©2016 Published in 4th International Symposium on Innovative Technologies in Engineering and Science 3-5 November 2016 (ISITES2016 Alanya/Antalya - Turkey)

Subspace based Speech Enhancement using Common Vector Approach

[1]Mehmet Hakan DURAK, [2]Erol SEKE, [3]Kemal OZKAN
[1]Faculty of Engineering, Department of Electrical-Electronics Engineering Gazi University, Ankara, Turkey
[2]Faculty of Engineering and Architecture, Department of Electrical-Electronics Engineering Eskisehir Osmangazi University, Eskisehir, Turkey
[3]Faculty of Engineering and Architecture, Department of Computer Engineering Eskisehir Osmangazi University, Eskisehir, Turkey

Abstract

In this paper, we propose a new speech enhancement method using the common vector approach. Common vector approach is a subspace method used in recognition applications. In the proposed method, we separate the noisy speech data into magnitude and phase in frequency domain. And also magnitude data is separated into common and difference parts using common vector. It is considered that difference part contains the noise. Therefore, this part is cleaned using Linear Minimum Mean Square Error Estimation. After this process, the magnitude data is reconstructed by combining common part. The frequency domain speech data is rebuilt by sum of the reconstructed magnitude data and kept phase data and transform to time domain on each frame. The proposed method was evaluated under various noise conditions. The results are compared with several methods in well-known quality measures.

Key words: speech, enhancement, cva, subspace based, frequency domain

1. Introduction

Speech enhancement applications such as mobile phone apps, speech recognition and hearing aids, voice automated systems, intelligent homes is more important day after day. The performance of such applications is dependent on how much the noise is removed and how much time is cost this process. These applications aim to improve speech quality, speaker's voice intelligibility or do both of them, carrying out that with minimal loss in signal energy. During the last decades, many subspace based approaches have been proposed to this problem, such as Singular Value Decomposition (SVD) (Dendrinos et al., 1991; Jensen et al.), Karhunen-Loéve Transform (KLT) (Ephraim and Van Trees, 1995; Mittal et al., 2000; Rezayee et al., 2001)

Subspace based methods depend on the assumption that the noisy data can be distributed into two or more components. A Singular Value Decomposition (SVD) based approach, proposed by Dendrinos et al. [1] constructed cleaned signal from singular vectors corresponding the largest singular value. In this method; it is believed voice and noise are in the largest and smallest singular vectors, respectively. This technique is developed by Jensen et al. [2] for colored noise on which the former method failed to reduce. Furthermore, their method with high computational complexity had several constraints for controlling residual noise. Ephraim et al. [3] designed to

*Corresponding author:Mehmet Hakan DURAK Address: Faculty of Engineering, Department of Electrical-Electronics Engineering Gazi University, Ankara TURKEY. E-mail address: mhdurak@gazi.edu.tr, Phone: +903125823311

optimize the estimator that minimizes distortion caused by residual noise. Noisy signal is decomposed into noise and signal subspaces using Karhunen Loeve Transform (KLT). Then, the components in noise subspace is zeroing and the signal subspace is restructuring using a gain function. Components in subspaces are merged again to obtain denoised signal through inverse KLT. Mittal et al. [4] and Rezayee et al. [5] developed this work for colored noise. They obtained better results by using different KLT matrices and converging covariance matrix of the noise vectors to a diagonal matrix respectively.

Common Vector Approach (CVA) is a subspace method used in recognition applications. In CVA, training data representing each subject to be discriminated are used to form its own class. In a speech recognition application, environment noise, ages and genders of speakers result in differences in a class [6]. CVA is depend on the common component of those, basically by eliminating these differences in the class. This component is called the common vector.

In this paper, we first describe Common Vector Approach (CVA), the proposed method based on this approach. Then, the experimental results of the proposed algorithm to some noisy speech signals are also reported. Finally, the proposed algorithm is compared with some well-known speech enhancement algorithms.

2. Common Vector Approach

In a group of vectors (class) each vector is assumed to be composed of a common and a difference vectors as

$$a_i = a_{i,diff} + a_{comm}\,,\ i = 1, 2, ..., m \tag{1}$$

where a_{comm} is common to all vectors in the class. Let n be the dimension of the vectors. The case where $m > n$ is called the sufficient case, for which the common vector is the mean vector. The insufficient case where $n \geq m$ is more common in many applications since the number of feature vectors are less than the dimension. For example, when dealing with image blocks, blocks consists of many pixels with insufficient number blocks in hand. Speech recognition and denoising approaches comprise such setups.
In the insufficient case, common vector of a class can be found as follows; $n - m + 1$ eigenvalues of the covariance matrix

$$\Phi = \sum_{i=1}^{m} (a_i - a_{avg})(a_i - a_{avg})^T \tag{2}$$

will be zero. Here, $a_{avg} = \frac{1}{m} \sum_{i=1}^{m} a_i$ is the average of the class members. The eigenvectors that correspond these zero eigenvalues and the remaining $n+1$ nonzero eigenvectors span the indifference $B^{\perp}$ and the difference B subspaces respectively. As implied by the notation, $B^{\perp}$ and B are orthogonal. The common component of a member vector a_i is the projection of it onto $B^{\perp}$ and is identical for all member vectors. Let u_j be the eigenvectors corresponding to zero eigenvalues. The projection matrix can be calculated using

$$P^{\perp} = \sum_{j=1}^{n-m+1} u_j u_j^T \,. \tag{3}$$

It follows that

$$a_{com} = P^{\perp} a_i \tag{4}$$

where a_i is any feature vector in the class. The difference component of a_i is then

$$a_{i,diff} = a_i - a_{comm} \,. \tag{5}$$

Using CVA, common and difference components can be separated as described and different operations can be applied on them before combining them back together.
Other subspace methods cannot handle the insufficient case since they require the inverse of the covariance matrix (eg. [7]). CVA does not have this problem.

3. Proposed Algorithm

The noisy speech can be expressed as $x_\eta = x + \eta$, where x is the original noiseless speech and η is the additive white Gaussian noise or colored noise. The speech data before any process is divided into frames, and the frames is illustrated as $a_\eta = a + \eta$. Then, each frames transformed into frequency domain is expressed as $a = a_r + ia_i$. Magnitude and phase data is calculated by $\sqrt{(a_r^2 + a_i^2)}$ and $\arctan(a_i/a_r)$, respectively. Because of considered to the phase data doesn't contain the noise, the class is constructed with neighborhood of each frame's magnitude data based on the Euclidian distance by ranking. In this class, common and difference are created by applying CVA. Insufficient case occurred when applying CVA and common, difference components are calculated as mentioned in the chapter 2. Assuming that difference component consists largely noise, the noise is extracted from the difference component with the Principal Component Analysis (PCA). The frequency domain speech frame is rebuilded by sum of the reconstructed magnitude data and kept phase data and transform to time domain. After applying this process on each frame, enhanced speech data is created by combining all frames. General scheme of the algorithm is given in figure 1.

During the processes, there are several parameters that can be tried for the best performance on the speech data such as number of frames in classes, frame size, overlap ratio. But best parameters are determined for a large speech database in order to avoid parameters for general usage.

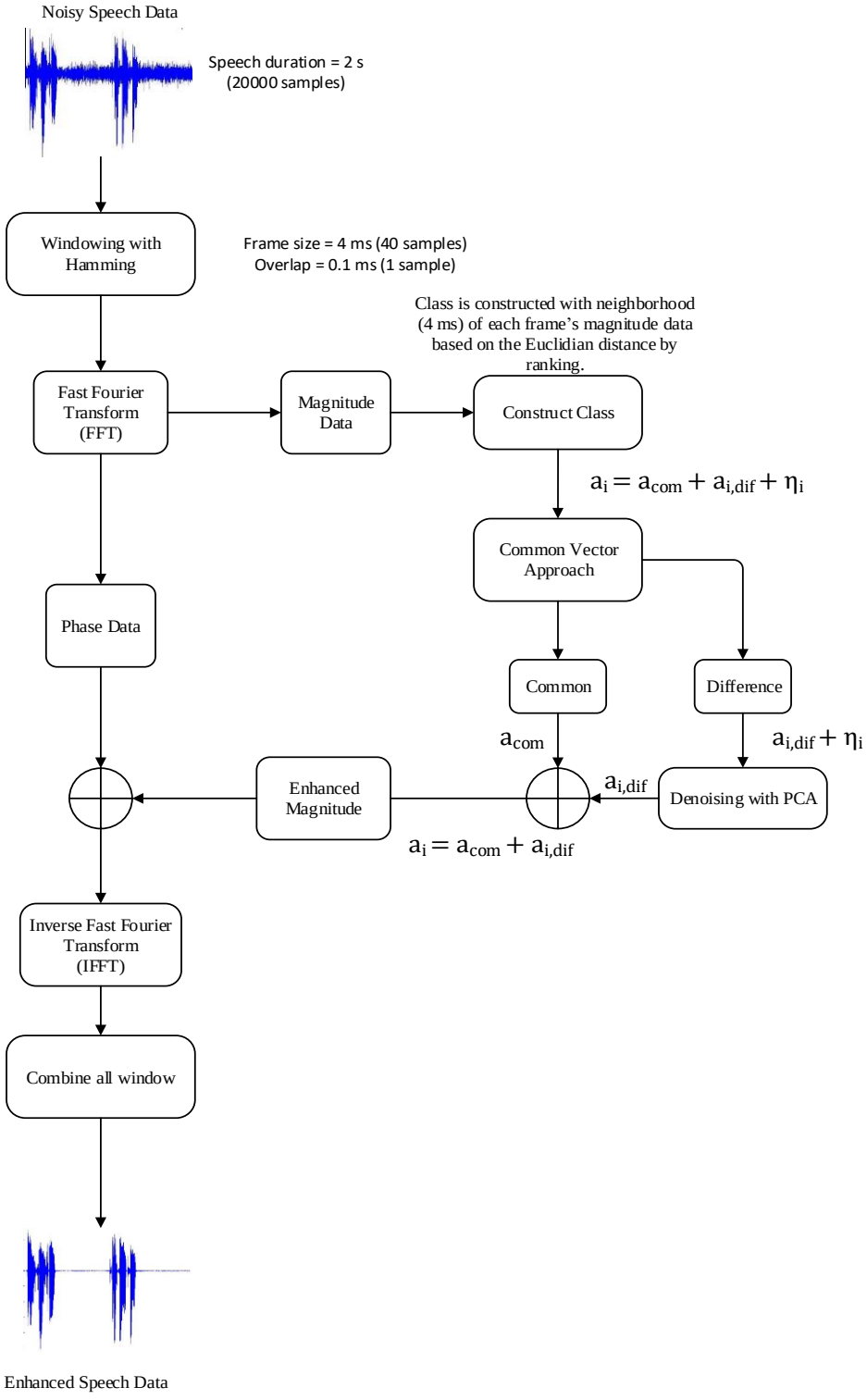

Figure 1. Flowchart of the proposed algorithm

4. Experiments

The proposed algorithm has been tested on the 30 English sentences of the NOISEUS speech database (Hu, and Loizou, 2007). The sentences is about 2 s with sampling rate of 8 kHz and spoken by 3 male and 3 female speakers. 9 different noise types (airport, crowd, car, exhibition hall, restaurant, train station, street, train, AWGN) and 4 different noise levels (0 dB, 5 dB, 10 dB, 15 dB) are used for analysis of performance of the proposed method. The results are compared with 3 different methods. These methods are psychoacoustically motivated statistical method (*stat*) [15], wiener filtering algorithm based on wavelet thresholding multi-taper spectra (*wien*) [16] and continuous spectral tracking (*spec*) [17]. Performance measures used in comparisons are Signal to Noise Ratio (SNR), frequency weighted segmental SNR (fwSNRseg), Itakuro-Saito distance (IS).

1) Signal / Noise Ratio (SNR)

The most common method used to measure the signal quality is signal/noise ratio (SNR). SNR is found by calculating the signal power divided by the noise power in decibel.

$$SNR_{dB} = 10\log\left(\frac{\sum_n s^2(n)}{\sum_n [s(n) - \hat{s}(n)]^2}\right) \tag{5}$$

$s(n)$: Clean signal $\hat{s}(n)$: Enhanced signal

Table 1 gives the results for CVA along with stat [15], wien [16], spec [17] in SNR measure. Best SNR values are marked with boldfaced characters.

Table 1. Comparison of 4 methods on SNR measure

SNR		*stat*	*wien*	*spec*	*cva*
airport	0 dB	2,81	3,60	2,11	**3,60**
	5 dB	6,52	7,37	6,82	**7,52**
	10 dB	10,34	11,35	10,96	**11,86**
	15 dB	14,25	15,54	14,35	**16,44**
crowd	0 dB	3,07	**3,97**	1,97	3,57
	5 dB	6,42	7,41	6,71	**7,61**
	10 dB	10,62	11,53	11,12	**11,99**
	15 dB	14,33	15,41	14,29	**16,32**
car	0 dB	3,65	4,03	3,94	**4,86**
	5 dB	6,63	7,76	8,13	**8,60**
	10 dB	10,15	11,71	11,91	**12,65**
	15 dB	14,25	15,95	14,80	**17,11**
ex. hall	0 dB	2,60	**3,68**	2,22	3,47
	5 dB	6,46	**7,81**	6,78	7,54
	10 dB	10,69	**12,02**	11,06	11,98
	15 dB	14,67	15,92	14,34	**16,74**
restrnt	0 dB	2,23	**2,94**	1,20	2,39
	5 dB	5,91	**6,74**	6,03	6,60
	10 dB	10,10	11,01	10,45	**11,23**
	15 dB	14,36	15,32	14,02	**15,97**
station	0 dB	3,08	3,52	3,03	**3,91**
	5 dB	6,24	7,12	7,49	**7,90**
	10 dB	10,45	11,58	11,53	**12,26**
	15 dB	14,05	15,50	14,41	**16,63**
street	0 dB	2,24	3,16	2,46	**3,92**
	5 dB	5,68	6,89	7,30	**7,67**
	10 dB	9,79	10,90	11,26	**12,01**
	15 dB	14,07	15,31	14,42	**16,57**
train	0 dB	1,70	3,72	2,43	3,57
	5 dB	5,88	7,97	7,14	7,72
	10 dB	10,36	12,28	11,32	11,93
	15 dB	14,90	16,02	14,43	**16,53**
white	0 dB	5,13	6,09	5,07	5,91
	5 dB	8,32	9,91	9,20	9,77
	10 dB	11,96	13,97	12,71	**13,77**
	15 dB	15,69	17,39	15,22	**18,07**

2) Frequency weighted segmental SNR (fwSNRseg)

The frequency domain segmental SNR measure is computed as:

$$fwSNRseg = \frac{10}{m}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} B_j \log\left[F^2(m,j)/(F(m,j)-\hat{F}(m,j))^2\right]}{\sum_{j=1}^{K} B_J} \quad (6)$$

Where B_J is the weight of the j^{th} frequency band, K is the number of band, M is the number of all segment, $F(m,j)$ is the filtered magnitude value of the clean signal in the m^{th} segment and j^{th} frequency band, $\hat{F}(m,j)$ is magnitude value of the enhanced signal in the same segment and same band.

Table 2, fwSNRseg values for CVA and 3 other methods are compared with boldfaces indicating the best fwSNRseg value. It is notable that CVA is almost superior to the compared methods for all noise types and levels.

Table 2. Comparison of 4 methods on fwSNRseg measure

fwSNRseg		*stat*	*wien*	*spec*	*cva*
airport	0 dB	4,65	4,11	4,84	**5,05**
	5 dB	6,49	6,04	6,50	**7,03**
	10 dB	9,22	8,65	8,67	**9,71**
	15 dB	11,96	11,91	10,50	**12,69**
crowd	0 dB	4,44	4,10	4,48	**4,94**
	5 dB	6,54	6,08	6,31	**6,99**
	10 dB	9,15	8,76	8,39	**9,61**
	15 dB	11,87	11,93	10,34	**12,73**
car	0 dB	4,47	3,72	4,46	**4,66**
	5 dB	6,23	5,73	6,23	**6,45**
	10 dB	8,45	8,01	8,27	**8,87**
	15 dB	11,20	11,16	10,27	**11,80**
ex. hall	0 dB	**4,92**	4,29	4,74	4,89
	5 dB	**6,78**	6,47	6,49	6,72
	10 dB	8,96	8,77	8,63	**9,24**
	15 dB	11,32	11,38	10,47	**12,18**
restrnt	0 dB	4,84	4,14	4,91	**5,30**
	5 dB	6,93	6,55	6,69	**7,34**
	10 dB	9,43	8,91	8,80	**10,10**
	15 dB	12,32	12,31	10,58	**13,20**
station	0 dB	4,47	3,91	4,64	**4,85**
	5 dB	6,59	6,00	6,44	**6,81**
	10 dB	8,75	8,14	8,39	**9,23**
	15 dB	11,46	11,16	10,34	**12,27**
street	0 dB	4,64	3,90	4,82	**5,08**
	5 dB	6,49	5,88	6,57	**6,83**
	10 dB	9,05	8,48	8,71	**9,54**
	15 dB	11,38	11,21	10,42	**12,36**
train	0 dB	4,92	4,55	4,90	**5,11**
	5 dB	6,36	6,24	6,48	**6,84**
	10 dB	8,71	8,75	8,37	**9,19**
	15 dB	11,42	11,68	10,38	**12,13**
white	0 dB	3,71	3,65	3,77	**4,15**
	5 dB	5,58	5,63	5,38	**5,73**
	10 dB	**7,83**	8,02	7,42	7,77
	15 dB	**10,28**	10,84	9,58	10,34

3) Itakura-Saito Distance Measure (IS)

The Itakura-Saito (IS) distance is a measure of the perceptual difference between an original spectrum and an approximation of that spectrum. The distortion measure is given by,

$$d_{IS}(\boldsymbol{a_p}, \boldsymbol{a_c}) = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\boldsymbol{a_p} R_c \boldsymbol{a_p^T}}{\boldsymbol{a_c} R_c \boldsymbol{a_c^T}}\right) + \log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \quad (7)$$

Where σ_p^2 is speech with linear prediction coefficient vector and σ_c^2 is processed speech coefficient vector which represents the all-pole gains for processed and clean speech.

As shown in Table 3, proposed algorithm is best or close to best according to other methods for all noise types and levels.

Table 3. Comparison of 4 methods on IS distance measure

IS		*stat*	*wien*	*spec*	*cva*
airport	0 dB	43,35	44,96	**2,24**	2,64
	5 dB	36,52	38,35	2,13	**1,87**
	10 dB	32,52	32,71	2,80	**1,28**
	15 dB	25,35	24,35	3,94	**0,95**
crowd	0 dB	44,87	40,28	**2,42**	2,68
	5 dB	38,48	37,00	2,14	**1,93**
	10 dB	29,13	30,58	2,40	**1,30**
	15 dB	23,99	23,99	4,03	**0,89**
car	0 dB	47,04	58,31	**2,37**	2,91
	5 dB	38,29	41,47	**1,98**	2,05
	10 dB	35,92	42,25	2,07	**1,34**
	15 dB	29,60	32,19	3,25	**0,95**
ex. hall	0 dB	43,30	49,29	**2,55**	2,98
	5 dB	33,93	35,98	**2,07**	2,18
	10 dB	31,71	38,03	2,05	**1,46**
	15 dB	31,02	34,29	2,77	**1,01**
restrnt	0 dB	44,74	43,75	**2,40**	2,58
	5 dB	31,86	29,64	2,12	**1,68**
	10 dB	30,33	30,18	2,75	**1,28**
	15 dB	25,34	23,88	4,04	**0,87**
station	0 dB	45,88	48,34	**2,30**	2,49
	5 dB	39,18	43,77	2,03	**1,83**
	10 dB	33,47	36,45	2,46	**1,34**
	15 dB	30,23	30,91	3,85	**1,01**
street	0 dB	47,99	49,95	**2,69**	3,17
	5 dB	38,46	37,36	2,29	**2,08**
	10 dB	28,30	30,21	2,41	**1,47**
	15 dB	27,46	29,08	3,56	**1,11**
train	0 dB	45,14	47,10	**2,89**	3,13
	5 dB	36,14	36,39	2,49	**2,35**
	10 dB	27,50	31,02	2,36	**1,58**
	15 dB	21,58	24,38	2,87	**1,07**
white	0 dB	38,39	59,56	**3,31**	3,36
	5 dB	34,49	53,29	2,69	**2,58**
	10 dB	31,59	45,48	2,39	**1,88**
	15 dB	29,82	38,79	2,65	**1,43**

Conclusions

Most of the tests performed on 30 sentences spoken by 3 male and 3 female speakers with 9 different noise types and 4 noise levels showed that the proposed CVA method is almost best against other 3 methods. That is, this approach can effectively enhance the speech while reducing noise.

References

[1] M. Dendrinos, S. Bakamidis and G. Carayannis, "Speech enhancement from noise: a regenerative approach," Speech Comm., vol. 10 (1), pp. 45-57, 1991.
[2] S. H. Jensen, P. C. Hansen, S. D. Hansen and J. A Sorensen, "Reduction of broadband noise in speech by truncated QSVD," IEEE Trans. Speech Audio Processing, vol. 3, pp. 439-448, 1995.
[3] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech Enhancement," IEEE Trans. on Speech Audio Processing, 3, 251. 166, 1995.
[4] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," IEEE Trans. Speech and Audio Proc., vol. 8 (2), pp. 159-167, 2000.
[5] A. Rezayee and S. Gazor, "An Adaptive KLT Approach for Speech Enhancement," IEEE Trans. Speech and Audio Proc., vol. 9 (2), pp. 87-95, 2001.
[6] S. Ergin, "The Improvement and Recognition of the Noisy Speech Parameters," M. Eng. Thesis, Eskisehir Osmangazi University, Eskisehir, Turkey, 2004, [Online]. Available: http://ulusaltezmerkezi.com/gurultulu-ses-parametrelerinin-iyilestirilmesi-ve-taninmasi/ .
[7] L. Zhang, W. S. Dong, D. Zhang, and G. M. Shi, "Two-stage image denoising by principal component analysis with local pixel grouping," Pattern Recognition, vol. 43, pp. 1531-1549, 2010.
[8] M. B. Gülmezoglu. and A. Barkana, "Text-dependent speaker recognition by using Gram-Schmidt orthogonalization method," Proc. of IASTED Int. Conf. on Sig. Proc. & App., 438-440, 1998.
[9] M. B. Gülmezoglu, V. Dzhafarov, M. Keskin and A. Barkana, "A novel approach to isolated word recognition," IEEE Trans. on Acoustic Speech and Signal Processing, vol. 7(6), pp. 620-628, 1999.
[10] M.B.Gülmezoglu, V. Dzhafarov and A. Barkana, "The Common Vector Approach and its relation to Principal Component Analysis," IEEE Trans. On Speech and Audio Processing, vol. 9(6), pp. 655-662, 2001.
[11] M. B. Gülmezoglu, V. Dzhafarov and A. Barkana, "The Common Vector Approach and its Comparison with Other Subspace Methods in Case of Sufficient Data," Computer Speech and Language, vol. 21, pp. 266-281, 2007.
[12] M. B. Gülmezoglu and S. Ergin, "An Approach for Bearing Fault Detection in Electrical Motors," European Transactions on Electrical Power, vol. 17(6), pp. 628-641, 2007.
[13] S. Günal, S. Ergin, M. B. Gülmezoglu and Ö. N. Gerek, "On Feature Extraction for Spam E-Mail Detection," Lecture Notes in Computer Science, vol. 4105, pp. 635-642, 2006.
[14] K. Özkan and E. Seke, "Image Denoising Using Common Vector
Approach, "IET Image Processing, vol. 9(8), pp. 709-715, 2015.
[15] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," IEEE Signal Processing Letters, vol. 11(2), pp. 270-273, 2004.
[16] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. on Speech and Audio Processing, vol. 12(1), pp. 59-67, 2004.
[17] G. Doblinger, "Computationally Efficient Speech Enhancement By Spectral Minima Tracking in Subbands," Proc. EuroSpeech, vol. 2, p. 1513-1516, 1995.
[18] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Speech Audio Processsing, vol. 16 (1), pp. 229-238, 2008.