Chapter 6

Hybrid Algorithm for Performance Improvement

It was suggested in last chapter that for better performance the STSA algorithm can be combined in some way with RASTA approach. The best performing STSA algorithm is MMSESTSA85 (MMSE-LSA) as discussed in chapter 4. It is combined with modified RASTA multiband filter approach which is evaluated in chapter 5. The hybrid algorithm is proposed here and also it is simulated and tested under different additive noise conditions using the NOIZEUS database and compared with the original algorithms. The results of performance evaluation using objective measures are described in this chapter. The comparison using alone objective measures is not sufficient as it will not ensure the quality of speech signal for human listeners and hence the subjective evaluation is also required to perform. The IEEE recommended and ITU-R BS.562-3 standard mean opinion score (MOS) listening test is carried out. The chapter describes the various guidelines followed to perform this test. The original and modified algorithms are compared based on this test and conclusion is made regarding quality of output of different algorithms.

Reverberation is one type of convolutive distortion that occurs commonly in communication systems. The speech enhancement algorithm must be able to tackle it. The proposed algorithm is also tested under different reverberation condition using the Aachen impulse response (AIR) database developed by RWTH Aachen University, institute of communication systems and data processing (India). It is a set of impulse responses that were measured in a wide variety of rooms. This database allows realistic studies of signal processing algorithms in reverberant environments. The comments are made about performance of algorithms in the simulated reverberant conditions.

6.1 Proposed New Approach

The proposed modified approach for speech enhancement uses combination of MMSE STSA85 algorithm and multiband RASTA filter. The connection is not simple cascade but the blocks are interacting as shown in figure 6.1. The noisy speech is presented simultaneously to both multiband RASTA and MMSE STSA85 algorithms. The VAD is required to estimate speech/silence segment for MMSE STSA85 algorithm. This block is responsible for malfunctioning of algorithm if the detection is false. The MMSE STSA85 algorithm is highly dependent of VAD false rate. So VAD is not directly getting the noisy speech for estimation but the output of multiband RASTA filter is given to VAD for estimation. The RASTA approach does not require VAD and reduce the noise moderately as discussed in chapter 5. Some speech

distortion and musical and residual noise remain in enhanced speech by RASTA algorithm. However, the VAD can now better detect the speech/silence segment compared to direct detection from noisy speech. But the white noise after RASTA filtering gets converted into colored noise with sharp spectral peaks. Hence, the accuracy in noise estimation reduces; this causes the rise in musical noise. So the noise power is estimated for RASTA filtered as well as original noisy speech spectrum. The ratio of original noise power to the filtered noise power (PR) is calculated and it is used to calculate a priori SNR. A mild linear compression is required to avoid over suppression. The modified decision direct rule taking this factor into consideration is given by following equation for frame *t*.

$$\xi^{(t)}(K) = \eta \frac{\left|\hat{X}^{(t-1)}(K)\right|^2}{\left|\hat{D}^{(t)}(K)\right|^2/PR} + (1-\eta)max(\bar{\gamma}^{(t)}(K) - 1,0) \tag{6.1}$$

$$where;\ \bar{\gamma}(K) = \frac{|\bar{Y}(K)|^2}{\left|\overline{\hat{D}}(K)\right|^2/\mathrm{PR}}$$

The enhanced speech obtained after this modification has almost no musical noise.

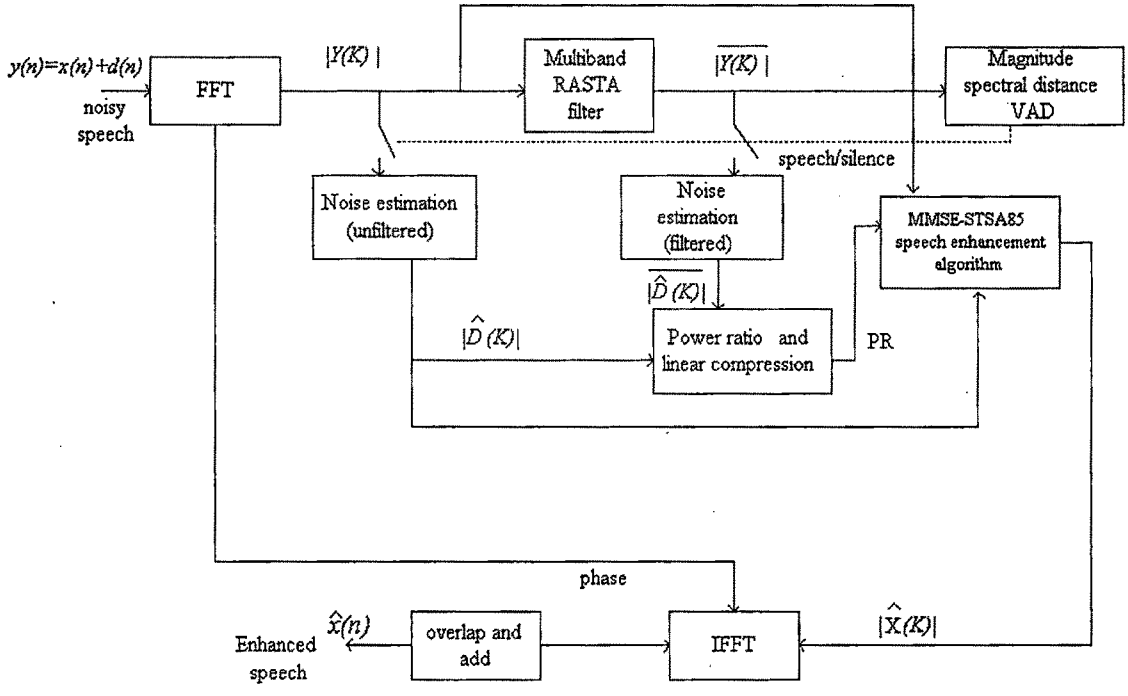

Fig. 6.1 Block diagram of proposed speech enhancement method

6.2 MATLAB Implementation of Proposed Algorithm

The input speech sampled at 8 KHz is applied to 32ms hamming window with 50% overlap and 256 point FFT is applied. From complex FFT the magnitude and phase are separated. Due to symmetry property 128 point spectral values are filtered using multiband RASTA with nonlinear compression parameter a=3/4 and expansion parameter b=4/3. The filter is initialized with zero values. The filtered input speech spectrum is used by magnitude spectral distance VAD to identify the current frame as speech/silence. If the current frame is silence frame, the filtered as well as unfiltered noise estimate is updated by using noise estimation rule described in section 4.2. The power ratio is calculated and linear compression is applied to avoid over suppression. The linear compression is implemented using straight line equation and it ensures the ratio to be between 1 and 2. The actual speech enhancement is performed by MMSE STSA85 method. The enhanced spectral values are combined with the phase of the noisy spectrum. 256 point IFFT is applied and overlap add synthesis is performed to reconstruct the speech signal as final output.

6.3 Spectrographic and Objective Evaluation of Proposed Algorithm

The spectrogram of the enhanced speech for the clean speech with spectrogram shown in figure 4.5 and subjected to 0dB white noise is enhanced by the proposed approach. The spectrogram of the proposed approach is shown in figure 6.2. Comparison of this with the spectrograms of speech enhanced by MMSE STSA85 (figure 4.6) and with the modified RASTA filter (figure 5.16) indicates that the speech enhanced by using proposed approach more closely resembles to the clean speech signal. Still there are some randomly distributed spots present in the enhanced speech spectrogram which results in small level of musical noise. The residual noise is very less compared to two original algorithms.

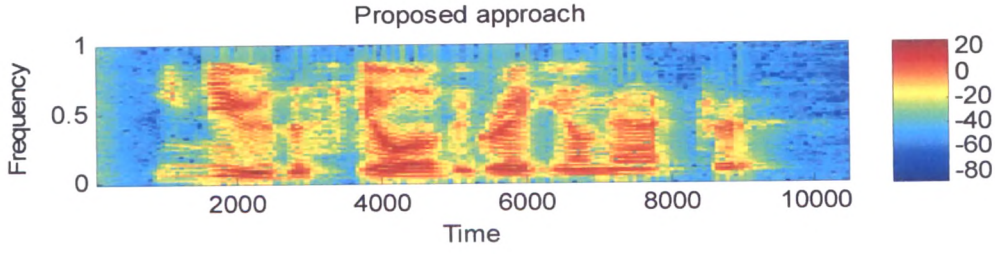

Fig. 6.2 Spectrogram of enhanced speech signal using proposed approach

The objective quality measures SSNR, WSS, LLR and PESQ observed over 0dB, 5dB and 10 dB SNRs using NOIZEUS database [1] are given in figures 6.3 to 6.7 in the form of bar chart. As mentioned in section 4.5 the number of test runs on each algorithm is 810. The comparison of proposed approach is done with MMSE STSA85 and Modified RASTA filter algorithms. The quality measures for noisy speech and maximum theoretical limits (obtained by using clean magnitude and noisy phase) are also included for comparison.

The WSS measure indicates the spectral distortion in the speech and the comparison shows that in all types of noise conditions at 0 and 5dB SNRs the proposed algorithm gives good improvement. Except in white noise and airport noise condition at 10 dB SNR the WSS for proposed algorithm is improved in all other noises at 10 dB SNR. The LLR measure is better for proposed approach in all noises at all SNRs compared to MMSE STSA85 algorithm. The PESQ score at 0dB SNR is comparable with MMSE STSA85 algorithm in most of the cases and in few cases it shows improvement. For restaurant noise it is noticeably improved. For 5 dB SNR this measure slightly degrades compared to original MMSE STSA85 algorithm in all cases but it is marginal. For 10 dB SNR this measure shows some degradation in all cases. Putting these results altogether; it is noticed that at low SNR levels like 0 to 5dB the proposed approach gives better performance while at higher SNR levels ($\geq$10 dB) the original MMSE STSA85 algorithm performs better. Also the proposed algorithm outperforms the original algorithm in car noise, restaurant noise and train noise conditions. In these kinds of noisy environments the person using communication equipment has to combat with surroundings from the confined area only and the SNRs in such situation are always weak. As the primary goal of this research work is to design an algorithm for low SNR conditions the proposed approach is recommended to use in such circumstances.[1] However, the comments made here are still based on objective measures only; but this needs to correlate well with subjective listening tests which involves the human beings. For that it is required to do the subjective evaluation of algorithms [2]. The procedure and the experiment conducted for this purpose is explained in next section.

[1] A paper entitled "Objective Evaluation of STSA Based Speech Enhancement Techniques for Speech Communication Systems with Proposed" is presented in IEEE International conference on Communication, Network and Computing (CNC 2010) Organized by ACEEE at Calicut in October 2010. IEEE CS- CPS ISBN: 978-0-7695-4209-6. Listed in IEEE Xplore by IEEE Computer Society, DOI:10.1109/CNC.2010.13, pp.19-23. Archived in ACM digital library.

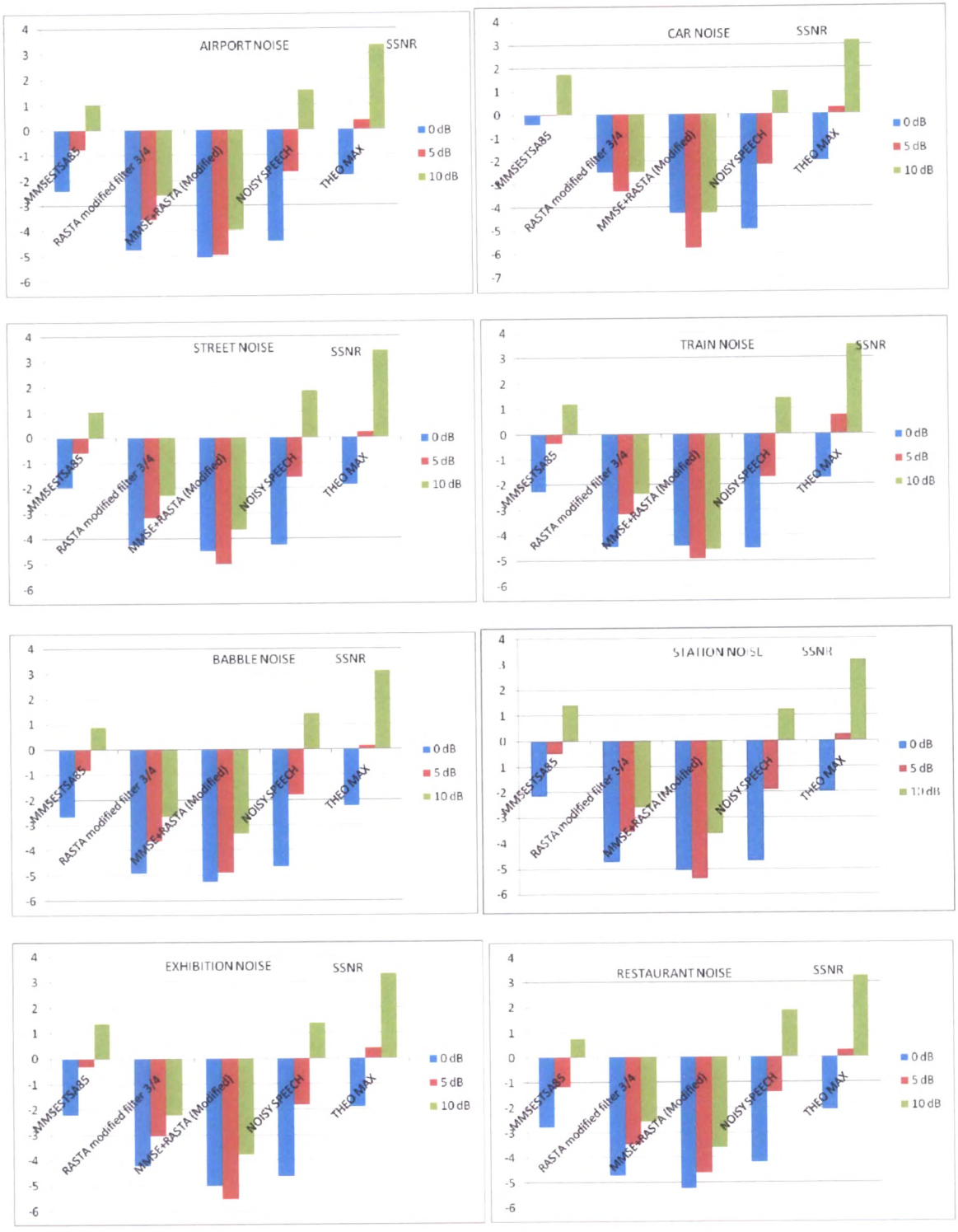

Fig. 6.3 SSNR comparison of proposed algorithm over NOIZEUS database

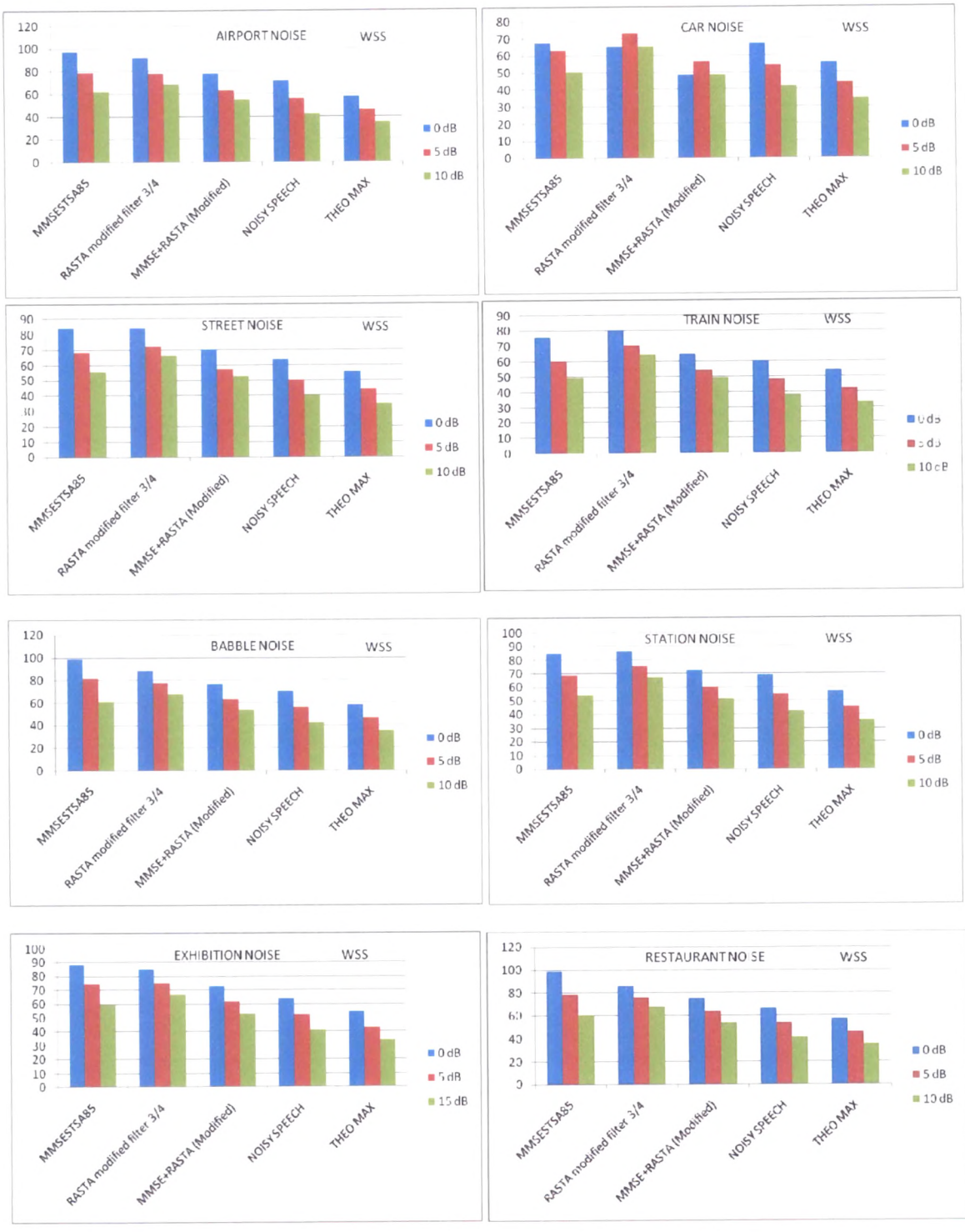

Fig. 6.4 WSS comparison of proposed algorithm over NOIZEUS database

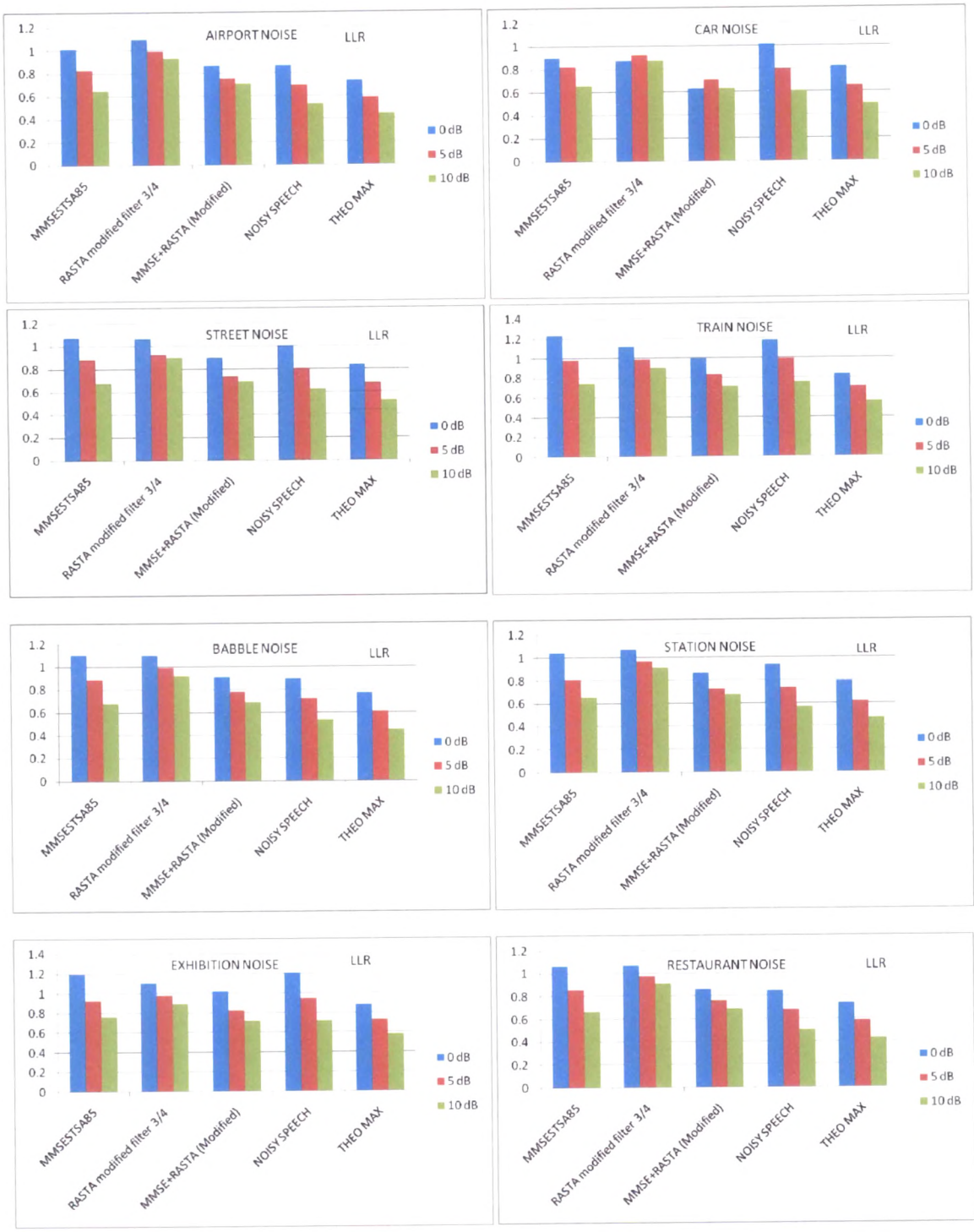

Fig. 6.5 LLR comparison of proposed algorithm over NOIZEUS database

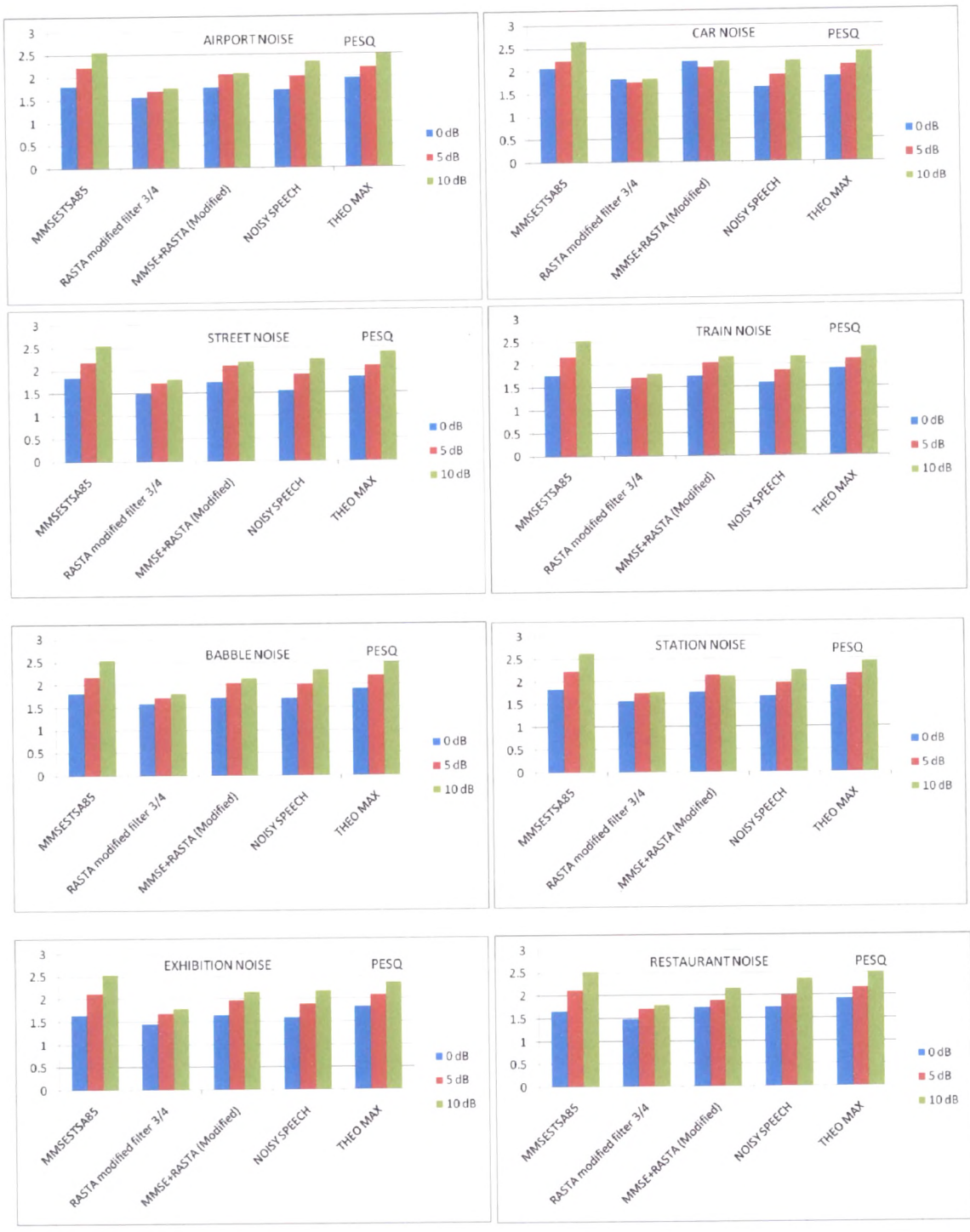

Fig. 6.6 PESQ comparison of proposed algorithm over NOIZEUS database

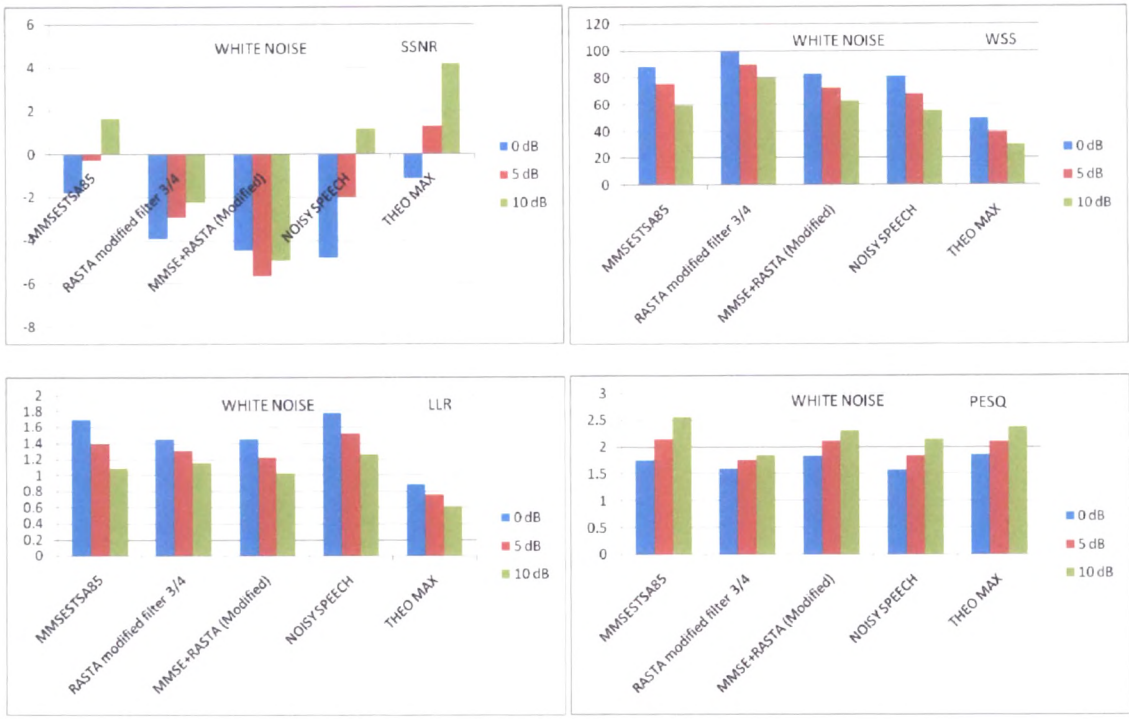

Fig. 6.7 Objective evaluation of proposed algorithm under white noise

6.4 Subjective Evaluation

Subjective evaluation involves comparisons of original and processed speech signals by a group of listeners who are asked to rate the quality of speech along a predetermined scale. The most widely used direct method of subjective quality evaluation is the category judgment method in which listeners rate the quality of the test signal using a five-point numerical scale as shown in table 6.1, with 5 indicating "excellent" quality and 1 indicating "unsatisfactory" or "bad" quality. This method is one of the methods recommended by IEEE subcommittee on Subjective Methods [3] as well as by ITU [5, 6]. The measured quality of the test signal is obtained by averaging the scores obtained from all listeners. This average score is commonly referred to as the Mean Opinion Score (MOS).

The MOS test is administered in two phases: training and evaluation. In the training phase, listeners hear a set of reference signals that exemplify the high (excellent), the low (bad), and the middle judgment categories. This phase, also known as the "anchoring phase," is very important as it is needed to equalize the subjective range of quality rating of all listeners- that is, the training phase should in principle equalize the "goodness" scales of all listeners to ensure, to

the extent possible, that what is perceived as "good" by one listener is also perceived as "good" by the other listeners. A standard set of reference signals need to be used and described when reporting the MOS scores [3]. In the evaluation phase, subjects listen to the test signal and rate the quality of the signal in terms of the five quality categories (1-5) shown in table 6.1. Reference signals can be used to better facilitate comparison between MOS tests conducted at different times, different laboratories, and different languages [4].

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Table 6.1 MOS rating scale

Detailed guidelines and recommendations for administering the MOS test can be found in the ITU-R BS.1116-1 standard [5] and include:

1. *Selection of listening crew*: Different number of listeners is recommended, depending on whether the listeners have had extensive experience in assessing sound quality. Minimum number of non expert listeners should be 20, and minimum number of expert listeners should be 10.
2. *Test procedure and duration*: Speech material (original and degraded) should be presented in random order to subjects, and the test session should not last more than 20 minutes without interruption. This step is necessary to reduce listening fatigue.
3. *Choice of reproduction device*: Headphones are recommended over loudspeakers, as headphone reproduction is independent of the geometric and acoustic properties of the test room. If loudspeakers are used, the dimensions and reverberation time of the room need to be reported.

6.5 Setup for Subjective Evaluation

For subjective evaluation four algorithms namely MMSE STSA85, wavelet de-noising, modified RASTA filter and proposed algorithm (combination of MMSE STSA85 and modified RASTA filter) are selected. The speech sentences from NOIZEUS database are selected contained in files sp02.wav (male speaker) and sp11.wav (female speaker) mentioned in table

4.1. The speech sentences corrupted by white noise, restaurant noise, car noise and airport noise at 0, 5 and 10dB SNRs are selected for enhancement. As mentioned in section 6.4 the performance of proposed algorithm is very good in car noise and restaurant noises but inferior in white noise and airport noise conditions compared to original MMSE STSA85 algorithm. So, all these four types of noises are selected for subjective evaluation. However, it can be extended for all other types of noises but to complete the test as per guidelines mentioned in section 6.5 within stipulated time the restrictions are applied.

For conducting the MOS test following procedure is obeyed:

1. *Selection of listening crew*: Total 20 listeners are selected having age in between 19 years to 38 years. It includes 9 undergraduate final year Electronics and communication engineering students, 9 faculty members of Electronics and communication engineering department and 2 laboratory assistants from S.V.M. Institute of Technology, Bharuch. The crew includes 13 male and 7 female listeners.
2. *Test procedure and duration*: The listeners are presented with clean speech file, noisy speech file and enhanced speech file by each algorithm. The care is taken when the enhanced speech files are named so that the identity of the algorithm remains undisclosed. The file names are not reflecting the type and name of algorithm by any means. The listeners are having freedom to play the clean, noisy and enhanced speech files at any time during the test. This is done to eliminate the overlay effect of the previously listened speech.
3. *Choice of reproduction device*: Good quality headphones are provided to each listener. The test is conducted in project laboratory of electronics and telecommunication engineering department of S.V.M. Institute of Technology, Bharuch in quiet environment.

The pro forma for filling up the MOS test score for different algorithms is shown in figure 6.8.

Subjective Evaluation (MOS) Test for Speech Enhancement Algorithms

Venue: Project Lab, EC Dept., SVMIT, Bharuch

Date:

Name of the listener: Group: Time:

Clean speech file name:

Type of Noise: AWGN

Algorithm	0 dB	5dB	10dB
E1			
E2			
E3			
E4			

Type of Noise: RESTAURANT

Algorithm	0 dB	5dB	10dB
E1			
E2			
E3			
E4			

Type of Noise: CAR

Algorithm	0 dB	5dB	10dB
E1			
E2			
E3			
E4			

Type of Noise: AIRPORT

Algorithm	0 dB	5dB	10dB
E1			
E2			
E3			
E4			

(Signature of the listener)

Fig. 6.8 Pro forma for filling up the MOS

6.6 Subjective Evaluation of Proposed Algorithm

Figure 6.9 shows the MOS test results obtained for various algorithms. The comparison shows that wavelet de-noising is the worst algorithm in all four algorithms. The proposed algorithm give high MOS scores for 0 and 5dB SNRs in all noise conditions. For 10dB SNR the performance of proposed algorithm is comparable with the original MMSESTSA85 algorithm. Hence the proposed algorithm performs well in low SNR conditions compared to original algorithm. This validates the results obtained from objective measures.

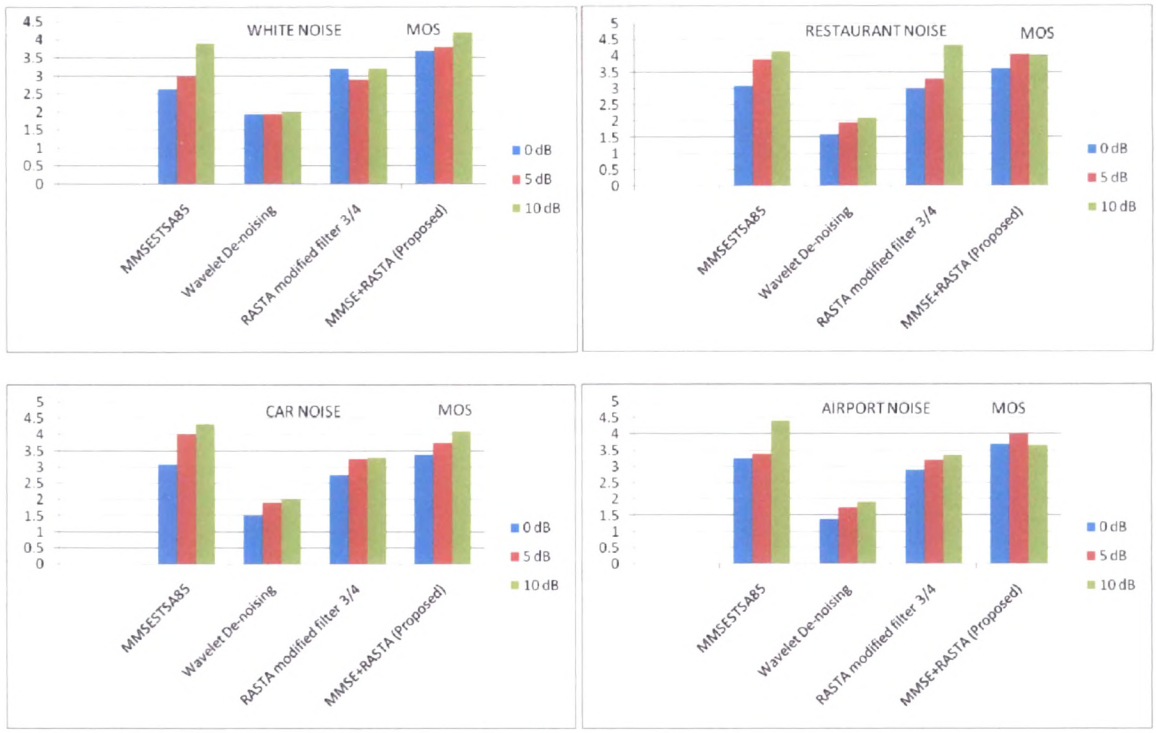

Fig. 6.9 Results of MOS test

6.7 Evaluation of Proposed Algorithm in Reverberant Environments

So far the objective and subjective evaluation is carried out using NOIZEUS database which contains the speech sentences corrupted with additive noise. In real circumstances the additive noise is not only the corrupting factor but some reverberation is also present. For wireless mobile communication systems the reverberant environment will change as the user moves from place to place. Hence it is required to test the proposed algorithm under different reverberant conditions. To test the algorithm in simulated reverberation environment a database called the Aachen Impulse Response (AIR) database is used [7].

It is a set of impulse responses that were measured in a wide variety of rooms. The initial aim of the AIR database was to allow for realistic studies of signal processing algorithms in reverberant environments with a special focus on hearing aids applications. It offers binaural room impulse responses (BRIR) measured with a dummy head in different locations with different acoustical properties, such as reverberation time and room volume. Besides the evaluation of de-reverberation algorithms and perceptual investigations of reverberant speech, this part of the database allows for the investigation of head shadowing influence since all recordings where made with and without the dummy head. Since de-reverberation can also be applied to telephone speech, it also includes (dual channel) impulse responses between the artificial mouth of a dummy head and a mock-up phone. The measurements were carried out in compliance with the ITU standards for both the hand held and the hands free position.

A MATLAB reference implementation is available at [7]. All impulse responses of the AIR database are stored as double precision binary floating point MAT-files which can be directly imported into MATLAB.

Table 6.2 shows the parameters to be specified to obtain a particular room impulse response. The clean speech signal can be convolved with this impulse response to generate the reverberant speech in particular environment. Table 6.3 specifies the combination of parameters used in the evaluation of proposed algorithm.

Parameter	Structure of parameter
Type of impulse response	rir_type '1': binaural (with/without dummy head) acoustical path: loudspeaker -> microphones next to the pinna '2': dual-channel (with mock-up phone) acoustical path: artificial mouth of dummy head-> dual-microphone mock-up at hand held or hands free position
Room type	room 1,2,..,10: 'booth', 'office', 'meeting', 'lecture', 'stairway','stairway1','stairway2', 'corridor','bathroom','lecture1' Available rooms for (1) binaural: 1,2,3,4,5 (2) phone: 2,3,4,6,7,8,9,10
Select channel	channel '0': right; '1': left
Select RIR with or without dummy head (for 'rir_type=1' only)	head '0': no dummy head; '1': with dummy head
Position of mock-up phone (for 'rir_type=2' only)	phone_pos '1': HHP (Hand-held), '2': HFRP (Hands-free)
RIR number (increasing distance, for 'rir_type=1' only)	rir_no Booth: {0.5m, 1m, 1.5m} Office: {1m, 2m, 3m} Meeting: {1.45m, 1.7m, 1.9m, 2.25m, 2.8m} Lecture: {2.25m, 4m, 5.56m, 7.1m, 8.68m, 10.2m} Stairway: {1m, 2m, 3m}

Table 6.2 Specification of parameters for generation of impulse response

Reverberant Environment	rir_type	Room	phone_pos	rir_no
Reverb1	dual-channel	Office	Hands-free	3m
Reverb2	binaural	Booth	Hand-held	1m
Reverb3	binaural	Meeting	Hands-free	2.25m
Reverb4	dual-channel	Bathroom	Hands-free	------
Reverb5	dual-channel	lecture1	Hands-free	2.25m

Table 6.3 Set of parameters for testing proposed algorithm in reverberant environments

Table 6.4 shows the comparison of MMSE STSA85 and proposed algorithm under the simulated reverberation environments. The table clearly indicates that the WSS and LLR score

for the proposed algorithm is very less compared to MMSE STSA85 which means very less distortion present in the enhanced speech. The PESQ score also shows improvement for proposed algorithm in all different reverberation conditions. However, the reverberation is not much bother to human listener as intelligibility is preserved in the reverberant speech; but it is much more significant for automatic speech recognizers. Hence the proposed algorithm can be used in speech communication systems as well as preferred as a preprocessing stage in ASR.

Reverberant Environment	Algorithm	SSNR	WSS	LLR	PESQ
Reverb1	MMSE STSA85	-0.9959	55.4408	0.9366	2.4614
	Proposed	-8.9914	34.6862	0.5034	2.6324
Reverb2	MMSE STSA85	-9.8105	47.9322	0.7945	3.1278
	Proposed	-8.4320	31.7140	0.2780	3.3966
Reverb3	MMSE STSA85	-9.8257	53.4960	0.8777	2.6979
	Proposed	-8.8671	37.8436	0.5962	2.8037
Reverb4	MMSE STSA85	-0.8166	62.3047	0.9325	2.7068
	Proposed	-8.8137	39.5675	0.5355	2.7890
Reverb5	MMSE STSA85	-0.4522	46.1798	0.8741	2.7540
	Proposed	-9.2058	24.8601	0.4645	2.9386

Table 6.4 Objective evaluation of proposed algorithm in reverberant environments

6.8 Summary

The combination of STSA and RASTA approach is termed here as hybrid approach which is proposed algorithm to improve the performance at lower SNRs (0-5dB). The performance evaluation using objective measures shows the improvement at lower SNRs compared to original STSA algorithm. The subjective listening tests also back the result. The proposed algorithm also found more superior compared to original algorithm under reverberant environments. Hence it is recommended to use hybrid approach in low SNR conditions and reverberant environments. However, the RASTA algorithm is non linear and non causal which throws the challenge for real time and hardware implementation. This is dealt in next chapter.