2.7444	2.6153	2.5349	2.3177	**1.3848**
PESQ	2.3295	2.3328	2.2804	2.2457	2.3865	2.3170	**2.5473**
SNRseg	1.0220	1.0221	1.1861	2.8779	3.4256	3.9122	**7.2316**

Table 1 The compared results with five other methods under four objective quality measures at SNR 5dB level.

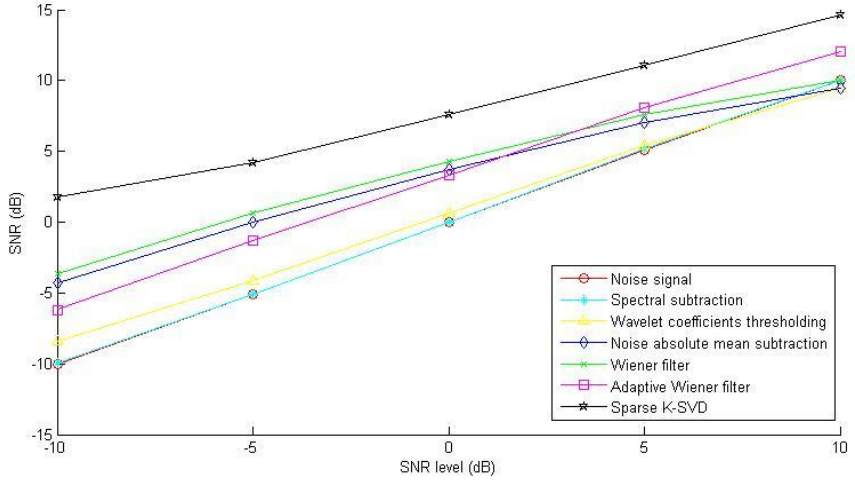

Fig.11 SNR assessment results with five other enhancement methods at SNR levels of -10, -5, 0, 5 and 10 dB.

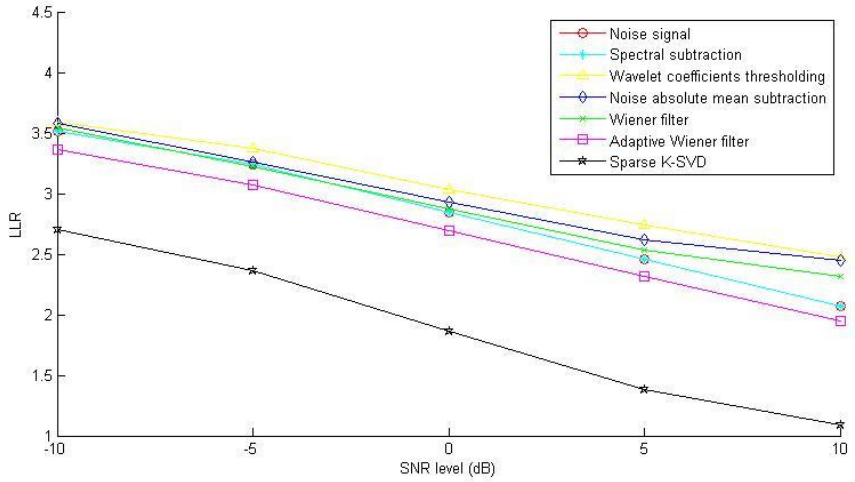

Fig. 12 LLR assessment results with five other enhancement methods at SNR levels of -10, -5, 0, 5 and 10 dB.

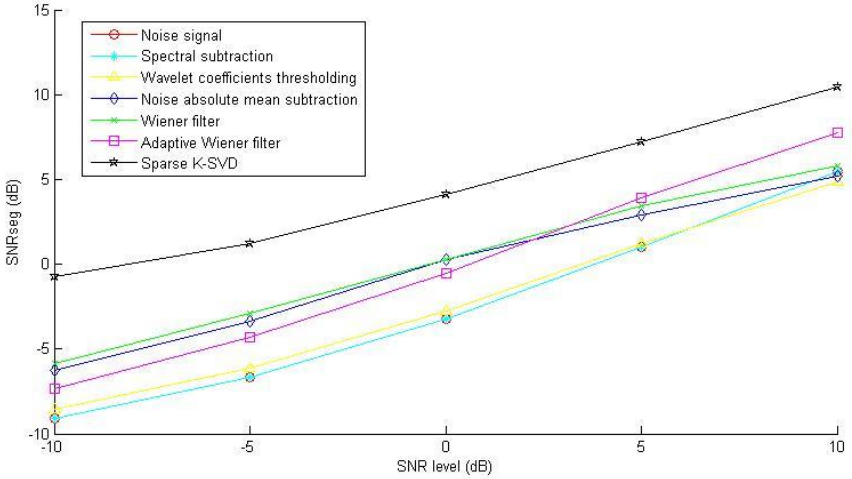

Fig. 13 SNRseg assessment results with five other enhancement methods at SNR levels of -10, -5, 0, 5 and 10 dB.

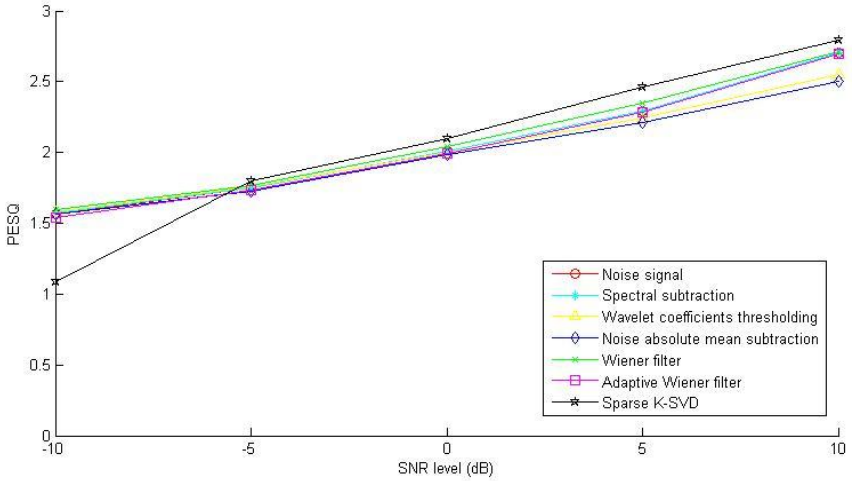

Fig. 14 PESQ assessment results with five other enhancement methods at SNR levels of -10, -5, 0, 5 and 10 dB.

4. Conclusion

In this paper, we proposed sparse representation for speech signal denoising. Training clean dictionary of each input noise signal based on K-SVD algorithm. Use OMP algorithm to find the best sparse coefficients for clean dictionary. In experimental results, we also show that our method has better effect on speech denoising. In the future work, we will use Jafari, M. G. and Plumbley, M. D. [13] method to reduce the computation time on the sparse coding. The proposed method will also be applied to color noise environment and the speech recognition.

Acknowledgments: This work was supported by the National Science Council under grant number MOST 103-2632-E-032 -001-MY3 and MOST 103-2410-H-032 -052.

5. REFERENCES

[1] Marwa, A. Abd, El-Fattah., Moawad, I., Dessouky, Alaa, M., Abbas, Salaheldin, M., Diab, El-Sayed, M., El-Rabaie, Waleed, Al-Nuaimy, Saleh, A., Alshebeili, Fathi, E., Abd and El-samie. "Speech enhancement with an adaptive Wiener filter," International Journal of Speech Technology, 17(1), pp. 53-64, 2014.

[2] Tanabe, N., urukawa, T., Matsue, H. and Tsujii, S. "Kalman Filter for Robust Noise Suppression in White and Colored Noises," IEEE International Symposium on Circuits and Systems. ISCAS. pp. 1172-1175, 2008.

[3] M. S. Lewicki and T. J. Sejnowski. "Learning overcomplete representations," Neural Comput., vol. 12, pp. 337-365, 2000.

[4] Aharon, M., Elad, M. and Bruckstein, A. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," IEEE Transactions on Signal Processing, 54(11), pp. 4311-4322, 2006.

[5] R. Gribonval and K. Schnass. "Some recovery conditions for basis learning by L1-minimization," in Proc. Int. Symp. Commun., Control, Signal Process. (ISCCSP), pp. 768-733, 2008.

[6] Michael, E. and Michal, A. "Image DenoisingVia Sparse and Redundant Representations Over Learning Dictionaries," IEEE Transactions on Image Processing, 15(12), pp.3736-3745, 2006.

[7] Zhimin, X. and Yuantao, G. "Adaptive Speech Enhancement using Sparse Prior Information," IEEE Intemational Conference on Acoustics, Speech and Signal Processing, pp. 7025-7029, 2013.

[8] Hu, Y. and Loizou, P. "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Speech and Audio Processing, 16(1), pp. 229-238, 2008.

[9] Ma, J., Hu, Y. and Loizou, P. "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", Journal of the Acoustical Society of America,125(5), pp. 3387-3405, 2009.

[10] Bahoura, M., Rouat, Jean. "Wavelet Speech Enhancement Based on the Teager Energy Operator," IEEE Signal Processing Letters, 8(1), pp. 10-12, 2001.

[11] J. F. Wang, S. H. Chen.and J. J. Lee. "Speech Signal Denoising Based on Multi-Type Wavelet Transforms," Asia Pacific Conference on Multimedia Technology and Applications, pp. 287-291, 2000.

[12] Barker, J., Vincent, E., Ma, N., Christensen, C. and Green, P. The PASCAL CHiME speech separation and recognition challenge. Computer Speech and Language..

[13] Jafari, M. G. and Plumbley, M. D. "Fast Dictionary Learning for Sparse Representations of Speech Signals,"IEEE jouranl of selected topics in signal processing, 5(5), PP. 1025-1031, 2011.