5th International Conference on Emerging Trends in Engineering, Technology, Science and Management
Institution of Electronics and Telecommunication Engineers, Ganganagar, Bengaluru, Karnataka, India (ICETETSM-17)
20th August 2017, www.conferenceworld.in ISBN: 978-93-86171-60-3

SPEECH ENHANCEMENT TECHNIQUES

Ramesh Nuthakki

(Asst. professor, Department of E&C, AIT, VTU, Bangalore, India,

ABSTRACT

Spectral subtraction is a very well-known and easier method to remove stationary background noise. Using this algorithm, a spectral noise bias is computed from segments of speech inactivity and is subtracted from noisy speech spectral amplitude by retaining the phase just as the same. In order to reduce the unpleasant acoustic effects due to spectral error, spectral subtraction follows essential methods. The major setback of this algorithm is that it is relevant only to speech corrupted by stationary noise. This paper intends mainly at studying the spectral subtraction and Weiner Filter techniques when the speech is corrupted by non-stationary noise. These two algorithms are studied in terms of non-stationary noise. To evaluate the time varying noise spectrum which produces better performance in terms of intelligibility and reduced musical noise, a decision directed approach (DD) is used. Nevertheless, the a priori SNR estimator of the present frame depends on the estimated speech spectrum from the earlier frame. The unpleasant consequence is that the gain does not correspond the current frame the result of which causes echoing effect. To overcome this problem, a Two Step noise reduction algorithm (TSNR) was used that traces immediately the non-stationarity of the signal without losing the benefit of the DD approach. The a priori SNR estimation was improved further for removing the bias ,thus discarding reverberation effect. The obtained output with TSNR is still subjected to harmonic distortions that are natural to all short time noise reduction techniques. The main reason is the defect in estimating the PSD in single channel systems. To solve this problem, a principle named Harmonic Regeneration Noise Reduction (HRNR) is applied that makes use of non-linearity for regenerating the missing harmonics. All the above mentioned algorithms are implemented and their performance was estimated in terms of both subjective and objective standards. There is a significant improvement in the performance using HRNR along with TSNR. Compared to other techniques, HRNR has the capability of restoration of missing harmonics.

Keywords: Directed Approach, Harmonic Regeneration, Speech Enhancement, Two-step Noise Reduction, Wiener Filtering

I. INTRODUCTION

Speech enhancement usually deals with processing of noisy speech signals in order for better understanding by humans or by decoding systems. Speech enhancement algorithms [1] concentrate on improving the performance of a system when its input speech is corrupted by noise. The simplest method to discard the ill effects of stationary background noise on clean speech is spectral subtraction [2]. In this algorithm, a spectral noise bias is subtracted from noisy speech spectral amplitude by calculating it from segments of speech inactivity while retaining the phase as it is. Secondary procedures succeed spectral subtraction, which decreases the unpleasant auditory effects due to spectral error. The limitation of spectral subtraction is that it can only be applied to speech corrupted by stationary noise. The technique of wiener filtering, adopted from communication theory has been applied to speech enhancement techniques for time varying properties of signal. This research aims at

ISBN: 978-93-86171-60-3

studying the spectral subtraction & wiener filter technique when speech degraded by non-stationary noise. The decision directed (DD) approach is used to estimate a priori SNR, a key parameter, resulted in better performance, both in terms of intelligibility and reduced musical noise. However, the estimated a priori SNR of the current frame is dependent on the estimated speech spectrum from the previous frame. The consequence of this mismatch is that the gain function doesn't correspond to the current frame resulting in a bias which causes unpleasant echoing effect. Therefore, a method called Two-step noise reduction (TSNR) [4] algorithm was used to solve the problem which tracks instantaneously the non-stationarity of the signal but, not by losing the advantage of the DD approach. The a priori SNR estimation was modified and made better by an additional step for removing the bias, thus eliminating reverberation effect. The output obtained even with TSNR still suffers from harmonic distortions, which are inherent to all short time noise suppression techniques, the main reason being the inaccuracy in estimating noise in single channel systems. To undo this problem, a concept called, Harmonic Regeneration Noise Reduction (HRNR) [5-6] is used wherein a non-linearity is made use of for regenerating the distorted/missing harmonics. All the above-discussed algorithms have been implemented and their performance evaluated using both subjective and objective criteria. The performance is significantly improved by using HRNR combined with TSNR, as compared to TSNR, DD alone, since HRNR ensures restoration of harmonics**.**

II. SPEECH ENHANCEMENT METHODS

2.1 SPECTRAL SUBTRACTION

Spectral subtraction [2], is one of the famous and widely used algorithm as it involves only a FFT and IFFT. This algorithm is very easy to implement with less complexity. The magnitude spectrum of the windowed data is calculated and the noise spectrum which is estimated from the segments of speech absence is subtracted. The resulting spectrum is biased down by the noise spectrum. Since the noise is approximated from the non-speech regions, there occurs spectral errors. This spectral error introduces what is called musical noise which is annoying to the listener. To minimize the musical noise the authors propose additional processing steps, like Magnitude averaging, half-wave rectification, residual noise reduction and additional signal attenuation.

Finally it is observed that noise reduction is done only in the magnitude spectrum whereas the phase remains the same. We assumed that the additive background noise is added digitally or acoustically to the speech and this noise remains locally stationary. Using Hanning window with an overlap of 50% the DTS is segmented into short frames. Then the magnitude spectrum of the windowed data is calculated and the noise spectrum which is estimated from the segments of speech absence is subtracted. The flow chart of spectral subtraction [8] is as shown in fig.1.

ISBN: 978-93-86171-60-3

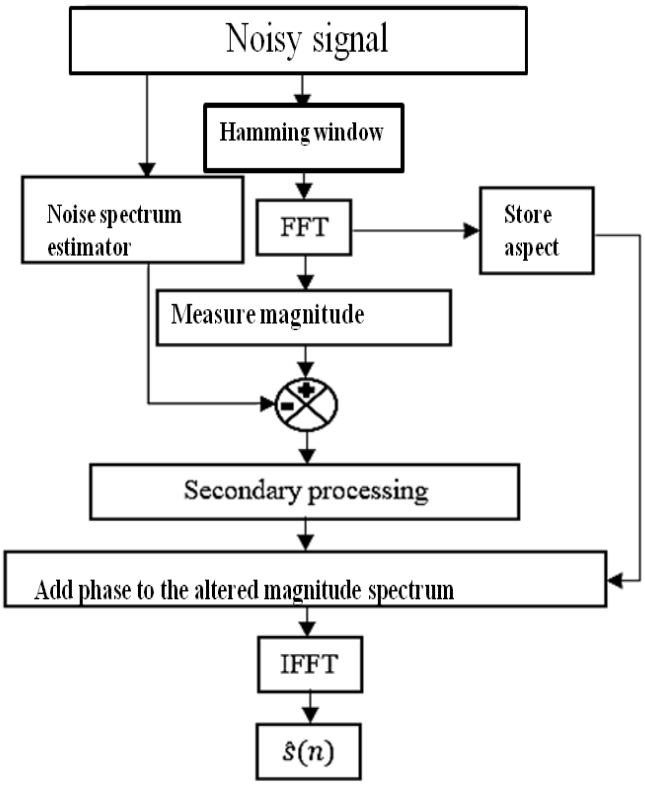


fig.1 spectral subtraction

2.1.1 Additive noise model

Considering a noisy speech signal x formed due to additive background noise d corrupting a clean speech s. It can be discretized and mathematically expressed as follows,

$$x(n) = s(n) + d$$

, the Short Time Fourier Transform of x is,

$$X(p,k) = S(p,k) + D(p$$

where,

$$X(p,k) = \sum_{m=-\infty}^{\infty} x(m)w(p-m)e^{-}$$

Where $w(n$ hanning window and

$X(p,k), S(p$ and $D(p$ stands for the k[th] spectral component of p[th] time window of x, s and d respectively.

2.1.2. Spectral subtraction estimator

Spectral subtraction estimator $\hat{S}(p$ is calculated by sending the clean speech through spectral subtraction filter $H(p$.

$$\hat{S} = H(p,k)X(p,k) \quad (4)$$

where, $H(p,k) = 1 - \frac{\mu(k)}{|X(p,k)|}$

ISBN: 978-93-86171-60-3

and $\mu(k) = E[\{|D(p,k)|\}]$

Substituting equations (5) and (6), equation (4) becomes,

$$\hat{S}(p,k) = [|X(p,k)| - \mu(k)]e^{j\theta_x(p,k)} \quad (7)$$

2.1.3. Spectral error

The difference between the clean speech and estimator is called as spectral error $\epsilon(p$ and is given by,

$$\epsilon(p,k) = \hat{S}(p,k) - S(p,k) \quad (8)$$

$$\epsilon(p,k) = D(p,k) - \mu(k)\,e^{j\theta_x} \quad (9)$$

For reducing the spectral error's auditory effects [2] include to further processing done: 1) Biasing down the noisy speech spectrum, 2) Reduction of Residual noise, and 3) Signal attenuation during speech absence.

2.1.4. Synthesis

Above mentioned all processing is done, noisy speech's phase was added to the processed magnitude spectrum and IFFT was applied to get the short time-domain signal. These short time frames are earlier divided with an overlap of 50%, the synthesized short time frames are added approximately with the overlap of 50% to construct the whole speech signal.

2.2 Wiener filtering

The optimal filter that minimizes the estimation error is called the Weiner Filter. In this, the mean square of the estimation error is commonly used as a criterion for minimization and the optimal filter coefficients can be derived in the time or frequency domain. Overall, the Weiner filters [8] are considered to be the linear estimators of the clean signal speech spectrum and they are optimal in the mean-square sense. In Wiener Filter the enhanced time-domain signal is obtained by sophisticating the noisy speech signal with a linear filter. Similarly in the frequency domain, the enhanced spectrum is obtained by multiplying the input (noisy) spectrum by the Weiner Filter.

Many methods are available for estimating the coefficients of clean speech, one such method based on MMSE estimation such as Wiener filter. If the noise is independent and additive with respect to speech, the minimization of $E\{(\hat{S}(p,k) - S(p,k)$ leads to,

$$G(p,k) = \frac{E\{|S(p,k)|^2\}}{E\{|S(p,k)|^2\} + E\{|D(p,k)|^2\}} = \frac{S\hat{N}R_{prio}(p,k)}{1 + S\hat{N}R_{prio}(p,k)} \quad (10)$$

The estimation of the *a priori* SNR, . $S\hat{N}R_{prio}(p$ is considered, which is required for the computation of G(p,k). $S\hat{N}R_{prio}(p$ is frequently estimated using the DD approach.

2.2.1 DD approach

In decision directed (DD) approach [5,6,8], a priori signal-to noise ratio (SNR) is the key parameter behind the reduction in musical noise. However, this estimated SNR is biased because *a priori* SNR of present frame follow the *a posteriori* SNR of previous frame. As a Consequence, the desired spectral gain was not achieved

ISBN: 978-93-86171-60-3

and the performance is decreased during speech activity. This bias is seen as a reverberation effect. This effect is eliminated by Two Step Noise Reduction method (TSNR).

Discrete time noisy signal is segmented into short-time frames using the Hamming window. In the additive noise model, the noisy speech is given by y(n) = s(n) + m(n) where s(n) and m(n) denote the speech and noise signal respectively. Let S(p, i), M(p, i) and Y(p, i) represent the ith spectral component of the short-time frame p of the speech signal s(n), noise m(n) and noisy speech y(n) respectively. The objective is to first derive an SNR estimate from the noisy features because no direct solution for the spectral estimation exists. An estimate of S (p, i) is obtained subsequently by applying a spectral gain H (p, i) to each short-time spectral component Y (p, i).

In practical implementations of speech enhancement systems, the PSDs of speech and noise are unknown since only the noisy speech spectrum Y(p, i) is available. Thus, both the *a posteriori* SNR and the *a priori* SNR have to be estimated.

The spectral gain H (p, i) is obtained by the function

$$H(p,i) = g\left(\hat{SNR}_{prio}(p,i), \hat{SNR}_{post}(p,i)\right) \quad (11)$$

The function *g* is chosen to be wiener filtering gain and the estimate of speech signal is obtained as

$$\hat{S}(p,i) = H(p,i)X(p,i) \quad (12)$$

Using the obtained noise PSD, the *a posteriori* and *a priori* SNRs are computed as follows:

$$\hat{SNR}_{post}(p,i) = \frac{|Y(p,i)|^2}{\hat{M}(p,i)} \quad (13)$$

$$\mathrm{SNR}_{\mathrm{prio}}^{\mathrm{DD}}(p,k) = \alpha \frac{\left[\left|\hat{S}(p-1,k)\right|^2\right]}{E\left[|D(p,k)|^2\right]} + (1-\alpha)P\left[\mathrm{SNR}_{\mathrm{post}}(p,k) - 1\right] \quad \text{for } 0 \le \alpha \le 1 \quad (14)$$

(4)

Where $\left[\left|\hat{S}(p-1,i)\right|^2\right]$ is the amplitude estimator of the ith spectral component of the (p −1) th frame, and the function *P* [·] is defined as *P*[*x*] = *x* if $x \geq 0$ and 0 otherwise and the parameter $\alpha = 0.98$.

Without the loss of generality, in the following the chosen spectral gain (function g in (4)) is the Wiener filter, and then the multiplicative gain function for this approach is

$$H_{DD}(p,i) = \frac{\hat{SNR}_{prio}^{DD}(p,i)}{1+\hat{SNR}_{prio}^{DD}(p,i)} \quad (15)$$

In DD algorithm, is the delay that is inherent and its effect, while speech transitions, i.e. onset and offset. This delay leads to a bias in estimating gain, which causes in a reverberation effect.

2.2.2 TWO STEP NOISE REDUCTION TECHNIQUE

TSNR [4] is a technique used to reduce to enhance the noise reduction performance and it delivers a greater SNR than the DD approach since the gain matches to the current frame whatever the SNR. Thus TSNR approach removes the drawbacks of DD approach while still maintaining the advantage i.e, highly reduced musical noise level. We propose a procedure in two steps in order to estimate the a priori SNR. The algorithm can be given in two steps.

In the first step using the DD approach spectral gain $G_{DD}(p, i)$ is computed. The second step includes estimation of the a priori SNR at frame p+1.

ISBN: 978-93-86171-60-3

Without loss of generality, spectral gain is given by

$$\widehat{SNR}_{prio}^{TSNR}(p,i) = \widehat{SNR}_{prio}^{DD}(p+1,i) = \beta' \frac{|H_{DD}(p,i)Y(p,i)|^2}{\gamma_n(p,i)} + (1-\beta')P[SNR_{post}(p+1,i)-1] \quad (16)$$

We chose β'=1 hence

$$H_{TSNR}(p,i) = \frac{\widehat{SNR}_{prio}^{TSNR}(p,i)}{1+\widehat{SNR}_{prio}^{TSNR}(p,i)} \quad (17)$$

This algorithm in two steps in the above equations (16) and (17) is called the TSNR technique.

TSNR approach is used to preserve speech onsets and offsets. TSNR technique successfully removes the annoying reverberation effect which is typical in DD approach. The reverberation effect can be reduced but cannot be suppressed but the TSNR approach makes it possible where the typical overlap is 50%. The a priori underestimation for high SNR which is introduced by DD approach due to delay is suppressed while the underestimation is preserved for low SNR in order to achieve the suppression of musical noise. Suppression of the a priori SNR estimation is also achieved in this approach. The only limitation of this approach is the presence of harmonics which can be removed by using the next technique i.e., HRNR technique [5].

2.2.3 Harmonic Regeneration And Noise Reduction

Harmonics is nothing but a signal whose frequency is an integral multiple of its reference signal. Usually in wireless communications all signals are transmitted such that it contains energy at its harmonic frequencies. We can get to know if the signal has energy at its harmonic frequencies by looking at the shape of the signal if no energy is stored at harmonic frequency the signal will be perfect sine wave if not the signal is not a perfect sine wave like saw tooth and square waves.

While transmitting these signals in which energy is stored at its harmonic frequencies gets attenuated and at the receiving side by using previous speech enhancement techniques these harmonics are usually considered as noise and they will be filtered and the signal information at the harmonic frequencies are lost, in order to prevent harmonics getting filtered we will be using a different method of speech enhancement technique called Harmonic Regeneration and Noise Reduction method.

In here, as the name says the harmonics are regenerated [5-6] in order to achieve that we need a non-linear function NF as non-linear function in time domain is used to restore harmonics.

This non-linear function is applied on time domain output signal of any of the previous speech enhancement techniques.

In here we will be using output of TSNR. By doing this all the harmonics which were attenuated earlier will be regenerated exactly at the same place as before.

$$s_{harmo}(t) = Nf_{(}\hat{s}(t)) \quad (18)$$

ISBN: 978-93-86171-60-3

Even though harmonics are regenerated this regenerated signal is not considered a clean speech signal as its amplitude is different from speech signal.

$$\widehat{SNR}_{prio}^{HRNR}(p,i) = \frac{\boldsymbol{\delta}(p,i)\left|\hat{S}(p,i)\right|^2 + \left(1-\rho(p,i)\right)\left|S_{harmo}(p,i)\right|^2}{\zeta(i)} \quad (19)$$

Where

$$\boldsymbol{\delta}(p,i) = H_{TSNR}(p,i) \quad (20)$$

This $\boldsymbol{\delta}$ parameter determines the mixing ratio of S (p, i) and $S_{harmo}(p,i)$.

Whenever the ouput signal given by TSNR is reliable this parameter is taken as 1 as the harmonics are not lost no regeneration process is required. If the output signal given by TSNR for harmonic regeneration is not reliable then it means harmonics are attenuated and it should be regenerated and hence this parameter takes up the value 0.

We choose $\boldsymbol{\delta}(p,i) = H_{TSNR}(p,i)$ to obtain this condition.

This $\widehat{SNR}_{prio}^{HRNR}(p,i)$ is used generate gain.

III. RESULTS

3.1 Objective Evaluation of Speech

Objective speech quality measures are generally calculated from the original undistorted speech and the distorted speech using some mathematical formulae. Some objective quality measures are highly correlated with subjective perceived quality, while others are more correlated with subjective intelligibility. All the algorithms discussed have been studied and implemented and their behavior have been analyzed for different kinds of background noises such as babble noise, car noise, added to clean speech with different SNRs.

To calculate the average segmental SNR [8] for measuring the performance of the implemented techniques, given by,

$$SNR_{seg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\sum_{l=Lm}^{Lm+L-1} s^2(l)}{\sum_{l=Lm}^{Lm+L-1}[\hat{s}(l)-s(l)]^2} \quad (21)$$

Where, M represents number of frames with active speech, and L denotes frame length. HRNR combined with TSNR technique achieves the best segmental SNRs for stationary and non-stationary noises the as shown in the Table 1.

The below table 1 shows the average segmental SNRs for the respective input SNRs of noisy speech and the obtained using all the techniques.

ISBN: 978-93-86171-60-3

Table 1: average segmental SNRs with different noise types, for various techniques implemented

Noise type	Input SNR (dB)	Spectral subtraction	DD approach	TSNR method	TSNR and HRNR
White noise	-5	8.54	6.94	8.72	9.8
	-2	10.5	9.72	11.20	12.04
	0	11.82	11.7	12.83	13.40
	2	13.09	13.68	14.35	45.27
	5	14.76	16.57	16.26	17.61
Car noise	-5	1.61	1.52	1.93	1.94
	-2	2.9	2.73	3.03	3.13
	0	3.98	4.04	4.36	4.47
	2	5.44	5.29	5.59	5.77
	5	7.06	7.28	7.37	7.72
Babble noise	-5	0.42	0.28	1.54	1.75
	-2	0.93	0.76	2.10	2.23
	0	1.48	1.26	2.58	2.67
	2	2.26	1.98	3.43	4.39
	5	3.59	3.43	4.79	4.78
Helicopter noise	-5	1.67	1.66	4.13	4.12
	-2	3.16	2.90	4.99	5.22
	0	4.46	3.97	5.82	6.12
	2	5.45	5.24	6.78	7.08
	5	7.26	7.40	8.54	8.75

3.2 Subjective Evaluation of Speech

Subjective quality measures are measures based on the subjective opinion of a panel of listeners on the quality of the speech output file. Listening tasks involved sentence recognition in noise. Speech intelligibility was assessed in terms of percentage of words identified correctly. The subjective measure tests are Mean Opinion Score (MOS). Conducted the test with 5 different listeners, who were asked to listen to enhanced speech using different scenarios, and then were asked give scores from 1 to 5 for the parameters listed in the below table 2. A score of 1 denotes poor and 5 represents excellent. The average scores for enhanced speech using TSNR combined with HRNR were better, while the scores for other techniques were considerably poorer and hence are not mentioned in the paper.

ISBN: 978-93-86171-60-3

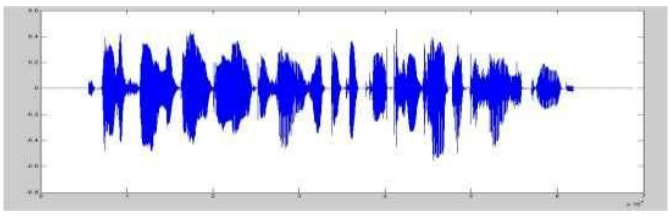
Table 2: Average Subjective test score for 5 persons

Noise type	Parameters	Input global SNR (dB)				
		-5	-2	0	2	5
White noise	Musical noise	4.2	4.4	4.8	5	5
	Intelligibility	4.5	4.6	4.7	5	5
	Quality	3.3	3.8	4.3	4.8	5
Car noise	Musical noise	2.6	3.1	3.6	3.9	4.2
	Intelligibility	3.2	3.2	3.5	4.0	4.5
	Quality	2.1	2.9	3.2	3.7	3.9
Babble noise	Musical noise	2.8	3.2	3.5	3.9	4.1
	Intelligibility	2.5	2.8	3.2	3.5	4.2
	Quality	2.5	2.9	3.3	3.9	4.2
Helicopter noise	Musical noise	3.1	3.4	3.8	3.9	4.2
	Intelligibility	3.3	3.3	3.8	4.1	4.2
	Quality	1.7	2.4	2.9	3.5	4.2

3.3 Spectrogram Analysis

For demonstration we are using a typical speech file "*In the fall of 1996 he took a coaching job at high school".*

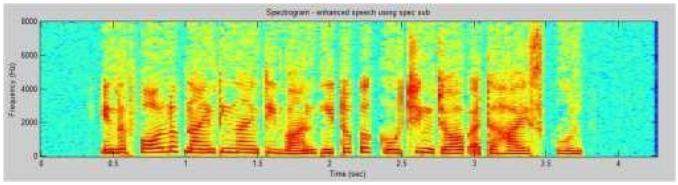
The time waveform of clean speech and the corresponding spectrogram is shown in the below figure.



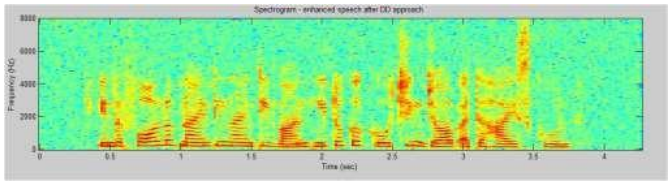
Enhancement of speech corrupted with 0dB white noise

i. Spectrogram obtained by using spectral subtraction.

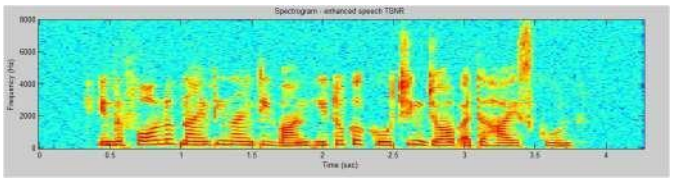
ISBN: 978-93-86171-60-3



ii. Spectrogram obtained by using DD method.



iii. Spectrogram obtained by using TSNR method.



iv. Spectrogram obtained by using HRNR method.

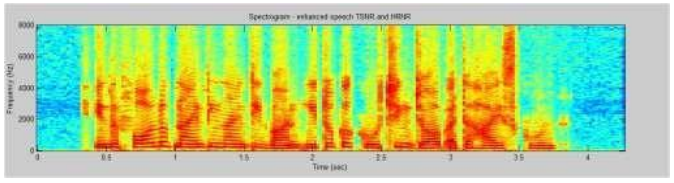


Fig 2: Spectrogram of noisy speech, enhanced speech using (i) spectral subtraction, (ii) DD approach, (iii) TSNR method, and (iv) TSNR method followed by HRNR method

IV. CONCLUSION

In this paper, a study on different noise reduction techniques was presented and their performance for different noise types and SNR's was determined. In spectral subtraction method, an estimate of noise spectrum is assessed from segments of speech absence and it is subtracted from noisy speech spectrum. However, this method is not productive for speech computed with non-stationary noise such as car noise, helicopter noise and babble noise. In Weiner filtering method, the multiplicative gain is estimated as a function of a priori SNR. The major setback for the DD algorithm is the frame delay which subsequently leads to reverberation effect giving rise to a need for a better method. To overcome the drawback of DD method, the TSNR technique was introduced. It comprises mainly of two steps, in the first step the musical noise was reduced and the second step removes the frame delay by preserving the speech transitions.

This TSNR technique accomplished so well in terms of reducing noise but brings a harmonic distortion because of the errors in evaluating noise PSD. To sort out this problem i.e to restore the missing harmonics back, a non-linearity was applied to produce an artificial signal in time domain. This signal was used for refining the a priori SNR that was used for computing the spectral gain. The results are presented for evaluation

ISBN: 978-93-86171-60-3

of performance of different techniques. The results clearly indicate that TSNR followed by HRNR technique is proved to be the best among the other techniques in terms of both subjective and objective tests

V. ACKNOWLEDGMENT

The authors are thankful for the management and the Principal of the Atria institute of Technology for carrying out this work in the institute. The authors are highly grateful for the financial grant given to the institute through K-FIST (L1) for FY 2014 for infrastructure development for the Advanced Signal Processing Laboratory to carry out this work.

REFERENCES

[1] Elsevier, Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016) Comparison of Speech Enhancement Algorithms, Siddala Vihari, A. Sreenivasa Murthy, Priyanka Soni and D. C. Naik .

[2] S F Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSp-27, No. 2, pp. 113-120, Apr.1979.

[3] P. Scalart, and J. Vieira Filho, "Speech Enhancement Based on a Priori Signal to Noise Estimation," IEEE Int. Conf. on Acoustics, Speech and Signal Proc., pp. 629–632, 1996.

[4] Cyril Plapous, Claude Marro1, Laurent Mauuary, Pascal Scalart, "A Two-Step Noise Reduction Technique" 2004.

[5] Cyril Plapous, Claude Marro, Pascal Scalart. Speech enhancement using harmonic regeneration. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2005), Mar 2005, Philadelphia, United States. 2005.

[6] Cyril Plapous, Claude Marro, Pascal Scalart. Improved Signal-to-Noise Ratio Estimation for Speech Enhancement. IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers, 2006.

[7] 2. Y. Ephra¨ım, and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109–1121, Dec. 1984.

[8] PiliposC.Loizou (2013). Speech Enhancement Theory and Practice, 2nd Edition, CRC press.