The Nucleus **50**, No. 2 (2013) 155-163

The Nucleus
A Quarterly Scientific Journal of Pakistan Atomic Energy Commission
NCLEAM, ISSN 0029-5698

SIMULATION OF MODIFIED HYBRID NOISE REDUCTION ALGORITHM TO ENHANCE THE SPEECH QUALITY

A. WAQAS, *T. MUHAMMAD and H. JAMAL

Department of Electrical Engineering, University of Engineering and Technology, Taxila, Pakistan

(Received March 26, 2013 and accepted in revised form May 27, 2013)

Speech is the most essential method of correspondence of humankind. Cell telephony, portable hearing assistants and, hands free are specific provisions in this respect. The performance of these communication devices could be affected because of distortions which might augment them. There are two essential sorts of distortions that might be recognized, specifically: convolutive and additive noises. These mutilations contaminate the clean speech and make it unsatisfactory to human audiences' i.e. perceptual value and intelligibility of speech signal diminishes. The objective of speech upgrade systems is to enhance the quality and understandability of speech to make it more satisfactory to audiences. This paper recommends a modified hybrid approach for single channel devices to process the noisy signals considering only the effect of background noises. It is a mixture of pre-processing relative spectral amplitude (RASTA) filter, which is approximated by a straightforward 4^{th} order band-pass filter, and conventional minimum mean square error short time spectral amplitude (MMSE-STSA85) estimator. To analyze the performance of the algorithm an objective parameter called Perceptual estimation of speech quality (PESQ) is measured. The results show that the modified algorithm performs well to remove the background noises. SIMULINK implementation is also performed and its profile report has been generated to observe the execution time.

Keywords: RASTA filter, Modified hybrid algorithm, PESQ score, MMSE-STSA85 algorithm, Performance analysis

1. Introduction

A standout amongst the most regular types of human communication is speech. It is produced by the speech handling models dependent upon the learning of human vocal tract framework such as speech creation system of a speaker [1]. To pass it to the sound-related arrangement of an audience it is, then, transmitted with a certain medium, which might be copper wires, fiber links or essentially the air. The speech signal may be debased by diverse sorts of commotions throughout this transmission. Additive noise is the one which comes about when background sound distortion includes itself in the desirable speech.

In numerous common requisitions like versatile telephones, speech distinguishment framework, VoIP and, portable hearing assistants it is profitable to smoother such underpinning added substance commotions [2, 3]. Consequently, improvement of preprocessing functional processes for talk upgrade is consistently of investment. The reason for the speech upgrade procedures is to enhance the quality and understandability of speech for the human audience. The purported technique is reputed to be noise lessening.

Over the years engineers are in a constant battle to advance a mixture of successful systems to warfare this issue. In any case, the restrictions of these systems, in signal processing, still represent an impressive challenge to the researchers. Evacuating diverse sorts of noise, due to inherent complications of speech and irregular nature of noise, is a difficult process. Noise diminishment methods more often than not have a tradeoff between the quantity of noise removal and speech mutilations added due to the handling of speech.

Speech is a non stationary signal whose frequency contents consistently change with time and it has certain particular properties [1]. Speech processing ordered systems ordinarily work on a frame by frame support with frames extending from 20 to 30 ms throughout which the signal is recognized as semi-stationary. Right by taking DTFT not gives any handy informative data as time qualified information is absent. To deal with time changing otherworldly informative data short time discrete Fourier transform (STDFT) handling methodology is for the most part utilized at present [4].

* Corresponding author : tahir.muhammad@uettaxila.edu.pk

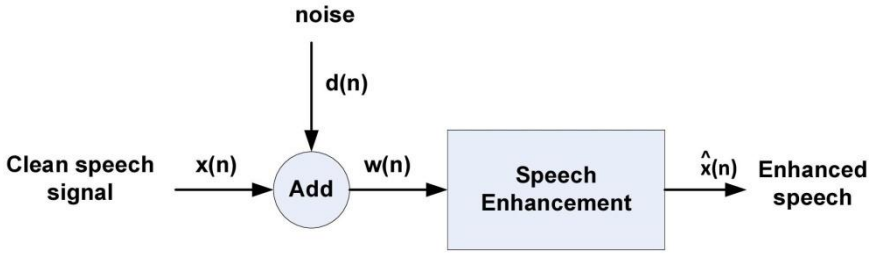

Figure 1. Speech enhancement additive noise model.

Speech observation is an unpredictable method which is not well grasped contrasted with speech generation. The sound-related framework is less sensitive to phase than to magnitude and the short-time magnitude is recognized vital for the discernment of speech. Along these lines, the most regular methodology is to evaluate the short time spectral amplitudes (STSA) to have a satisfactory speech quality. The phase of the noisy speech is separated then afterward consolidates it with improved spectral values to yield an appraisal of the STFT.

The well known illustrations of short time spectral attenuation (STSA) based systems are spectral subtraction and Wiener filtering [5, 6]. In spectral subtraction ordered system, an assessment of the clamor spectrum is subtracted from the spectrum of noisy speech [6]. By utilizing Weiner filtering method, an optimal MMSE appraisal of the intricate Fourier transform coefficients is first obtained to acquire the estimator. Both of these are not optimal spectral amplitude estimators from perceptual perspective. Ephraim and Malah [7, 8] figured an optimal amplitude estimator pointed to as MMSE- STSA85 to appraise the magnitude of every intricate Fourier coefficient from the processed frame of noisy speech. In any case, it experiences musical noise impacts [9].

The Relative Spectral Amplitude (RASTA) ordered system recommended by Hermanskey and Morgan is an additional methodology for speech improvement for programmed speaker distinguishment frameworks to decrease the impact of additive and convolutional noise [10]. Assessment of initial RASTA filter and adjustment in its parameters is done by S.K.Shah et al. [11]. By utilizing RASTA musical noise situation is settled up to some degree and additionally remaining noise stays in the initial segment of speech signal. To defeat this situation a blend of RASTA and MMSE_STSA85 is utilized to upgrade the exhibition. The technique demonstrate its imperativeness at flat SNRs, on the other hand, it is not overall suited for heightened SNRs which limit its requisition for high SNR speech signal [12,13].

This paper gives a modified hybrid approach to improve the quality of speech. It throws some light upon problem formulation aspect of speech enhancement. The paper assesses the exhibition of the algorithm regarding perceptual estimation of speech quality (PESQ) measures. At last, the SIMULINK execution is performed and profile report is created to watch the execution time for further advancement of the code.

2. Problem Formulation

In certain requisitions, for example portable hearing assistants and multi-station teleconferencing, a few microphones could be utilized and hence numerous renditions of the noisy signal are ready synchronously. In any case, to point of confinement the framework's requirements, one and only one microphone is put forth in the larger piece of the pie of requisitions. The methodology of suppressing noise when one singular origin of the noisy signal is accessible is pointed to as single channel speech enhancement. They generally do not improve the intelligibility of speech.

A noise free signal x(n) is collected to be corrupted by background noise d(n). The resultant noisy signal got by the microphone is figured as

$$w(n) = x(n) + d(n) \quad (1)$$

Where n is the index in discrete time.

The objective for any noise suppression framework is then to estimate the clean signal $\hat{x}(n)$ having access just to the noisy signal w(n). It is charming to constrict the noise signal however much as could be allowed while keeping the distortions of the speech signal as level as could reasonably be expected in the meantime. This is shown in Figure 1.

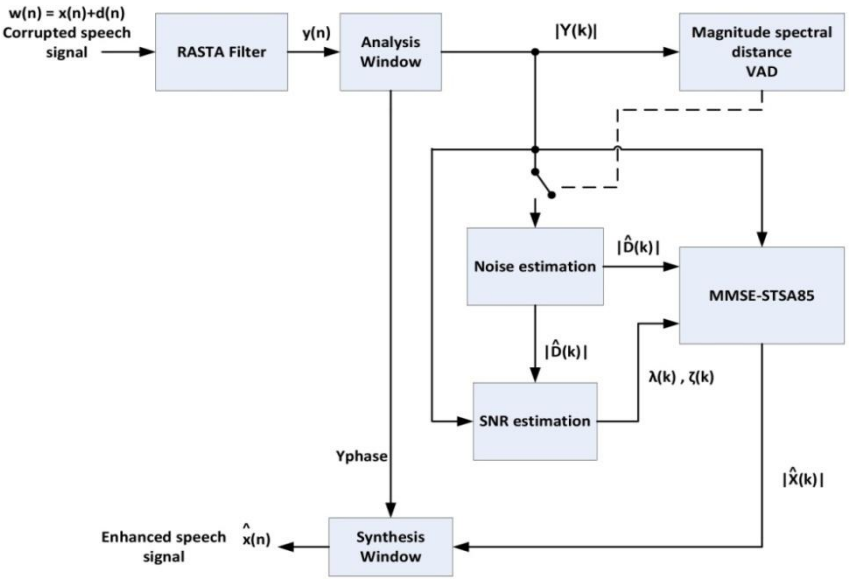

Figure 2. Block diagram of proposed modified hybrid algorithm.

3. Modified Hybrid Algorithm

The algorithm is shown as block diagram in Figure 2. It utilizes a blending of pre-processing RASTA filter and Log spectral amplitude estimator. The noise contaminated signal w(n), comprising of noise free signal x(n) and background noise d(n), is presented to the algorithm. The point, as described earlier, is to gauge the improved signal $\hat{x}(n)$ at the yield. The functionality of the entire process relies on diverse blocks utilized as a part of it. The following section describes them one by one.

3.1. RASTA Filter

The rate of change of vocal tract is closely related to linguistic information. The speech signal reveals this rate of alterations. Non linguistic informative data regularly lies outside this rate of alteration. The RASTA filter exploits this certainty and smoothers the level or towering frequency segments that are outside the regular reach of rate of change of speech. The rate of change could be depicted by the modulation spectrum. The written works declares that a large portion of the advantageous phonetic informative content is in frequency components from the extent between 1Hz and 16Hz, with the prevailing part at around 4 Hz. The range around 4Hz is handy both in clean and noisy nature. In noisy nature, the range beneath 2Hz or above 10Hz is less imperative [10].

Thinking about these frequency characteristics the RASTA filter utilized for this trial is portrayed by the transfer function given in Eq. (2).

$$H(z)=z^4\left(\frac{0.078-0.157z^{-2}+0.078z^{-4}}{1-2.991z^{-1}+3.393z^{-2}-1.78z^{-3}+0.379z^{-4}}\right) \quad (2)$$

It is approximated by an effortless 4th order band pass IIR filter. The modulation and sampling frequency of the filter is 50Hz, 100Hz respectively. The magnitude and phase response of RASTA filter is demonstrated by Figures 3 and 4.

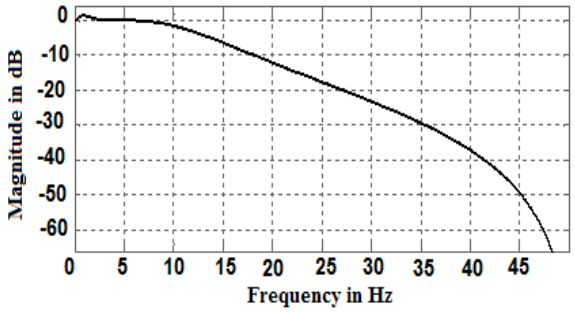

Figure 3. RASTA filter's magnitude response.

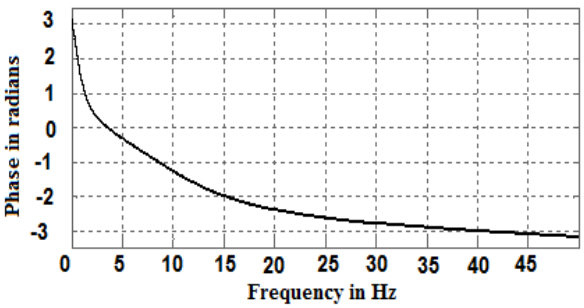

Figure 4. RASTA filter's phase response.

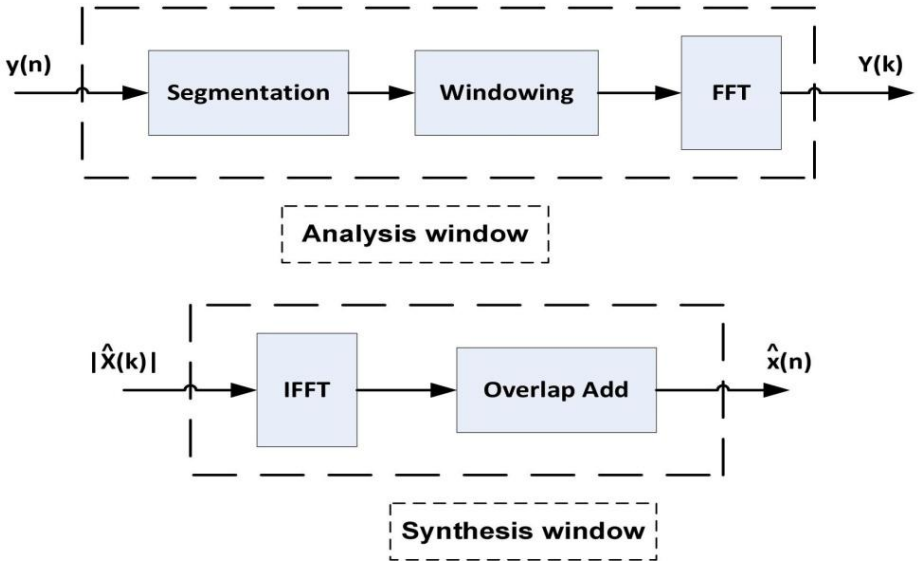

Figure 5. Analysis synthesis framework.

3.2. Analysis Synthesis Framework

The speech signal got from the RASTA filter is further processed in frequency domain. This is being as how the frequency domain processing demands less effort as compared to time domain [14]. For this reason, analysis-synthesis framework with near to ideal reconstruction is applied.

At first, the pre-processed noisy speech y(n) is partitioned into frames, each of length 256 samples, with overlapping of 128 samples. With a specific end goal to balance the spectral leakage effect, the samples in every frame are reproduced by a tapered hamming window which is the most generally utilized window method. After division, windowing and if essential zero padding, the noisy short-time fragments are converted into the frequency domain utilizing a Fast Fourier Transform (FFT) of length 256. The method is demonstrated in Figure 5. Changing the noisy signal into the frequency domain is a broadly received system for speech quality improvement as it strongly compares to the handling happening in the human sound-related system. Be that as it may, as the human ear is noticeably insensitive to phase distortions, most estimators just alter the spectral magnitudes and utilize the phase of the noise containing signal for reproduction.

The estimator's gain G lies in the reach between 0 and 1. To get the improved signal in time domain, all operations linked with the analysis window are switched in succeeding synthesis window. As demonstrated in Figure 5, an Inverse Fast Fourier Transform (IFFT) and overlap-add are used for this reason.

3.3. Voice Activity Detector and Noise Estimation

The usefulness of MMSE-STSA85 is remarkably reliable on speech silent indicator as it is acknowledged indispensible to appraise the speech or silent areas. The difference between the improved signal and the clean signal is minimized if the assessment of the noise spectrum is correct. Subsequently, it is pleasing to evaluate the noisy signal at each ready moment to get a more precise assessment of the noise spectrum. This is not a situation in double channel usage since the noisy signal is only made accessible in a second channel. Nonetheless, in single channel routines, the noisy signal needs to be gauged from the mixture of clean and noisy signal itself because of the non accessibility of a second clamor channel. Subsequently a voice activity detector (VAD) is needed that will recognize the aforementioned frames of the information signal in which the speaker is silent. These information signals are recognized to just hold the meddling noise signals and the noise spectrum is upgraded. The VAD needs to exactly distinguish such times of hush to forestall computing a wrong overhaul of the noise spectrum with parts of the speech signal, since resulting subtraction will evacuate the speech signal from the succeeding input fragments. Thusly, the speech silent indicator is just needed to distinguish pauses between statements or sentences and not moves between phonemes or

statements. Thus, the system is in addition called the speech silent indicator.

The technique investigates the mean worth of the spectral magnitude distance with a predefined threshold η to confirm the speech or silent areas. The distance value is computed as

$$\text{Dist}=\text{Mean}\left[20\left(\log Y(k)-\log\hat{D}(k)\right)\right] \quad (3)$$

$$\begin{Bmatrix}\text{Dist}<\eta & \text{speech absent}\\ \text{Dist}>\eta & \text{speech present}\end{Bmatrix} \quad (4)$$

$$\hat{D}(k)=\text{Mean}[(Y(:,1:\text{NIS})').\hat{}\,2]' \quad (5)$$

NIS is the number of initial silence segments in the speech signal.

The value of η is ordinarily situated to 3. At the point when only silent region is present in the processed frame, the noise spectrum is upgraded as

$$|\hat{D}(k)_{curr}|^2=\beta.|\hat{D}(k)_{prev}|^2+\frac{|Y(k)_{curr}|^2)}{(1+\beta)} \quad (6)$$

Where β=9.

3.4. Decision-Directed SNR Estimation

The apriori SNR ζ(k) is obscure and we need to evaluate it so as to actualize the log spectral amplitude estimator. The excuse for why ζ(k) is unfamiliar is being as how the clean signal is occupied. The decision directed approach has been utilized to gauge ζ(k) [10].

Let λ(k), ζ(k) and, Y(k) indicate the aposteriori SNR, apriori SNR, the magnitude, individually of the comparing frame. The definitions of aposteriori SNR λ(k) and ζ(k) and their relation is the basic requirement of inference of the apriori SNR estimator.

$$\zeta(k)=\frac{E\{Y(k)^2\}}{|\hat{D}(k)|^2} \quad (7)$$

$$\zeta(k)=E\{\lambda(k)^2-1\} \quad (8)$$

Also

From Eq. (7) and Eq. (8) we get

$$\zeta(k)=E\left[\frac{1}{2}\frac{Y(k)^2}{|\hat{D}(k)|^2}+\frac{1}{2}\{\lambda(k)^2-1\}\right] \quad (9)$$

The resultant estimator in an iterative shape might be derived from Eq. (9) as

$$\zeta(k)=\gamma\frac{|\hat{X}(k)|^2_{prev}}{|\hat{D}(k)|^2_{prev}}+(1-\gamma)\max\{\lambda(k)_{curr}-1,0\} \quad (10)$$

The above estimator for ζ(k) is a decision directed sort estimator. By utilizing

$$\hat{X}(k)=G(k).Y(k) \quad (11)$$

Where G is a gain function which comes about because of the amplitude estimator, Eq. (10) might be composed in an iterative form as

$$\zeta(k)=\gamma.|G(k)|^2.\lambda(k)_{prev}+(1-\gamma).\max\{\lambda(k)_{curr}-1,0\} \quad (12)$$

The a posteriori SNR $\lambda(k)$ is defined as

$$\lambda(k)=\frac{|Y(k)|^2}{|\hat{D}(k)|^2} \quad (13)$$

Here γ is the smoothing parameter which lies in the extent between 0 to 1 and is regularly situated to .99 (.95 to .99) for optimal exhibition [12].The starting condition that minimizes initial transition effects in the improved signal is given in Eq. (14).

$$\zeta(k)=\gamma+(1-\gamma).\max\{\lambda(k)-1\} \quad (14)$$

3.5. MMSE Log Spectral Amplitude Estimator

Taking into account the surmise that logarithmic compression is performed by the human sound-related framework of short time spectral attenuation. As a result the logarithm of the STSA is more perceptually related than the STSA [8]. The MMSE of the logarithm of the STSA is suggested in [7] for which the gain function is computed as

$$G(k)=\frac{\zeta(k)}{1+\zeta(k)}\exp\left(\frac{1}{2}\int_{Uk}^{\infty}\frac{e^{-t}}{t}dt\right) \quad (15)$$

Where U*k* is defined by

$$U_k=\frac{\zeta(k)}{1+\zeta(k)}.\lambda(k)_{curr} \quad (16)$$

At last, the improved speech signal can be attained by multiplying the gain G(k) with spectral values Y(k) as given in Eq. (11).

4. Implementation Results and Performance Analysis

4.1. Database Used to Test the Algorithm

A noisy speech signal's database is advanced to help examination of speech upgrade contrivances right around exploration bunches. The database holds 30 IEEE sentences defiled by eight distinctive noises at distinctive SNRs. The noise has taken from the AURORA database incorporating several sorts of noises.

Table 1. PESQ score comparison.

Signal	Noisy Signal	RASTA Filter	MMSE-STSA85	Modified Hybrid
Car noise-0dB	1.3712	1.7376	1.7832	1.9894
Car noise-5dB	1.6626	1.9693	2.1840	2.4188
Car noise-10dB	1.8797	2.1991	2.5315	2.7366
Car noise-15dB	2.2509	2.5123	2.6922	2.8664
Restaurant noise-0dB	1.5669	1.8576	1.6795	1.8942
Restaurant noise-5dB	1.8164	2.0242	1.8354	2.0270
Restaurant noise-10dB	2.245	2.4409	2.6050	2.6419
Restaurant noise-15dB	2.5586	2.7797	2.8561	3.0267
Airport noise-0dB	1.6899	1.9027	1.6989	1.7928
Airport noise-5dB	2.0136	2.1982	2.2355	2.3902
Airport noise-10dB	2.0997	2.3241	2.3804	2.4674
Airport noise-15dB	2.4462	2.6214	2.6516	2.7543
Train noise-0dB	1.2126	1.7575	1.5388	1.8969
Train noise-5dB	1.5018	2.0316	1.9236	2.2405
Train noise-10dB	1.7873	2.3521	2.4441	2.7029
Train noise-15dB	2.2221	2.6832	2.7451	2.9889
Street noise-0dB	1.5585	1.7958	1.8052	1.9569
Street noise-5dB	1.6044	2.0581	1.8239	2.3101
Street noise-10dB	1.8726	2.2408	2.4304	2.8018
Street noise-15dB	2.2638	2.5055	2.7372	2.8514
Babble noise-0dB	1.4400	1.7774	1.5087	1.6583
Babble noise-5dB	1.8065	2.0851	2.1181	2.1334
Babble noise-10dB	2.0093	2.3349	2.2216	2.4573
Babble noise-15dB	2.4767	2.6769	2.7298	2.8041

A sound proof environment, such as Davis Tucker technology, is used to record thirty IEEE sentences. These sentences are recorded by three female and three male speakers. The thirty sentences from 720 IEEE sentences' database are chosen in such a way so that they contain all phonemes in the American language. All the sentences are down sampled to 8 kHz from the original 25 kHz sampling frequency. Noise is artificially augmented the speech signals at distinctive SNRs [15].

4.2. Objective Evaluation

The goal value measure test; perceptual estimation of speech quality (PESQ) is carried out to test the functional process [16]. The noisy signal's test files recognized for this experiment are of several sorts having SNR values from 0dB to 15dB and are defiled with three different noises. The first ever MMSE-STSA85, RASTA and modified hybrid functional processes are examined dependent upon this test. The PESQ measure for noisy speech is likewise incorporated for illustration. The PESQ score for different techniques is demonstrated in Table 1.

The PESQ score of modified hybrid approach at 0dB SNR is to some degree tantamount (little bit degradation in some cases) with RASTA and initial MMSE_STSA85 though it demonstrates to certain enhancements if there should arise an occurrence of car noise. For 5dB and 10dB SNRs the outcomes are perceptibly enhanced in all cases. In Contrast to existing hybrid methodology exhibited in [11, 13], in any case, we can again state that PESQ measure is tantamount at 0dB SNR. For 5dB SNR the outcomes have enhanced from 6% to 25% for distinctive scenarios. For 10dB SNR the enhancement is 20%. For the most part; it is watched that at flat SNR levels such as 0dB the modified methodology is comparable with existing strategies while it performs best for SNR values greater or equal to 5dB. The modified hybrid methodology is suggested to utilize within flat and also at high SNR conditions.

For easy analysis Figure 6, 7 and, 8 show some of the results in the form of a bar graph chart.

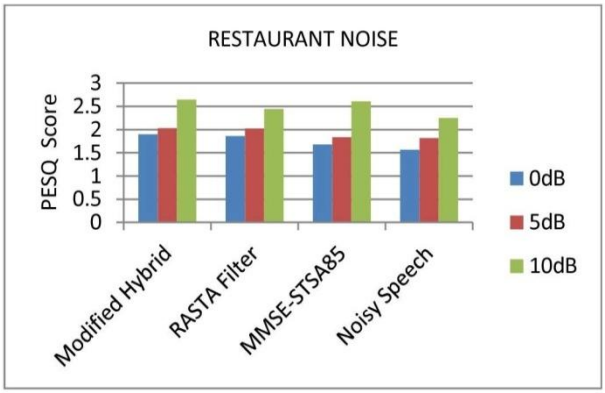

Figure 6. PESQ score connection in restaurant clamor.

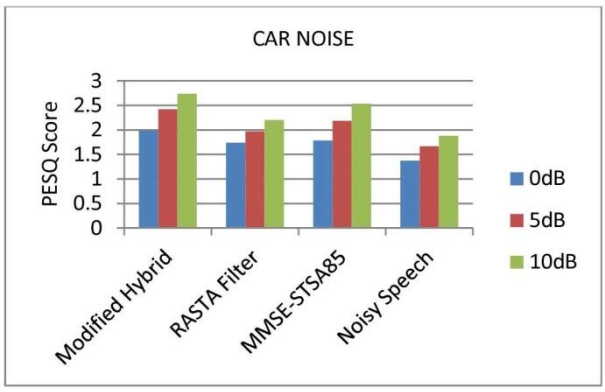

Figure 7. PESQ score connection in car clamor.

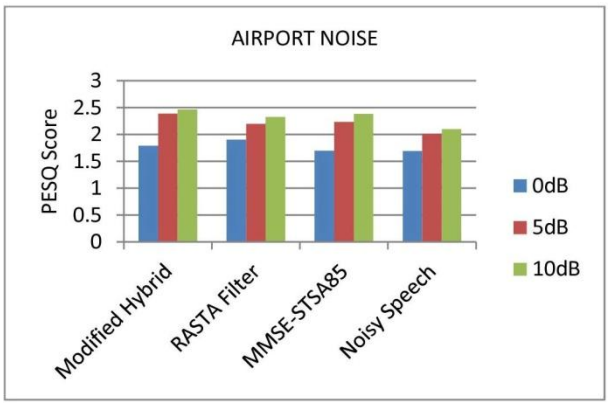

Figure 8. PESQ score connection in airport clamor.

5. SIMULINK Implementation

The SIMULINK model is planned as indicated in figure 9. It contains RASTA filtering, segmentation with FFT, MMSE-STSA85 estimator and, overlap-add blocks as a major part of it. Each block is represented by a different user defined embedded MATLAB function. After pre-processing RASTA filtering the division square, by utilizing the Hamming window, outlines the incoming

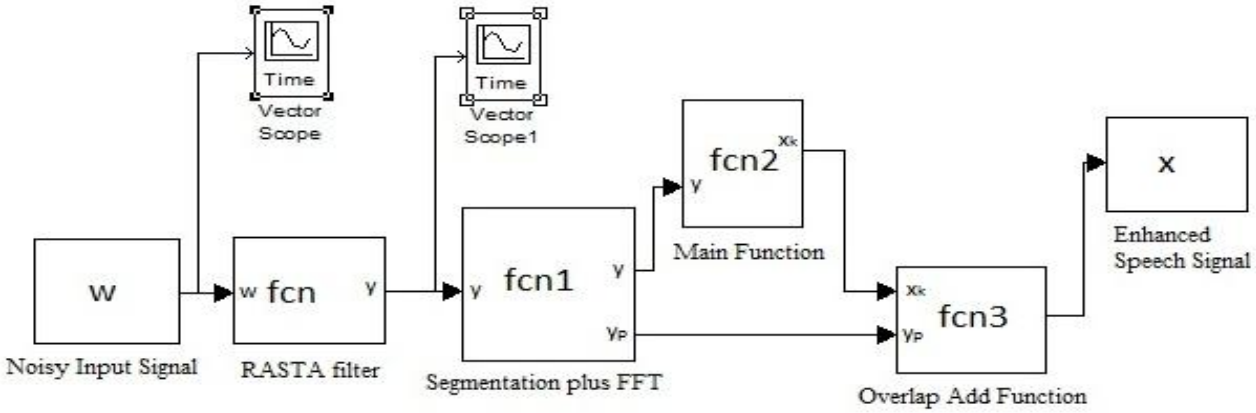

Figure 9. SIMULINK Implementation model.

information into a 256 samples' frame with 128 samples of frame shift. After fitting confining the FFT block simply takes 256 point FFT and magnitude is segregated from complex values for further handling and phase is given for remaking with upgraded isolated magnitude. The log spectrum amplitude estimator performs its operation to uproot underpinning noises. At last, from early separated noisy phase and resultant magnitude the improved spectrum is acquired for a particular frame. To get the enhanced speech signal in time domain all the operations are reversed utilizing overlap-add function.

The profiler report has generated and included in the paper to observe how long each block took to execute as it gives insight information for optimization of the code. Table 2 gives the profiler summary.

Table 2. Profiler report.

Function/Block	SIMULINK Implementation
Clock speed	3000MHz
RASTA filter	0.8%
Segmentation and FFT	2.3%
MMSE-STSA85	4.0%
Overlap add	2.2%
Scopes	0.2%
Model initialize	90.2%
Model execute	9.6%

From Table 2 the total time taken by the designed RASTA filter and MMSE estimator is 4.8% whereas it was 12.4% in the existing approach [13]. Here the RASTA filter is being used for pre-emphasis which limits the need to estimate the large number of parameters and makes the iterative decision directed approach quite simple.

6. Conclusions

A modified hybrid approach is presented to cope up with the challenges of background noises which might augment the human communication devices like mobile phones, hearing aids etc. The algorithm is used to improve the speech quality by removing such background distortions. It can be used for flat as well as for high SNR conditions. The PESQ score shows 6-25% increase in performance for different scenarios. Moreover, it is less complex. It can also be implemented using hardware platforms such as FPGA or DSP.

References

[1] B. Gold and N. Morgan, Speech and Audio Signal Processing, John Wiley & Sons, Inc. (2000) Vii, 3, 4, 15.

[2] C. H. You, S. N. Koh and S. Rahardja, Adaptive β-order MMSE Speech Enhancement Application for Mobile Communication in a Car Environment, in Proc. 4th Int. Conf. on Information, Communications and Signal Processing, 4th Pacific Rim Conf. on Multimedia (2003) pp. 1629–1632.

[3] T. Yamada, M. Kumakura and N. Kitawaki, IEEE Trans. Audio, Speech, Language Processing **14**, Nov. (2006) 2006.

[4] J. Benesty, S. Makino and J. Cheng, Speech Enhancement, Springer Series of Signals and Communication Technology (2005).

[5] F. Toledo, P. C. Loizou and A. Lobo, Subspace and Envelope Subtraction Algorithms for Noise Reduction in Cochlear Implants, IEEE Annual International Conference of Engineering Medicine and Biology Society (2003) pp. 213-216.

[6] S. F. Boll, IEEE Trans. on Acoustic, Speech and Signal Processing, **ASSP-27**, April (1979) 113.

[7] Y. Ephraim and D. Malah, IEEE Trans. on Acoustics, Speech and Signal Processing, **ASSP-32** (1984) 1109.

[8] Y. Ephraim and D. Malah, IEEE Trans. on Acoustics, Speech and Signal Processing, **ASSP-33** (1985) 443.

[9] C. Breithaupt and R. Martin, MMSE Estimation of Magnitude-squared dft Coefficients with Super Gaussian Priors in Proc. IEEE Int. Conf. Acoust, Speech, Signal Processing (2003) pp. 848-851.

[10] H. Hermanskey and N. Morgan, IEEE Trans. on Acoustics, Speech and Signal Processing **2** (1994) 578.

[11] S.K. Shah, J.H. Shah and N.N.Parmar, Evaluation of RASTA Approach with Modified Parameters for Speech Enhancement in Communication Systems, Proc. IEEE Symposium on Computers and Informatics (March 2011) pp. 159-162.

[12] J. Shah and S. Shah, International Journal of Computer Science **9**, Issue 4, No 2 (2012) 230.

[13] S.K. Shah and J.H. Shah, Real Time and Embedded Implementation of Hybrid Algorithm for Speech Enhancement, Information and Communication Technologies (WICT), Mumbai, India (2011) pp. 341-345

[14] D. Burshtein and S. Gannot, IEEE Trans. Speech, Audio Process **10** (2002) 341.

[15] The NOIZEUS database (2009) available at: http://www.utdallas.edu/~loizou/speech/noize

[16] A. Hu and P. Loizou, Subjective Comparisons of Speech Enhancement Algorithms, Proc. IEEE International conference on Acoustics, Speech and Signal Processing (May 2006).