JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Spectral Subtraction based Speech De-noising using Adapted Cascaded Median Filter

Dhiraj Nitnaware[1*]

[1]Dept. of Electronics & Telecommunication, Institute of Engineering and Technology, Devi Ahilya University, Indore, India

[*]*Corresponding Author:* dnitnawwre@ietdavv.edu.in *Tel.: +00-12345-54321*

Available online at: www.ijcseonline.org

Accepted: 23/Nov/2018, Published: 30/Nov/2018

***Abstract*—** In this paper, a new method is proposed for improvement of speech which is distorted by acoustic noise. Acoustic noise reduction is done through a proposed post processed adapted cascaded median filter based on spectral subtraction technique. This method use two stages of filter, in which background noise is eliminated by first stage cascaded median filter and then output speech is post processed by second stage adaptive filter, to reduce musical and residual noise. Proposed post processing algorithm is compared to conventional single stage cascaded median filter based on subjective listening tests and perception evaluation of speech quality (PESQ) scores. Simulation is done in Matlab-15 and results show that enhanced speech generated by proposed algorithm has better quality than conventional cascaded median filter.

***Keywords*—** Speech Enhancement, Noise Estimation, Spectral Subtraction, Cascaded Median Filter, Musical Noise

I. INTRODUCTION

Many types of noises are present in environment that is uncontrollable in nature. Speech signals are generally corrupted by environmental noise. There is requirement of a proper speech quality for communication in the fast growing world. So, speech enhancement is a technique of reducing noise that is generated from various types of sources. Various classes of algorithms to enhance speech have been already developed to improve the quality of speech.

Nature of noise may be stationary, pseudo stationary and non- stationary. Many of the existing speech enhancement algorithms were failed to deal with non-stationary nature of noise where noise characteristics change with respect to time. Speech enhancement algorithms are classified on the basis of three categories. Firstly, based on channel, as one channel, two channel and multichannel algorithms. One microphone is used in one channel algorithms whereas array of microphones are used in multi-channel algorithms. Second category is based on type of noise and time-frequency scale falls under third category. Noise estimator is the main block in any speech enhancement technique. This paper is organized as follows: Related works are discussed in section 2, while proposed method is explained in section 3. Section 4 describes the simulation results while conclusion is explained in section 5.

II. RELATED WORK

Boll introduced Spectral Subtraction method in 1979 [1]. The principle used in this method is that speech can be processed and improved by subtracting the estimated noise magnitude spectrum from input noisy speech magnitude spectrum. Phase is left unchanged as phase errors are insensitive to human ears. Conventional spectral subtraction is based on the supposition that speech signal and noise signal are uncorrelated with each other. Noise is generally additive in nature. It is estimated using apposite noise estimation algorithm. Boll's spectral subtraction [1] equation can be written as,

$$|X_w^{\hat{}}(k)|=\begin{cases}|S_w(k)|-|G_w^{\hat{}}(k)|, |S_w(k)|>|G_w^{\hat{}}(k)|\\ 0, \text{otherwise}\end{cases} \quad (1)$$

Where $|G_w^{\hat{}}(k)|$ is estimated noise spectrum, $|X_w^{\hat{}}(k)|$ is estimated enhanced speech spectrum and k is discrete frequency index. But background noise is not eliminated completely due to inaccurate noise estimation and moreover it also suffers from the problem of musical artifacts and residual noise.

Berouti et.al. have proposed modifications to overcome above noise to conventional spectral subtraction (CSS) method [2]. Two parameters namely over subtraction factor α and spectral floor factor β were introduced in Berouti spectral subtraction (BSS) method. Here, enhanced speech can be obtained by over subtracting the estimated noise spectrum by input noisy speech spectrum. Value of α is generally kept larger than 1. This helps in removing background noise effectively and increases the signal to noise ratio (SNR) of output enhanced speech. But at the same time, there will be distortion in some speech parts if α is kept too high which may reduce quality and intelligibility of speech. In CSS, all the negative spectral components are set to zero which removes some parts of speech and results in musical artifacts. Introduction of spectral floor factor β prevents the spectral components of the enhanced speech spectrum to go below the lower bound β|$\hat{G_w}$(k)| thereby reducing musical noise. The mathematical equation describing BSS [2] is written as,

$$|\hat{X_w}(k)| = \begin{cases} [|S_w(k)|^{\gamma} - \alpha|\hat{G_w}(k)|^{\gamma}]^{\frac{1}{\gamma}}, \\ if\ |S_w(k)|^{\gamma} - \alpha|\hat{G_w}(k)|^{\gamma} > \beta|\hat{G_w}(k)|^{\gamma} \\ \beta|\hat{G_w}(k)|^{\gamma},\ otherwise \end{cases} \quad (2)$$

Where, γ is an exponent factor. Magnitude spectral subtraction is obtained at γ =1 and power spectral subtraction is obtained at γ =2. Block diagram of BSS is shown in Figure 1.

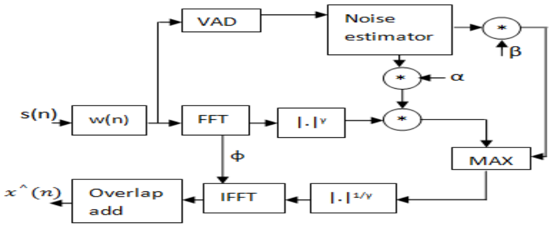


Fig. 1: Spectral Subtraction Method Block Diagram [2]

Noise estimator is the basic building block in any speech enhancement algorithm. Noise has to be eliminated from speech signal to increase SNR but at the same time, shape and characteristics of the speech signal must be retained. The proposed algorithm is also used for noise estimation. Various noise estimation methods were developed by different authors. Some of them are discussed here. Spectral subtraction method and adaptive filters are used to cancel wideband noise [1] while Comb filters are used to separate two mixed speech [2]. Kamath et. al. have explained multi-band spectral subtraction approach to reduce colored noise effect and improve the speech quality [3].

Weighting function introduced in [4] is used to reduce musical noise to some extent. They have also perform listening test to check quality and intelligibility of speech. Martin et. al. have proposed minimum statistics algorithm which tracks non-stationery noisy speech spectrum over fixed window length [5]. Spectral subtraction and wiener filtering in frequency domain approach given by Stahl et. al. They have developed quantile based algorithm to estimates noise spectrum by selecting a proper quantile value from histogram of past frames [6].

The author of paper [7] have developed a noise estimation algorithm in highly non-stationary conditions by noise estimation using time-frequency smoothing factors. Cohen et. al. discussed minima controlled recursive averaging (MCRA) algorithm to estimate noise by finding only in noise regions and comparing noisy speech to local minimum ratio [8]. Cascaded median based noise estimation is given in [9] which improves SNR of speech signal by taking non-stationary noise.

Wavelet based approach on time-frequency domain which is proposed in [10], shows that there is reduction in musical noise but small degradation in the quality of the speech. Y. Hu et. al. have explained noise suppression method to predict the speech quality [11]. Doblinger et. al. presented continuous minima tracking algorithm by tracking minimum of noisy speech to update the noise estimation in every frequency bin [12]. While Hirsch and Ehrlicher [13] has given weighted average technique to estimate noise but fails if there is sudden change in the noise. Ris and Dupont have merged the techniques proposed Doblinger and Hirsch with narrow band spectrum and estimated the noise levels but fails to adopt fast changes in the noise levels [14].

III. PROPOSED METHOD

The proposed method uses two stages of filter to enhance the speech. First stage is cascaded median filter which estimates the background noise spectra and cancel it by using spectral subtraction filter. But resulted enhanced speech has some contents of musical noise which is annoying for listeners. So the output speech is post processed by adaptive filter to reduce musical noise. Block diagram of proposed method is shown in Figure 2.

Let input noisy speech is s(n) that is summation of clean speech x(n) and noise signal g(n).

$$s(n) = x(n) + g(n),\ 0 \le\ n \le N-1 \quad (3)$$

Where, n is discrete time index and N is total number of samples in input speech. Following nine steps are involved for implementation of the proposed algorithm.

3.1 Windowing:

Here the input noisy speech s(n) is multiplied by an appropriate window function with Hamming window w(n). Noise g(n) is generally considered as additive in nature, therefore, input noisy speech can be written as summation of clean speech x(n) and noise g(n) as described by equation (3). Hamming window can be mathematically expressed as:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, otherwise \end{cases} \quad (4)$$

3.2 Fast Fourier Transform:

Time domain speech is converted into frequency domain by using Fourier transform which is described by equation (5) as follows:

$$S_w(k) = X_w(k) + G_w(k) \quad (5)$$

FFT length is chosen as 1024 but computation is done only for 512 bins. Other 512 bins can be calculated directly by using complex conjugate property of FFT for real signals.

$$S_w(k) = |S_w(k)|\, e^{j\angle S_w(k)} \qquad (6)$$

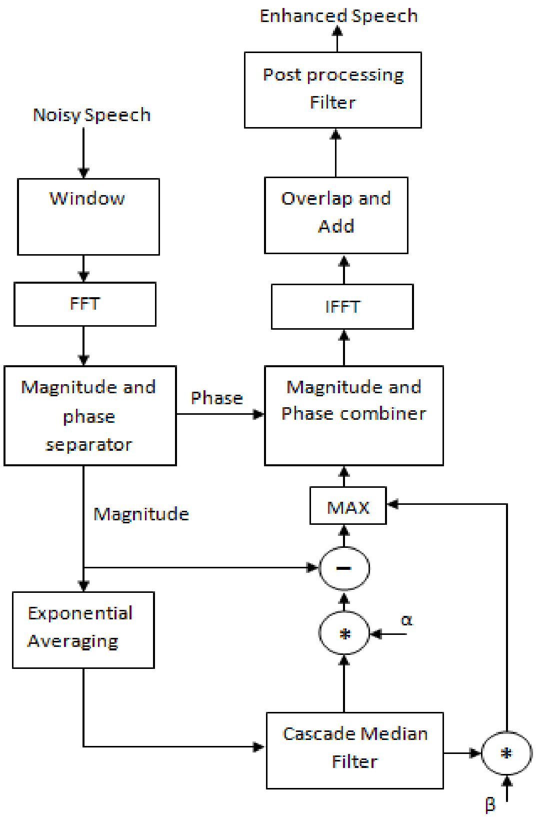


Fig. 2: Block diagram of proposed method

3.3 Magnitude and Phase Separation:

It is assumed that phase errors are insensitive to human ears and does not affect quality and intelligibility of speech. Hence complex noisy spectra $S_w(k)$ can be written in polar form as given in equation (6):

3.4 Exponential Averaging:

Exponential smoothing is done to remove high frequency noise. Exponential average is different from simple moving average. Equal weights are assigned to each past sample in simple moving average whereas exponential averaging assigns exponentially decreasing weights as sample gets older. It can be mathematically described by equation (7) as follows:

$$|Y_w(k)| = \beta'|S_w(k)| + (1-\beta')|Y_w(k-1)| \qquad (7)$$

Where, $|Y_w(k)|$, $|Y_w(k-1)|$ and β' represents exponential average, past exponential average and smoothing factor respectively.

3.5 Noise Estimation:

Cascaded median based noise estimation uses median which are connected in cascade mode with 'r' number of frames in single stage and 'S' being the total number of stages in cascade. Each stage contains two dimensional integer arrays of size r*f, f being the number of samples in each frame. In first stage, frames 'r' of exponential average magnitude noisy spectra are stored which ensemble median, these noisy spectra are calculated at each position within the frame. These median are stored as a row in the next stage i.e. for every 'r' frames in first stage, a row of median is generated in second stage. Process is continued till the final stage i.e. S^{th} stage as shown in figure 3. Hence a total of r^S frames are required for noise estimation. Let $|G_w^*$(k) be the estimate of noise spectrum by cascaded median filter.

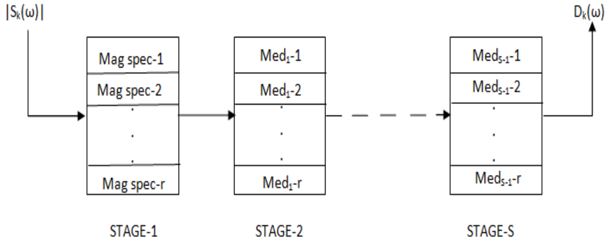


Fig. 3 Cascaded median filter

3.6 Spectral Subtraction:

An estimate of clean speech spectrum $|X_w^*$(k)| can be generated by subtracting the noise estimated magnitude spectra multiplied by over subtraction factor α from original noisy speech magnitude spectrum. Spectral floor β controls the musical noise. Spectral subtraction expression is described in equation (2).

3.7 Complex Spectrum:

The estimated clean speech spectrum is combined with original noisy phase spectrum using magnitude phase combiner (MPC) as follows:

$$X_w^*(k) = |X_w^*(k)|\, e^{j\angle S_w(k)} \qquad (6)$$

3.8 Resynthesized Speech:

An inverse Fourier transform (IFFT) followed by 50% overlap add is used to reconstruct the enhanced speech in time domain. 50% overlap and 75% overlap and add are giving almost same results, that's why 50% overlap and add is selected.

3.9 Post processing filter:

Adaptive filter is used as a post processed filter. Most of the background noise has been eliminated by cascaded median filter. Now fine tuning is done by the adaptive filter and it also eliminates musical noise present in output speech which is generated at the end of step 8. So speech $x^*(n)$ generated after overlap and add is passed through a tenth order adaptive filter whose impulse response coefficients will update until

$x^{\wedge}(n)$ becomes almost close to our desired clean speech signal d(n) as given in figure 4.

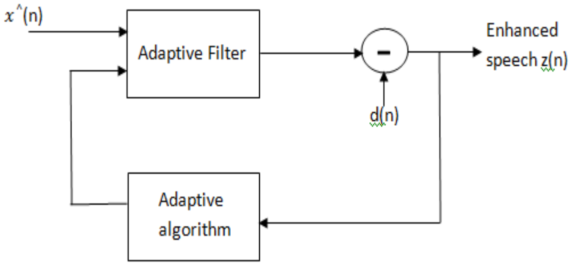


Fig. 4: Post- Processed Filter in Proposed Algorithm

Various adaptive algorithms were applied but NLMS algorithm gives best results. Output of adaptive filter z(n) can be expressed as:

$$z(n) = h(n)^{T} * x^{\wedge}(n) \qquad (8)$$

Where h(n) is the impulse response whose coefficients are updated by the following relation:

$$h(n+1) = h(n) + \frac{\mu x^{\wedge}(n)}{\rho + \|x^{\wedge}(n)\|^{2}} e(n) \qquad (9)$$

Where, μ is step size, ρ is small positive number and e(n) is the error signal which is given as a feedback to adaptive filter and can be expressed as:

$$e(n) = d(n) - z(n) \qquad (10)$$

IV. RESULTS AND DISCUSSION

The parameters used in the proposed algorithm at each steps are listed in table 1. Input noisy speech is taken from Noizeous database of sampling frequency 8 KHz and duration of 2.6 seconds [16].

Table 1: Simulation Parameters

Name of parameter	Description
Window function used	Hamming window of length 256 msec
Fast Fourier transform	FFT of length 1024 bins.
Exponential averaging parameters	Present exponential average of length 512 bins and averaging time of 1 sec
Cascaded median filter	3-frame 6 stage median filter
Over subtraction factor (α)	Varies on the base of type of noise and SNR.
Spectral floor factor (β)	0.001
Overlap add	50% overlap i.e. overlap size of 128 msec
Adaptive filter	Direct form FIR filter
Order of adaptive filter	10
Adaptive algorithm	NLMS
Step size (μ)	0.1
Leakage factor	0.001
Offset	0.0001

Three different arrangement were used to test the performance of the proposed algorithm. Following section explain these tests in detail.

4.1 Time Based Signal and Spectrograms Test:

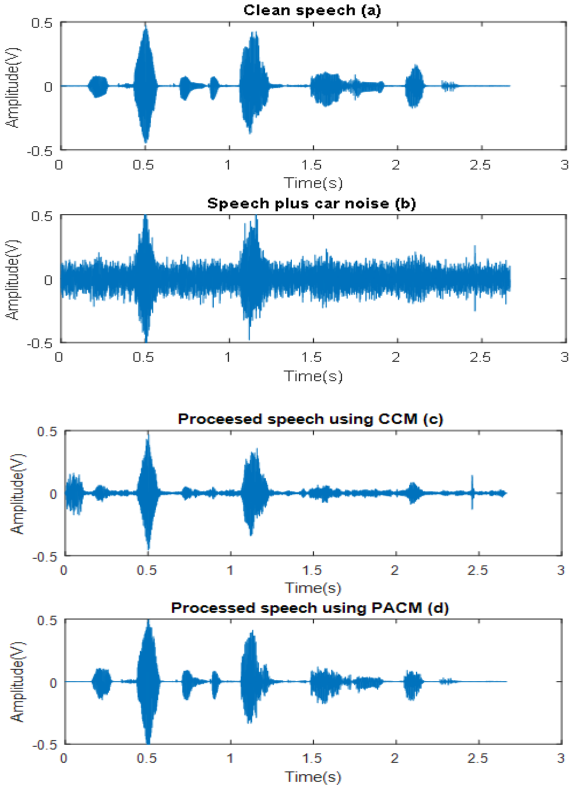
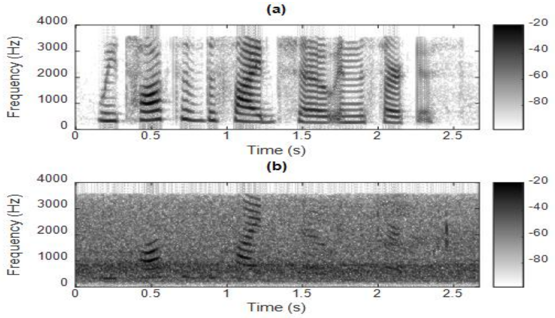


Fig. 5 (a) Clean speech sc-01(b) Noisy speech
(c) Processed speech using CCM and (d) Processed speech using PACM.

Time domain waveforms and spectrograms of speech signals in presence of noise and enhanced speech using conventional cascaded median filter (CCM) and by using proposed adapted cascaded median filter (PACM) are plotted in order to notice the reduction in strength of noise bursts. Two clean speech files named as sc-01 and sc-02 with text 'we talked of the sideshow in the circus' and 'the stray cat give birth to kittens' respectively are taken. Fig. 5 and 6 shows time domain signals and spectrograms respectively for speech, added with car noise at input SNR of 0 dB.



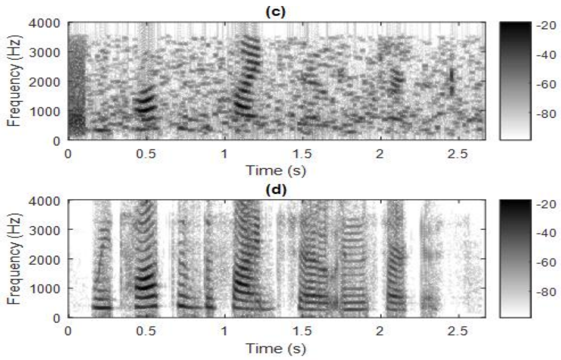


Fig. 6 Spectrograms of (a) Clean speech sc-01(b) Noisy speech (c) Processed speech using CCM and (d) Processed speech using PACM.

The result shows that time domain signal and spectrograms of proposed method PACM is almost equal to clean speech. Similarly Fig. 7 and 8 gives the signal and spectrograms for speech that is added with airport noise at input SNR of 5 dB. The same results are observed.

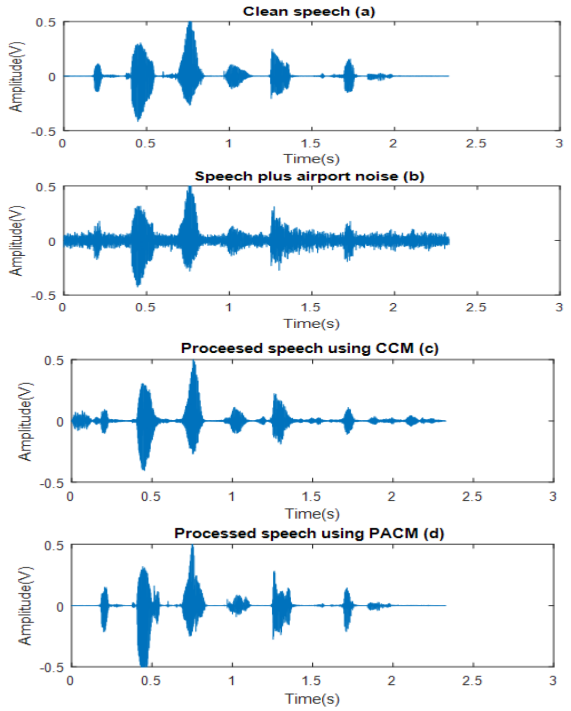
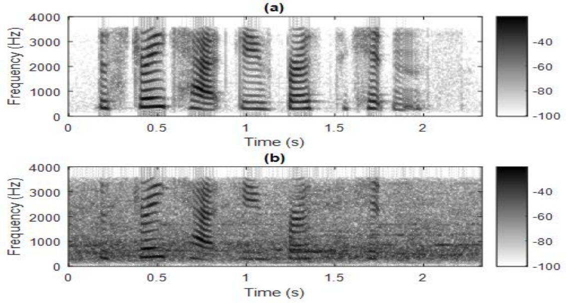


Fig. 7 (a) clean speech sc-02 (b) Noisy (airport) speech (c)) processed speech using CCM (d) processed speech using PACM.

4.2 Listening Test:



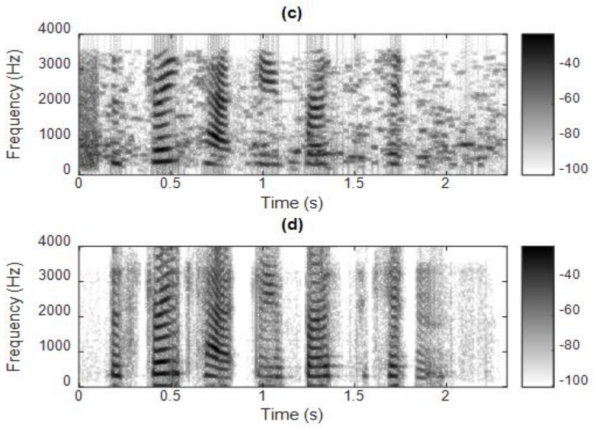


Fig. 8 Spectrograms of (a) Clean speech sc-02 (b) Noisy (airport) speech (c)) Processed speech using CCM (d) Processed speech using PACM

Listening tests involve hearing of clean speech, unprocessed speech, processed speech using conventional cascaded median based noise estimation and processed speech using proposed post processed algorithm. Ten to fifteen listeners including both male and female listeners were invited to check the quality and intelligibility of speech.

The listeners were asked to write down the text corresponding to speech. It is found that there is a problem in writing down text version of processed speech using CCM noise estimation because of the musical artifacts. Only Five listeners were able to write down the text correctly. Proposed method shows better performance over CCM noise estimation. More than 10 out of 15 listeners were able to write correctly the text corresponding to speech.

4.3 Objective Measure Test:

Table 2: PESQ score with SNR at 0 Db

S.no	Alpha (α) optimal	Type of Noise	PESQ Score of Noisy Speech	PESQ Score of enhanced speech	
				Using CCM	Using PACM
1	2	CAR	1.477	2.062	2.451
2	2.25	TRAIN	1.543	2.036	2.386
3	2.5	AIRPORT	1.820	1.969	2.556
4	2.4	STATION	1.388	1.95.	2.222
5	2.4	STREET	1.539	2.132	2.552
6	2.4	RESTAURANT	1.352	1.606	2.390
7	2.4	BABBLE	1.494	1.863	2.476

There are several objective measures for assessment of quality of speech. Perception Evaluation of Speech Quality (PESQ) score is calculated to compare the quality of speech generated by proposed method with conventional cascaded median filter.

Table 3: PESQ score with SNR at 5 dB

S.no	Alpha (α) optimal	Type of Noise	PESQ Score of Noisy Speech	PESQ Score of enhanced speech	
				Using CCM	Using PACM
1	1.8	CAR	1.819	2.273	2.564
2	2	TRAIN	1.779	2.257	2.557
3	2.2	AIRPORT	1.860	2.238	2.661
4	1.8	STATION	1.785	2.255	2.544
5	1.8	STREET	1.866	2.319	2.418
6	1.8	RESTAURANT	1.859	2.036	2.595
7	1.8	BABBLE	1.882	2.231	2.542

Table 4: PESQ score with SNR at 10 dB

S.no	Alpha (α) optimal	Type of Noise	PESQ Score of Noisy Speech	PESQ Score of enhanced speech	
				Using CCM	Using PACM
1	1.5	CAR	2.120	2.525	2.651
2	1.7	TRAIN	2.130	2.235	2.640
3	1.9	AIRPORT	2.230	2.512	2.773
4	1.5	STATION	2.213	2.513	2.672
5	1.95	STREET	2.172	2.622	2.723
6	1.30	RESTAURANT	2.323	2.452	2.770
7	1.6	BABBLE	2.200	2.392	2.672

Table 5: PESQ score with SNR at 15 dB

S.no	Alpha (α) optimal	Type of Noise	PESQ Score of Noisy Speech	PESQ Score of enhanced speech	
				Using CCM	Using PACM
1	1.4	CAR	2.354	2.707	2.715
2	1.2	TRAIN	2.450	2.643	2.766
3	1.3	AIRPORT	2.420	2.705	2.744
4	1.3	STATION	2.380	2.660	2.786
5	1.3	STREET	2.441	2.609	2.761
6	1.6	RESTAURANT	2.417	2.660	2.725
7	1.6	BABBLE	2.418	2.662	2.730

Only one way signal distortion is computed in this measure while two way interactions are neglected. Table 2, 3, 4 and 5 shows PESQ scores of unprocessed and processed speech at SNR of 0, 5, 10 and 15 dB respectively. Speech with different types of noise such as car, airport, station, street, restaurant and babble are considered for analysis. PESQ score of PACM showing greater improvement with all types of noise over CCM method.

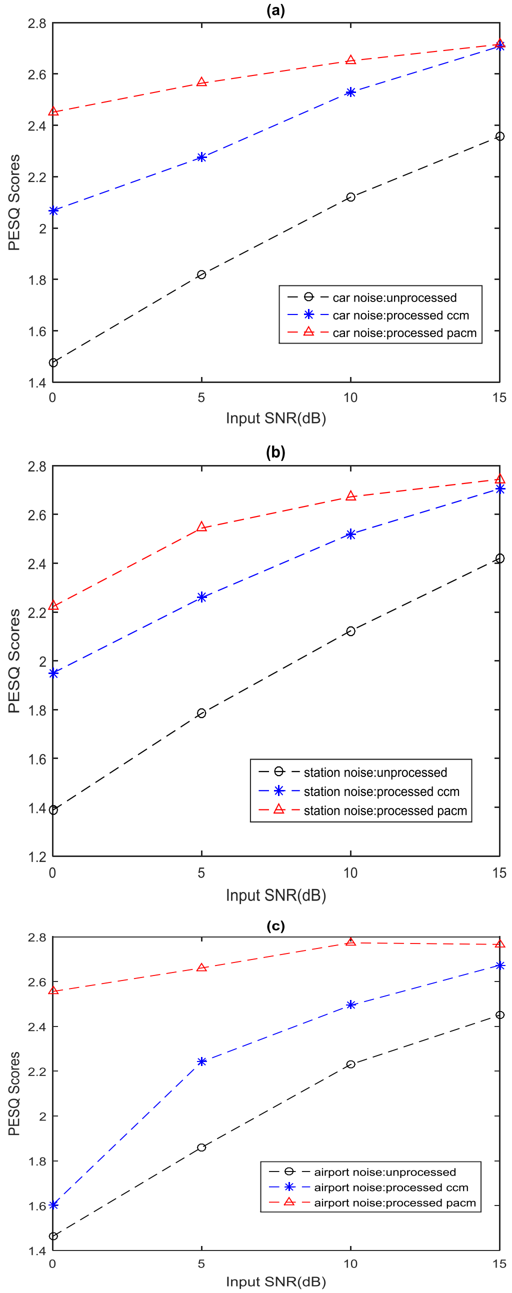


Fig. 9: PESQ scores Vs input SNR for (a) Car noise (b) Station noise (c) Airport noise.

Figure 9(a), (b) and (c) shows a plot of PESQ scores against input SNR for car, station and airport noises. The post processed proposed method is also compared with conventional cascaded median based noise estimation using subjective listening tests. PESQ scores of unprocessed speech increases with the increase in input SNR. PESQ scores in all the cases are 2.5 or close to 2.5 which is considered as a good score. Large value of α is required for high SNR and vice-versa for low SNR. PESQ scores of proposed method are correspondingly higher than conventional cascaded median filter in all the cases. There is a large improvement in PESQ scores of proposed method over conventional CCM filter at low SNR and small improvement at high SNR.

V. Conclusion and Future Scope

Based on the simulation, conclusion can be drawn that the proposed adapted cascaded median filter based on spectral subtraction method algorithm removes musical as well as background noise. Proposed method shows an improvement in PESQ scores by 15-55% for different types of noises over conventional cascaded median filter. Proposed post processed algorithm can also be applied to other speech enhancement algorithms. Real time implementation of this algorithm is yet to be carried out. Experiments with other databases will also be carried out in future to arrive at more conclusive statements about the advantages of the proposed method. Computational complexity of proposed algorithm is higher than CCM filter. So, several steps shall be taken in future to reduce the computational complexity of the proposed method.

References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transaction Acoustic, Speech, Signal Process., vol.27, no. 2, pp. 113-120, 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", In the Processing of IEEE International Conference of Acoustic, Speech and Signal Processing, pp. 208-211, 1979.

[3] S. Kamath and P. Loizou, "A multiband spectral subtraction method for enhancing speech", In the Processing of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Volume 4, pp IV-4164, 2002.

[4] Ekaterina Verteletskaya and Boris Simak, "Noise reduction based on modified spectral subtraction method", IAENG International Journal of Computer Science, 38:1, IJCS_38_1_10, 2011.

[5] Martin, R, "Spectral subtraction based on minimum statistics", In the Processing of Eur. Signal Process. Conf., pp. 1182-1185, 1994.

[6] V. Stahl, A. Fisher, and R. Bipus, "Quantile based noise estimation for spectral subtraction and wiener filtering", In the Processing of ICASSP, pp. 1875-1878, 2000.

[7] S. Rangachari, and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments", Speech Communication., vol. 48, pp. 220-231, 2006.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging", IEEE Trans. Speech Audio Process., vol. 11, no. 5, pp. 466-475, 2003.

[9] Santosh K. Waddi, Prem C. Pandey, and Nitya Tiwari, "Speech Enhancement Using Spectral Subtraction and Cascaded-Median Based Noise Estimation for Hearing Impaired Listeners", NCC, Department of Electrical Engineering Indian Institute of Technology Bombay, pp 1 - 5 , 2013.

[10] Kun-Ching Wang, "Wavelet-based speech enhancement using time-frequency adaptation", EURASIP Journal on Advances in Signal Processing, 2009.

[11] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Transactions on Audio, Speech, and Language Processing, volume. 16, pp. 229-238, 2008.

[12] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands", In the Processing of IEEE International Conference on Eurospeech, vol 2, pp 1513–1516, 1995.

[13] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition", In the Processing of IEEE International Conference of IEEE International Conf. on Acoustic, Speech Signal Process., pp 153–156, 1995.

[14] C. Ris and S. Dupont, "Assessing local noise level estimation methods: application to noise robust ASR", IEEE Transaction on Speech Comm. Vol 34, pp 141–158, 2001.

[15] B. Widrow and S. D. Stream, Adaptive Signal processing, New York: Prentice-Hall, 1985.

Authors Profile

Dr. Dhiraj Nitnawre is working as Assistant Professor in E & TC Dept. at IET, DAVV, Indore (M.P.). He received his B.E. degree from GEC Ujjain in 2000 and M.Tech in 2003 in Instrumentation from Devi Ahilya University, Indore. He has also completed Ph.D degree in 2011 from Devi Ahilya University. He has published more than 25 research papers in reputed international journals including and 20 conferences including IEEE and it's also available online. . He has 15 years of teaching experience and 10 years of Research Experience. His research area includes Wireless Adhoc and Sensor Network, Wireless Communication and Image Processing.