ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 4, Issue 12, December 2015

Speech Enhancement Using Basic and Modified Spectral Subtraction

Gitu Geevarghese, Milind Shah

***Abstract*— In this paper, a noisy speech enhancement method based on basic and modified spectral subtraction performed on short time magnitude spectrum is presented. The first method involves subtracting an estimate of the noise spectrum from the noisy speech spectrum and setting the negative differences to zero. This method reduces the broadband noise however introduces a new type of noise known as musical noise. Hence, a modified spectral subtraction method is reported which eliminates the musical noise as well as the background noise. In both the method, the initial silence frames are taken as an estimate of noise for subtraction. The noise reduced magnitude spectrum is then recombined with the unchanged phase spectrum to produce a modified complex spectrum enhanced speech with overlap add method. Noise was generated based on AURORA database and was added in clean speech simulating noisy environment. Both the algorithms were applied on noisy database for performance evaluation. Investigations were carried out for obtaining optimum values of the parameters in the algorithm such as window length, noise estimation duration and spectral subtraction parameters based on objective evaluation test using PESQ score.**

***Index Terms*— Spectral subtraction, over-subtraction, spectral floor parameter, PESQ score.**

I. INTRODUCTION

Speech is a natural and basic way for humans to convey messages to each other. However, many times speech gets corrupted as it has to be recorded in the presence of undesirable background noise. Normal persons and Persons with sensorineural impairment experience great difficulty in speech perception in noisy environments. The solution to this problem is to use a noise suppression technique so as to improve the speech quality and intelligibility. Speech enhancement is a technique which is used to recover back the original signal from its corrupted version. In hearing aid applications, the speech enhancement technique has been employed mainly for reducing the additive background noise. In order to overcome the background noise, various methods have been employed for increasing the speech quality of hearing aid application. They are classified on the basis of the type of application as single channel speech enhancement, multi-channel enhancement and model based speech enhancement [1]. Spectral subtraction is one of the traditional methods that have been implemented for single channel speech enhancement [1]. The spectral subtraction offers a simple and efficient tool for the suppression of an additive noise in a speech signal [2]. The key idea of spectral subtraction is to estimate back-ground noise and then to subtract this estimate from the noisy speech signal.

In this report two spectral subtraction methods are being reported, a basic spectral subtraction method which involves windowing, magnitude subtraction using FFT, half-wave rectification and obtained enhanced speech with overlap add method. And another modified spectral subtraction method which differs from the previous method in two major ways: an oversubtraction factor α which varies from frame to frame and spectral floor parameter β which prevents the spectral floor component from going below the lower bound. The objective of this research work is implementation of the above algorithm and to carry out the investigations for obtaining optimum values of the parameters such as window length, noise estimation duration, and spectral subtraction parameters based on an objective evaluation test using Perceptual Evaluation of Speech Quality (PESQ) score [16]. The complete implementation and analysis of the basic spectral subtraction and modified spectral subtraction method were carried out to select the optimal set of parameter values for noise suppression.

The paper is divided into five sections, sections II and III explaining the basic and modified spectral subtraction algorithm that were implemented, section IV discusses the results obtained followed by conclusion drawn reported in section V.

II. BASIC SPECTRAL SUBTRACTION (BSS) METHOD

Spectral subtraction is a single-input noise reduction method based on the short time estimation of the magnitude spectrum of the noise. The concept of spectral subtraction is based on the basic concept that the spectral magnitude of speech plus noise can be effectively approximated as the sum of magnitudes of speech and noise. The processing of this method involves estimating the magnitude spectrum of the noise, estimating the magnitude spectrum of the speech signal, and re-synthesizing the speech using the enhanced magnitude spectrum along with the phase spectrum of the noisy speech signal[4][5] as shown in Fig.1.

It includes Hamming windowing of speech with window length of 30ms, FFT calculation with a length of 512 point-FFT, Noise estimation using 5 silence frames, magnitude spectral subtraction, complex spectrum calculation with noisy phase and re-synthesis using IFFT with overlap-add [3].The detailed processing steps in Fig.1 are explained below

Manuscript received Dec , 2015.
***Gitu Geevarghese**, Electronics and telecommunication (EXTC), Fr.C.R.I.T, Mumbai, Maharashtra.*
***Milind Shah**, Electronics and telecommunication (EXTC), Fr.C.R.I.T, Mumbai, Maharashtra.*

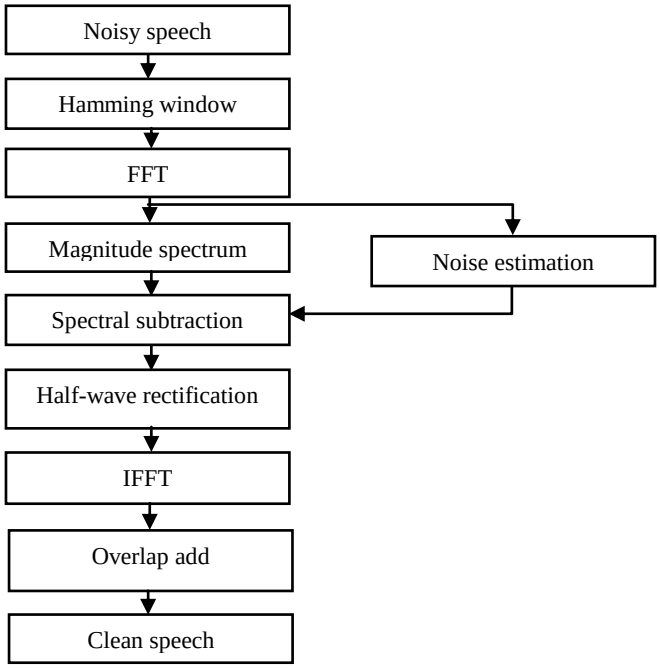


Fig 1 Flowchart of basic spectral subtraction algorithm [10]

A. *Noisy speech generation*

Noisy speech $x(n)$ is obtained by adding a speech signal $s(n)$ to a noise signal $d(n)$.

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

To evaluate the performance of the algorithms, noisy speech [17] with different SNR values ranging from -5, -3, 0, 3 and 5 dB are generated. The SNR calculations for noisy speech is done by taking the root mean square (RMS) value of each and every sample in the frame of clean speech file. Similarly this method was used to calculate the RMS value of the noise file. Based on the calculated speech and noise RMS values, a scaling factor for noise was determined to get an appropriate SNR.

$$SNR_{db} = 10\log_{10}\frac{P_{signal}}{k \times P_{noise}} \tag{2}$$

where k is the scaling factor which is used to obtain the required SNR value. The SNR value is computed first without the use of the scaling factor and according to the required SNR, the noise power is scaled using a scaling factor k. The scaled noise was then added to clean speech signal to get the desired noisy speech signal. This noisy speech signal is used for further processing and to evaluate the appropriate parameters for a particular SNR condition.

B. *Windowing:*

The input signal is windowed by using a Hamming window with an overlap of 50%. The window length was changed from 15ms to 40ms to evaluate the optimum value. In general it was observed that the PESQ score was high in the case when window length was 30 ms and so it was used for further processing. In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each one.

C. *FFT length:*

One of the window related parameter is the order of the FFT [11]. Investigations with FFT length of 256, 512 and 1024 was carried out and the minimum FFT corresponding to a given frame size of 30 ms was adequate with no noticeable improvement in going into a higher order. So a length of 512 samples was set as the order of FFT, which is similar to that reported by boll in [10].

D. *Magnitude spectrum:*

The magnitude part of the obtained FFT is noisy speech estimate which is used further for subtracting from the noise estimate to obtain the clean speech spectrum. The magnitude estimate of noisy speech is given as:

$$|X_n(k)| = \sum_{n=0}^{L} x(n)e^{-j\omega n} \tag{3}$$

where $x(n)$ is the noisy speech, $e^{-j\omega n}$ is the phase function of the DFT of the input noisy speech and L is the length of the FFT.

E. *Noise estimate:*

The spectral estimate of the noisy speech was obtained using FFT of the windowed frame. The noise estimate is obtained by taking a mean of past 5 frames and subtracted from the noisy speech estimate to obtain the clean speech which is added with the phase of the noisy speech. The noise estimate $Dn(k)$ is obtained by averaging the signal magnitude spectrum during non-speech activity (i.e. initial silence period of 0.5 sec) which is given as,

$$D_n(k) = E\{|X_n(k)|\} \tag{4}$$

For investigation on the noise estimation duration, the numbers of frames to be used for estimation of noise were varied from 5, 10, and 15. The mean of the past 5 frames showed best results and so this value was stored in a matrix which was further used for subtraction.

F. *Subtraction rule:*

The subtraction rule in basic spectral subtraction is as shown below:

Let $$Y_n(k) = |X_n(k)| - D_n(k) \tag{5}$$

Where

$$Y_n(k) = \begin{cases} Y_n(k) & if \quad Y_n(k) \geq 0 \\ 0 & otherwise \end{cases} \tag{6}$$

Where $Yn(k)$ is the enhanced signal spectrum, $Xn(k)$ is the spectrum of the input noise corrupted speech, and $Dn(k)$ is the estimate of the noise spectrum. The value of $Y_n(k)$ according to the Eq.6 is made zero for all the negative values of $Y_n(k)$. The magnitude spectrum or the power spectrum after spectral subtraction may contain some negative values due to errors in the estimated noise spectrum. These values are rectified using half-wave rectification (set to zero).

G. *IFFT and Overlap-add:*

The enhanced speech signal is obtained from both $Yn(k)$ and the original phase of the noisy speech signal by an inverse Fourier transform:

$$y(n) = F^{-1}\{|Y_n(k)|e^{j\theta(w)}\} \tag{7}$$

Since it is assumed that speech signal and noise are uncorrelated, some of the components of the processed spectrum may be negative. These negative values are set to

zero as shown in Eq 7. This can lead to further distortions in the resulting time signal. Using overlap-add method the enhanced signal is reconstructed, resulting in a segment synchronization criterion for overlap-add that solves the problem of phase-inconsistency between overlapping segments. The number of samples after IFFT is 512.

III. MODIFIED SPECTRAL SUBTRACTION

A major problem with basic spectral subtraction method is that a new type of noise is introduced in the processed speech signal which is known as the "musical noise". Though the noise is removed there still remains some considerable broadband noise in the processed speech. The broadband noise remains in the form of peaks and valleys (points which are lower than the estimate). Thus after subtraction valleys are set to zero by half-wave rectification and peaks remain as it is. The wider peaks are perceived as time-varying broadband noise and the narrower ones which are relatively large spectral excursions because of the deep valleys that define them are referred to as time varying tones which are referred to as musical noise. To overcome the shortcomings of basic spectral subtraction, Berouti et al. [11] developed a modified spectral subtraction. The modified spectral subtraction [11] steps are as shown in the flowchart in Fig 2.

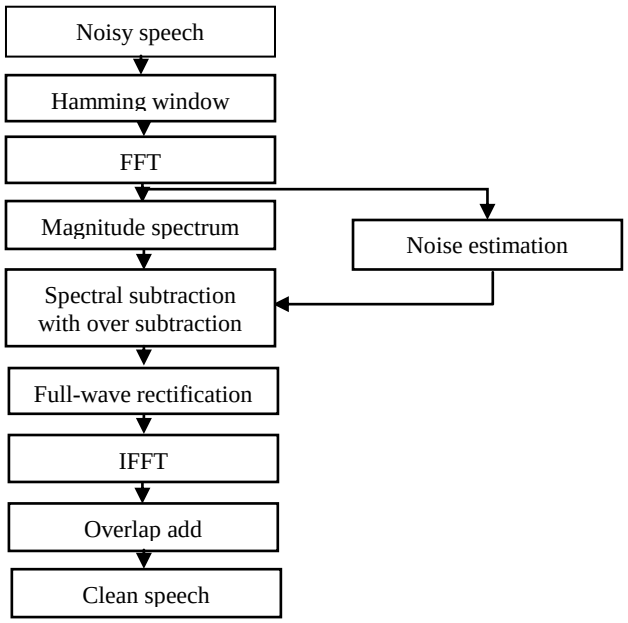


Fig 2 Flowchart of modified spectral subtraction [11]

The windowing, FFT, magnitude averaging, noise estimation and IFFT remains the same as the of the basic spectral subtraction method. Only the subtraction rule and rectification changes in modified spectral subtraction.

A. *Subtraction rule:*

The enhanced magnitude spectrum $|Yn(k)|$ computed using modified spectral subtraction is given as following

Let $$Y_n(k) = |X_n(k)|^{\gamma} - \alpha D_n(k)^{\gamma} \quad (8)$$

Here γ is an exponent factor, resulting in power subtraction if γ = 2 and magnitude subtraction if γ = 1. So γ is set to 1 so as to have magnitude subtraction. The parameter α is used as an over-estimation factor, as there are spectral errors present in the recovered speech. Use of subtraction factor $\alpha > 1$ reduces the broadband peaks in the residual noise, but it may result in deep valleys, causing warbling or musical noise and adversely affecting the speech quality. The parameter α is the over-subtraction parameter which changes according to the frame-wise SNR, if the SNR is too low then a high subtraction might be required which may be as large as 5 and for a high SNR value >20dB, the subtraction factor can be just 1.
The oversubtraction factor can be computed as:

$$\alpha = \begin{cases} 5 & SNR \le -5dB \\ 4-(3/20)SNR & -5 \le SNR \le 20 \\ 1 & SNR \ge 20 \end{cases} \quad (9)$$

The above given equation of α follows the graph as shown in figure,

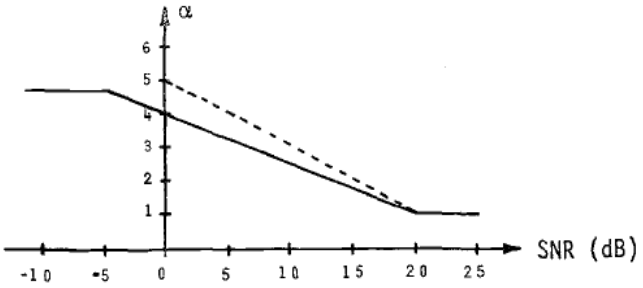


Fig. 3 plot of versus SNR in dB [11]

The dotted line in the figure shows the plot of the value of α used in an experiment conducted by Berouti [11], where several sentences at different SNR's were processed.

B. *Full-wave rectification:*

The full wave rectification is a process in which the negative spectral values obtained after the subtraction is made to an absolute value as given in the eq.10.

$$Y_n(k) = \begin{cases} Y_n(k)^{\frac{1}{\gamma}} & if \quad Y_n(k)^{\frac{1}{\gamma}} \ge \beta D_n(k) \\ \beta D_n(k) & otherwise \end{cases} \quad (10)$$

where β is the spectral floor parameter. The musical noise is masked by a floor noise controlled by the spectral floor factor. The negative values are set to an absolute value with the help of full wave rectification. As spectral subtraction involves modification of the STFT, there may be discontinuities between the signal segments corresponding to the modified complex spectra of the consecutive frames.

Typical values for α is 3 to 6 for SNR>20dB. The spectral floor depends upon the average segmental SNR of the input. For high noise levels (SNR=-5dB) β should be in the range of 0.02 to 0.06 and for lower noise levels (SNR=0 or 5dB) β should be in the range of 0.005 to 0.02 [11]. The optimal values for α and β were 2 to 3 and 0.001 respectively Assuming that the phase error does not significantly affect the intelligibility and quality of speech, the enhanced magnitude spectrum is combined with the original noisy phase, to get the complex spectrum

IV. RESULT AND DISCUSSION

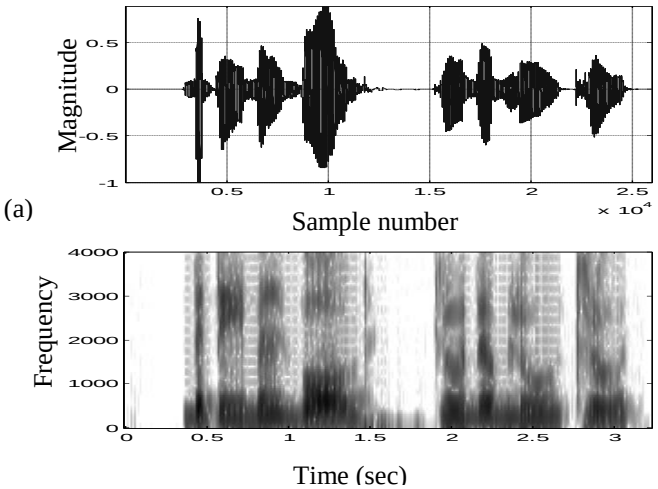
For the evaluation of the basic and modified algorithms, utterances from three male and three female were recorded which involved /a/-/i/-u/, VCV syllables /aba/, /afa/ and /aya/

ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 4, Issue 12, December 2015

and sentences “The table walked through the blue truth”, “The strong way drank the day” and “Never draw the house and the fact”. However, the results related to sentences are only reported in this paper. Noisy speech was generated by adding white, babble, street, pink, car and train noises at SNR of 5, 3, 0, -3, and -5 dB. The above mentioned noise was taken from the AURORA database [40]. The AURORA database is intended for the evaluation of algorithms for front-end feature extraction algorithms in background noise but may also be used more widely by speech researchers to evaluate and compare the performance of noise robust speech recognition algorithms.

The evaluation of the proposed technique was carried out using perceptual evaluation of speech quality (PESQ) measure. Investigations included the comparison of parameters like window length, noise estimation duration and spectral subtraction parameters with respect to the PESQ score of the unprocessed noisy speech signal. The clean speech recordings were done by taking initial 0.15s as silence. Speech was mixed with different types of noises at different SNR values and processed by the spectral subtraction techniques.

The time waveform and spectrogram of clean speech, noisy speech and recovered speech for a sentence “The table walked through the blue truth” at an SNR of 0dB is shown in the fig 4.The behavior of both the algorithms is shown with the plot of SNR vs PESQ score for an utterance of a sentence from a male speaker keeping window length as 30 ms, noise estimation duration of 5 frames with varying SNR values for both BSS and MSS is shown in fig 5. The Table I give the scores for the unprocessed and processed speech with optimal over-subtraction α and spectral floor parameter β for modified spectral subtraction, where the window length of 30 ms, noise estimation duration of 5 frames and a FFT length of 512 point is kept constant.



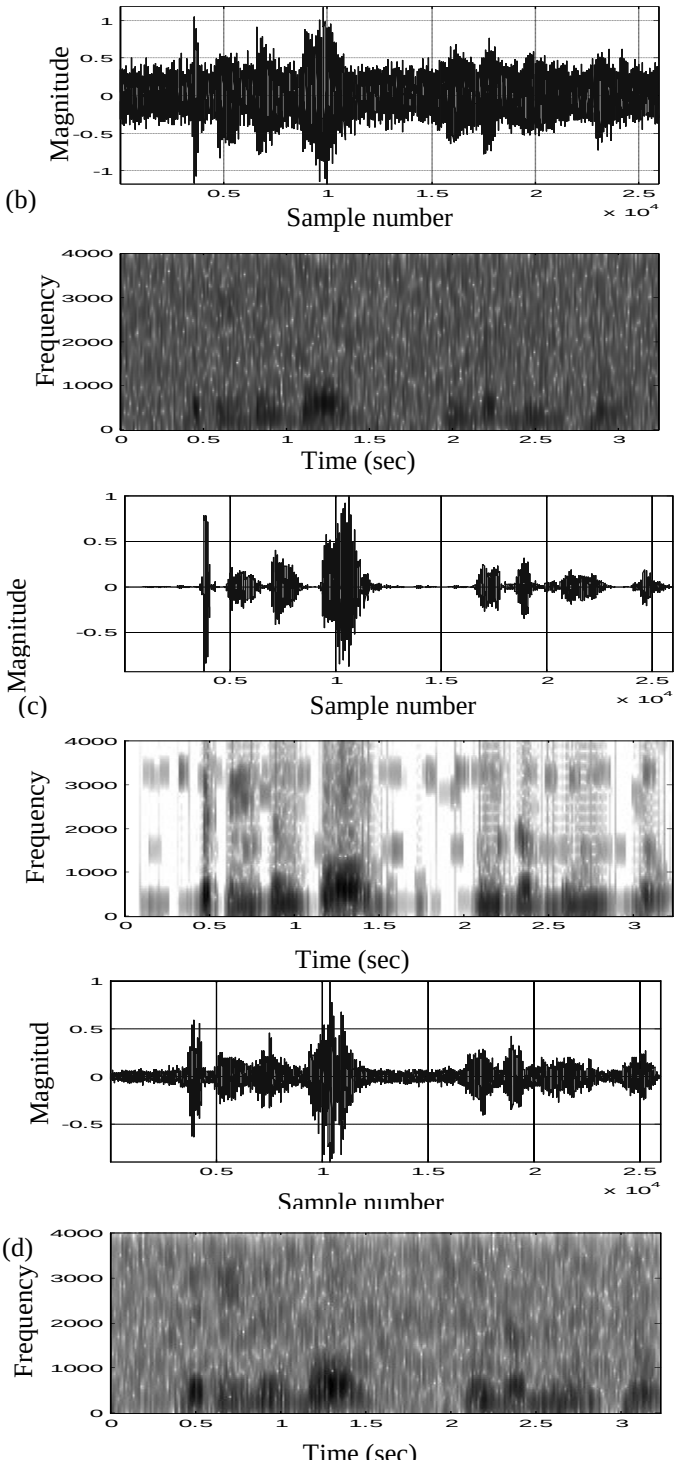
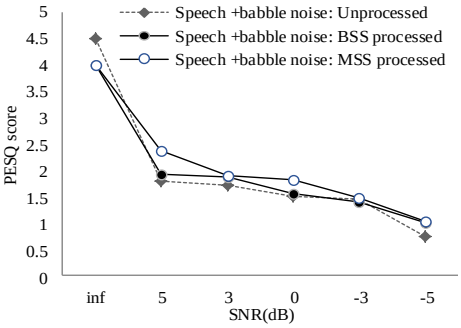
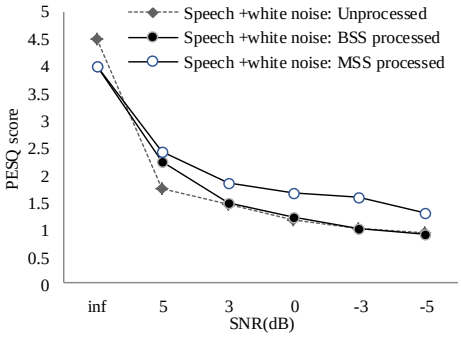


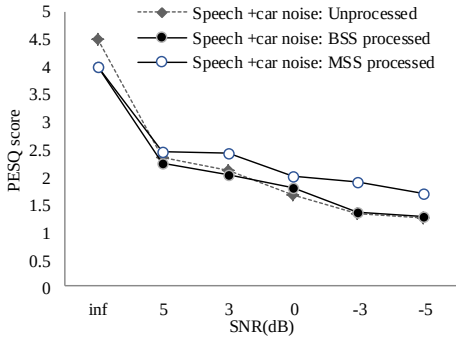
Fig 4 time waveform and spectrogram of (a) clean speech (b) noisy speech (c) recovered speech using MSS (d) recovered speech using BSS.



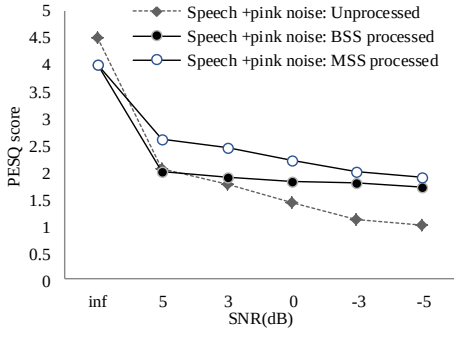
(a) Noise: babble



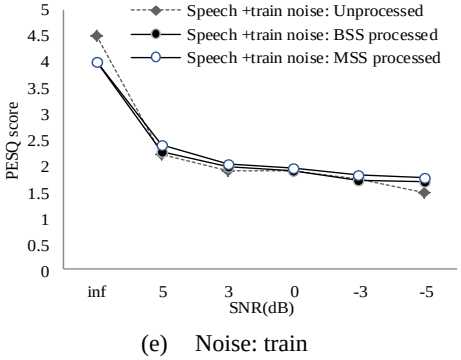
(b) Noise: white



(c) Noise: car



(d) Noise: pink



(e) Noise: train

Fig 5 PESQ vs SNR for 1 male speaker for an utterance of sentence for BSS and MSS method (a) Noise: babble (b) Noise: white (c) Noise: pink (d) Noise: car (e) Noise: train

Table 1 PESQ score for the enhanced speech using modified spectral subtraction for different type of noise .speech material: sentence of a male speaker

(a) Noise: babble

α	PESQ score						
	Beta 0.001	Beta 0.005	Beta 0.008	Beta 0.01	Beta 0.02	Beta 0.04	Beta 0.06
3	1.53	1.52	1.52	1.58	1.45	1.44	1.32
4	1..66	1.63	1.58	1.56	1.43	1.29	1.211
5	1.26	1.25	1.27	1.34	1.14	1.20	1.20

(b) Noise: white

α	PESQ score						
	Beta 0.001	Beta 0.005	Beta 0.008	Beta 0.01	Beta 0.02	Beta 0.04	Beta 0.06
3	1.70	1.71	1.70	1.82	1.70	1.70	1.69
4	1.89	1.89	1.89	1.89	1.89	1.89	1.89
5	1.56	1.56	1.56	1.56	1.56	1.56	1.56

(c) Noise: car

α	PESQ score						
	Beta 0.001	Beta 0.005	Beta 0.008	Beta 0.01	Beta 0.02	Beta 0.04	Beta 0.06
3	1.52	1.53	1.53	1.54	1.56	1.55	1.54
4	1.46	1.67	1.65	1.67	1.65	1.65	1.52
5	1.45	1.45	1.45	1.46	1.45	1.46	1.45

(d) Noise: pink

α	PESQ score						
	Beta 0.001	Beta 0.005	Beta 0.008	Beta 0.01	Beta 0.02	Beta 0.04	Beta 0.06
3	1.66	1.64	1.66	1.66	1.62	1.62	1.61
4	1.58	1.57	1.58	1.59	1.57	1.57	1.57
5	1.59	1.59	1.59	1..62	1.61	1.607	1.606

(e) Noise: train

α	PESQ score						
	Beta 0.001	Beta 0.005	Beta 0.008	Beta 0.01	Beta 0.02	Beta 0.04	Beta 0.06
3	1.89	1.89	1.85	1.92	1.85	1.90	1.81
4	1.92	1.92	1.92	1.94	1.85	1.85	1.85
5	1.83	1.87	1.867	1.865	1.85	1.66	1.66

VI. CONCLUSION

A basic and modified spectral subtraction for suppression of additive noise is being reported. From the various experiments conducted, basic spectral subtraction (BSS) gave a PESQ score improvement of 0.3 - 0.8 and modified spectral subtraction (MSS) gave a PESQ score improvement of 0.4 - 1.1 for sentences spoken by three male and three female speakers. In comparison with other noises, speech corrupted with white noise showed maximum improvement. The observation was valid in both male and female cases. For babble noise it was observed to have least improvement when compared to other noises. The reported speech enhancement methods can be combined with various other noise estimation techniques and signal processing methods for improved perception for hearing impaired listeners.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York: CRC, 2007.
[2] S. V. Vaseghi, *Advanced Digital Speech Processing and Noise Reduction*, 2nd ed., Chichester: John Wiley & Sons ltd., 2000.
[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, New Jersey: Prentice Hall, 1978.
[4] R. Martin, "Spectral Subtraction Based On Minimum Statistics," *in Proc*. EUSIPCO '94, 1994, pp. 1182-85.
[5] R. Chen, "Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking," *IEEE Trans. Audio Speech Lang. Process*., vol. 32, no.10, pp.1-4, 2005.
[6] R. J. McAulay and M. L. Malpass, Speech enhancement using a soft decision noise suppression filter, *IEEE Trans. on Acoust. Speech Signal Process.*, vol. 28, pp. 137-145, 1980.
[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, pp. 1109-21, 1984.
[8] M. Dendrinos, S. Bakamides, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Comm*., vol. 10, pp. 45-57, 1991.
[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust. Speech Signal Process*., vol. 27, no.2, pp.113-120, 1979.
[10] S. F. Boll, "Suppression of noise in speech using the SABER method," *in Proc*. ICASSP, 1978, pp. 606—609.
[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP* , Washington, DC, 1979, pp. 208–211.
[12] S. K. Waddi, P. C. Pandey, and N. Tiwari, "Speech enhancement using spectral subtraction and cascaded-median based noise estimation for hearing impaired listeners," *in Proc. Nat. Conf. Commun. (NCC 2013)*, Delhi, India, 2013, pp. 1–5.
[13] P. Boersma and D. Weenink. (2015) Praat: doing phonetics by computer. [Online], Available: www.praat.org
[14] K. Paliwal, K. Wojcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun*., vol. 52, no. 5, pp. 450–475, 2010.
[15] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *in Proc. ICASSP*, Detroit, MI, 1995, pp. 153–156.
[16] ITU, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec*., P.862, 2001.
[17] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in Proc. Int. conf. spoken language process.*, Beijing, China, 2000, pp. 29–32.