Real-time Enhancement of Noisy Speech Using Spectral Subtraction

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Master of Technology

by

Santosh Kumar Waddi
(10307932)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology Bombay
June 2013

Indian Institute of Technology Bombay

M. Tech. Dissertation Approval

This dissertation entitled **"Real-time Enhancement of Noisy Speech Using Spectral Subtraction"** by **Santosh Kumar Waddi** (Roll No. **10307932**) is approved, after the successful completion of *viva voce* examination, for the award of the degree of **Master of Technology** in **Electrical Engineering.**

Supervisor	.....	(Prof. P. C. Pandey)
Examiners	.....	(Prof. P. Rao)
	.....	(Prof. M. S. Shah)
Chairperson	.....	(Prof. P. Rao)

Date: 27 June 2013

Place: Mumbai

Declaration

I declare that this dissertation represents my ideas in my words and where ideas or words are taken from others, I have adequately cited and referenced the original sources. I declare that I have adhered to all principles of academic honesty and integrity and I have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

(Santosh Kumar Waddi)

Date: 27 June 2013

Place: Mumbai

Santosh K. Waddi / Prof. P. C. Pandey (Supervisor): “Real-time enhancement of noisy speech using spectral subtraction”, *M. Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology Bombay, June 2013.

ABSTRACT

Persons with sensorineural loss experience great difficulty when the speech is contaminated by noise. This thesis presents investigations for real-time enhancement of noisy speech using spectral subtraction for suppressing the external noise in hearing aids and sensory aids for the hearing impaired. Investigation using offline processing for enhancing the noisy speech with different types of noise and SNR values is carried out to select the optimal set of steps and parameters for real-time processing. Results show that median based noise estimation is effective in estimating noise from noisy speech without a voice activity detector, for different SNRs and types of stationary and non-stationary noises. It is shown that a cascaded-median can be used as an approximation to median for significantly reducing the computation and memory requirement. Speech enhancement using magnitude spectrum subtraction with 3-point 4-stage cascaded median for noise estimation and resynthesis using noisy phase resulted in improvements of 0.11 – 0.43 in PESQ scores for speech material from NOIZEUS database and different types of additive stationary and non-stationary noises at 6 dB SNR. Resynthesis using phase estimated from the enhanced magnitude spectrum did not result in any further improvement in the scores. The technique is implemented and tested for satisfactory operation, with sampling frequency of 10 kHz, 30 ms analysis window with 50% overlap, using a DSP board based on 16-bit fixed-point processor with on-chip FFT hardware. The implementation uses data transfer and buffering operations devised for an efficient realization of analysis-synthesis and codec and DMA for acquisition of the input signal and outputting of the processed output signal. The real-time operation is achieved with signal delay of approximately 48 ms and using about one-seventh of the computing capacity of the processor.

CONTENTS

List of Abbreviations

Abbreviation	Explanation
ADC	analog-to-digital converter
CMBNE	cascaded-median based noise estimation
DAC	digital-to-analog converter
DFT	discrete Fourier transform
DMA	direct memory access
DSP	digital signal processor
FFT	fast Fourier transform
I/O	input/output
IDFT	inverse discrete Fourier transform
IFFT	inverse fast Fourier transform
IMCRA	improved minima controlled recursive averaging
LSSE	least square error estimation
MBNE	median based noise estimation
MOS	mean opinion score
MSBNE	minimum statistics based noise estimation
MSE	mean square error
NOIZEUS	noisy speech corpus
PC	personal computer
PESQ	perceptual Evaluation of Speech Quality
QBNE	quantile based noise estimation
RMS	root mean square
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
USB	Universal Serial Bus
VHSES	vowel, Hindi sentence, English sentence

List of Symbols

Symbols	Explanation
b_{ei}	ending frequency sample of the *i*th band
b_{si}	beginning frequency sample of the *i*th band
$D_n(k)$	estimated noise magnitude spectrum
f_i	upper frequency of the *i*th band
F_s	sampling frequency
$H_{GA}(\xi, \mu)$	suppression function
L	window length
N	size of the DFT
S	window shift
$x(n)$	noisy speech signal
$y(n)$	enhanced speech signal
α	over-subtraction factor
β	spectral floor factor
γ	exponent factor
δ_i	tweaking factor for *i*th band
ζ_{dB}	relative RMS error
μ	posteriori SNR
ξ	priori SNR

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Problem overview

Hearing aids generally provide frequency-selective amplification to compensate for the elevated hearing thresholds. Hearing aids for persons with sensorineural loss employ multi-channel dynamic range compression with configurable attack time, release time, number of channels, and compression ratios to compensate for the reduced dynamic range [1]–[4]. Sensorineural impairment is also associated with increased spectral masking due to widened auditory filters. Several techniques, such as binaural dichotic presentation [5], [6], spectral contrast enhancement [7], and multiband frequency compression [8], [9], have been reported for reducing the adverse effect of increased spectral masking on speech perception. Despite these advances, hearing aid users with sensorineural impairment experience great difficulty in speech perception in noisy environments. Similar difficulty is faced by users of cochlear prostheses and other sensory aids for the hearing impaired [1]. Use of a second microphone in these aids to provide reference input for noise suppression by adaptive filtering is impractical. Hence, single-input noise suppression is the most practical solution for improving speech quality and intelligibility.

The noise suppression technique should have low algorithmic delay and low computational complexity to permit its implementation on a low-power processor in a sensory aid. Spectral subtraction is a single-input speech enhancement technique developed for use in audio codecs and speech recognition [10]–[20]. It involves estimating the noise spectrum, subtracting it from the noisy speech spectrum, and re-synthesizing the speech signal. As the interfering noise is non-stationary, its spectrum needs to be dynamically estimated. Under-estimation of the noise results in residual noise and its over-estimation results in distortion leading to degraded quality and reduced intelligibility. Noise can be estimated during the silence intervals identified by voice activity detection [10]. But the

detection may not be satisfactory under low-SNR conditions and the method may not correctly track the noise spectrum during long speech segments. Several statistical techniques for estimating the noise spectrum, without voice activity detection, have been reported [10], [20]–[27]. Their computational complexity and memory requirement pose difficulty in real-time processing using a low-power processor.

1.2 Project objective

The project objective is to implement a system for real-time enhancement of noisy speech for use in hearing-aids and other sensory aids for the hearing impaired. Towards this end, various noise estimation techniques for use in spectral subtraction for speech enhancement are investigated. A spectral subtraction technique for speech enhancement using cascaded-median based continuous updating of the noise spectrum, without using voice activity detection, is presented. It is implemented for real-time operation on a 16-bit fixed-point DSP processor, with on-chip FFT hardware.

1.3 Dissertation outline

Chapter 2 describes the generalized spectral subtraction along with different noise estimation techniques. Investigations on enhancement of noisy speech with different types of noises, using Matlab based offline processing and results are presented in the next chapter. Chapter 4 presents a DSP processor-based real-time implementation of speech enhancement and results are discussed. The last chapter gives a summary and conclusions of the work.

Chapter 2

SPEECH ENHANCEMENT USING SPECTRAL SUBTRACTION

2.1 Generalized spectral subtraction

Spectral subtraction is a single-input noise reduction method based on the short-time estimation of the magnitude spectrum of the noise. Processing involves estimating the magnitude spectrum of the noise, using it for 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. A block diagram of speech enhancement using spectral subtraction is shown in Fig. 2.1. Windowed frames of the noisy speech signal $x(n)$, which is a sum of noise-free speech and noise, are given to a FFT block to find magnitude and phase spectra. The magnitude spectra of the past frames are used to estimate the noise magnitude spectrum $D_n(k)$. The noise is estimated during non-speech segments using a voice detector or it is dynamically estimated using statistical methods. The enhanced magnitude spectrum $|Y_n(k)|$ is computed using spectral subtraction. IFFT is taken for the complex spectrum formed by enhanced magnitude spectrum and noisy phase spectrum. Enhanced signal is reconstructed using overlap-add method. Several investigations have been reported, providing different methods for each of these steps [10]–[27]. The effectiveness of the noise removal process is dependent on obtaining an accurate spectral estimate of the noise from the noisy speech signal. Significant differences between the estimated noise and the actual noise present in the short-time speech spectrum may result in the presence of isolated residual spectral peaks of large variance. These residual spectral contents manifest themselves in the reconstructed signal as varying tonal sounds known as “musical noise” and may result in an unnatural quality.

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) or full-wave rectification (set to its absolute value). This can lead to

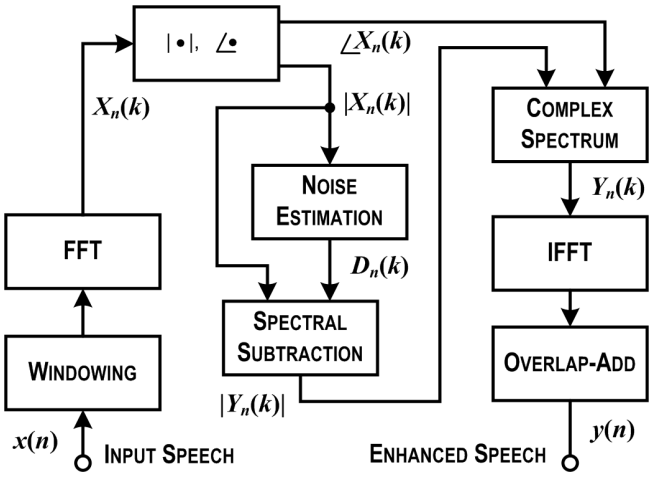


Fig. 2.1 Speech enhancement by spectral subtraction [29]

further distortions in the resulting time signal. To overcome the shortcomings of spectral subtraction, Berouti *et al.* [11] developed a generalized spectral subtraction. The enhanced magnitude spectrum $|Y_n(k)|$ computed using generalized spectral subtraction may given as the following

$$|Y_n(k)| = \begin{cases} [|X_n(k)|^\gamma - \alpha(D_n(k))^\gamma]^{1/\gamma} & \text{if } |X_n(k)| > (\alpha+\beta)^{1/\gamma} D_n(k) \\ \beta^{1/\gamma} D_n(k), & \text{otherwise} \end{cases} \tag{2.1}$$

Here γ is an exponent factor, resulting in power subtraction for $\gamma = 2$ and magnitude subtraction for $\gamma = 1$. 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 musical noise is masked by a floor noise controlled by the spectral floor factor β. These two factors offer a great flexibility in the algorithm. Several methods, with different computational complexity, using frequency-dependent factors and factors as functions of *a posteriori* estimate of SNR have been reported [10].

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

$$Y_n(k) = |Y_n(k)| e^{j\angle X_n(k)} \tag{2.2}$$

In order to avoid phase calculation, the complex spectrum is calculated using

$$Y_n(k) = |Y_n(k)| X_n(k) / |X_n(k)| \tag{2.3}$$

The resulting complex spectra are used to re-synthesize the speech signal. As spectral subtraction involves modification of short-time Fourier transform, there may be discontinuities between signal segments corresponding to the modified complex spectra of the consecutive frames. Use of overlap-add in the re-synthesis helps in masking and in reducing the perceived distortions related to their discontinuities.

In the generalized spectral subtraction [11], it is assumed that the noise affects the entire spectrum uniformly, which is generally not valid in the case of real-world noise. A multi-band spectral subtraction has been proposed by Kamath and Loizou [12]. In this method, different subtraction factors are used in different frequency bands on the basis of SNR estimated in the bands. Further, the generalized spectral subtraction is based on the assumption that speech and additive noises are uncorrelated and hence cross terms are made zero. However, this assumptions is not valid when the speech is processed on frame-to-frame basis [10]. Lu and Loizou have proposed a geometric approach [13] for spectral subtraction without setting the cross terms as zeros.

2.1.1 Multi-band spectral subtraction

Spectral subtraction proposed by Berouti *et al.* is based on the assumption that the entire spectrum is uniformly affected by noise. Setting $\gamma = 2$ in (2.1), results in power spectral subtraction. Here the subtraction factor α is constant for the entire spectrum. But the real-world noises (e.g., car noise, cafeteria noise) do not uniformly affect different frequency regions [10]. Kamath and Loizou [12] proposed a multi-band approach to spectral subtraction. Here the speech spectrum is divided into *B* non-overlapping bands, and spectral subtraction is performed independently in each band. Power spectral subtraction using multi-band approach for *i*th band is given as

$$\begin{aligned} |Y_n(k)| &= \beta^{1/2} |X_n(k)|, \quad \text{if } |X_n(k)| < (\alpha_i \delta_i)^{1/2} |D_n(k)| \\ &[|X_n(k)|^2 - \alpha_i \delta_i |D_n(k)|^2]^{1/2} \quad \text{otherwise} \end{aligned} \tag{2.4}$$

where α_i is subtraction factor for band i and depends on the SNR of the corresponding band and δ_i is a tweaking factor. SNR_i in dB is estimated as

$$\mathrm{SNR}_i(\mathrm{dB}) = 10\log_{10}\left[\left(\sum_{b_{\mathrm{si}}}^{b_{\mathrm{ei}}}|X_n(k)|^2\right)/\left(\sum_{b_{\mathrm{si}}}^{b_{\mathrm{ei}}}|D_n(k)|^2\right)\right] \tag{2.5}$$

where b_{si} and b_{ei} are the beginning and ending frequency samples of the ith band. Subtraction factor α_i depends on the SNR (in dB) of the particular band and is given as

$$\alpha_i = \begin{cases} 5 & \mathrm{SNR}_i < -5 \\ 4-(3/20)\mathrm{SNR}_i & -5 \le \mathrm{SNR}_i \le 20 \\ 1 & \mathrm{SNR}_i > 20 \end{cases} \tag{2.6}$$

Tweaking factor δ_i is empirically set for each frequency band to customize the noise removal properties as

$$\delta_i = \begin{cases} 1 & f_i \le 1\,\mathrm{kHz} \\ 2.5 & 1\,\mathrm{kHz} < f_i \le (F_s/2) - 2\,\mathrm{kHz} \\ 1.5 & f_{\mathrm{i}} > (F_s/2) - 2\,\mathrm{kHz} \end{cases} \tag{2.7}$$

where f_i is the upper frequency of the ith band and F_s is the sampling frequency. They evaluated the method using ten sentences from HINT (Hearing In Noise Test) database as the speech material and added speech-shaped noise at 0 dB and 5 dB SNR. The Itakura-Saito (IS) distance was used as an objective measure to evaluate the performance. The method showed an improvement in the objective measure over the conventional power spectral subtraction and the informal listening showed that the processed output had a very little trace of musical noise.

2.1.2 Geometric approach to spectral subtraction

In the spectral subtraction method proposed by Boll [14] and Berouti *et al.*[11], the cross term is taken as zero, assuming the speech and noise to be uncorrelated. This assumption is generally not valid for processing using short-time windowed speech [10]. Lu and Loizou [13] proposed a geometric approach to spectral subtraction without assuming the cross terms to be zero. In this method, the enhanced spectrum of a frame n is estimated as

$$|Y_n(k)| = H_{\mathrm{GA}}(n,k)|X_n(k)| \tag{2.8}$$

where $H_{\mathrm{GA}}(n, k)$ is known as suppression function. Its value is limited to 1 and is calculated as

$$H_{GA}(n,k)=\sqrt{\left(1-\frac{(\hat{\mu}_{n,k}+1-\hat{\xi}_{n,k})^2}{4\hat{\mu}_{n,k}}\right)\left(1-\frac{(\hat{\mu}_{n,k}-1-\hat{\xi}_{n,k})^2}{4\hat{\xi}_{n,k}}\right)^{-1}} \quad (2.9)$$

where $\mu_{n,k}$ and $\xi_{n,k}$ are the posteriori and priori SNRs, calculated as the following

$$\mu_{n,k}=|X_n(k)|^2/|D_n(k)|^2 \quad (2.10a)$$

$$\xi_{n,k}=|\mathrm{Y}_n(\mathrm{k})|^2/|D_n(k)|^2 \quad (2.10b)$$

The smooth estimates of μ and ξ for frame n for gain function are obtained

$$\hat{\mu}_{n,k}=\lambda.\hat{\mu}_{n-1,k}+(1-\lambda).\min\left[\left(|X_n(k)|^2/|D_n(k)|^2\right),20\right] \quad (2.11)$$

$$\hat{\xi}_{n,k}=\max\left[0.05,\sigma.\left(|\mathrm{Y}_{n\text{-}1}(\mathrm{k})|^2/|D_{n-1}(k)|^2\right)+(1-\sigma).\left(\sqrt{\hat{\mu}_{n,k}}-1\right)^2\right] \quad (2.12)$$

where σ and λ are smoothing constants. They evaluated the method using speech from NOIZEUS database (30 sentences spoken by 3 male and 3 female speakers) and added multi-talker babble, street noise, and car noise taken from the AURORA database and white noise at 0, 5 and 10 dB SNR levels. Mean square error (MSE) of true and estimated magnitude spectra was calculated and compared with the traditional spectral subtraction. The proposed method resulted in much less MSE than the traditional spectral subtraction. The results showed that cross terms can be ignored at very low and high SNRs but not near to 0 dB. Objective evaluation was carried using PESQ and log likelihood ratio (LLR) to compare GA with traditional spectral subtraction and MMSE algorithms. The PESQ score of GA was significantly higher than that of the spectral subtraction in all the cases. MMSE algorithm has higher PESQ scores than the GA except in babble at 0 dB and 5 dB SNR. Use of LLR measure also showed the same pattern. GA algorithm has advantage over MMSE in terms of computational requirement [13]. On the basis of informal listening, the authors have reported that the processed output of GA had no audible musical noise and the residual noise was smooth and pleasant.

2.2 Noise estimation

Because of the non-stationary nature of most of the interfering noise, the noise spectrum needs to be dynamically estimated. An under-estimation results in residual noise while an

over-estimation results in distortion leading to degraded quality and possibly a loss in intelligibility. Noise estimation is carried out as a moving average over several overlapping windows during the silence intervals, identified by a speech/non-speech classifier or a voice activity detector [10]. Estimated noise is assumed to remain stationary during speech segments. This method may not work satisfactorily under conditions of low SNR and it may not track the variation in noise spectrum during the speech segments. Hence, it is desirable to have a method that does not depend on voice activity detection. Several statistical techniques for dynamically estimating the noise spectrum without involving voice activity detection have been reported, e.g. minimal-tracking algorithms, time-recursive averaging algorithms, histogram based algorithms, and quantile-based algorithms [10], [20]–[27].

2.2.1 Minimal-tracking algorithms

Minimal-tracking algorithms are based on the assumption that in an individual frequency band the power of noisy speech signal decays to the power level of the noise, even during the speech activity. Hence tracking the minimum of the noisy speech power in a frequency band can roughly estimate the noise level in that band. Minimum statistics (MS) algorithm [20] tracks the noise as minima of the past frames. Drawback of this algorithm is that it cannot respond to fast changes of the noise spectrum. Algorithm is suitable for real-time operation, but it often under-estimates the noise and requires a high subtraction factor. As a constant subtraction factor may result in removal of some speech parts in weaker segments, a SNR-dependent subtraction factor is required. In minimum tracking algorithm [21], the noise is updated continuously by smoothing the noisy speech power spectra in each frequency bin using a nonlinear smoothing. In [21], the noise estimation was combined with spectral amplitude estimator [22] and informal listening showed that the algorithm performed well compared to conventional spectral subtraction. The combined technique was implemented on a floating-point DSP processor (Analog Devices, ADSP-21020) for real-time processing. The processor utilization was 14% – 23%.

2.2.2 Time-recursive averaging algorithms

In time-recursive averaging algorithms, the noise spectrum is estimated as a weighted average of past noise estimates and the present noisy speech spectrum. The weights are updated adaptively based on either on the effective SNR of each frequency bin or on the speech-presence probability. The algorithm reported includes SNR-dependent recursive averaging [23], weighted spectral averaging [24] and minima-controlled recursive averaging [25].

In the method reported by Lin *et al.* [23], the noisy speech is decomposed in sub-band signals based on auditory critical bandwidths [16]. The noisy signal power in each subband is smoothened, and the noise is estimated adaptively. The smoothing parameter is a function of the estimated signal-to-noise ratio (SNR). The performance was tested using a sentence with additive noise from the Noisex92 database at 2.5 dB SNR. Comparison of estimated *a posteriori* SNR with ideal *a posteriori* SNR resulted in a frame-by-frame average estimated error of 4.85%. Testing with pink noise, F16 noise, and car noise gave satisfactory results for the SNR's from -5 to 15 dB. The algorithm needs additional computations for calculating SNR in each sub-band.

Hirsch and Ehrlicher [24] reported a noise estimation algorithm based on 400 ms past noisy speech segments. In this method, the noise level in each sub-band is estimated as a first order recursive weighted average of past spectral magnitude values which are below an adaptive threshold. The method has a low computation complexity.

Cohen [25] proposed noise estimation using an improved minima-controlled recursive averaging (IMCRA) using a smoothing parameter. In this method, the smoothing parameter is frequency-dependent and is dynamically adjusted by the signal presence probability. The speech presence probability is controlled by the minima values of smoothed periodogram. Algorithm comprises two iterations of smoothing and minimum tracking. In first iteration, rough voice activity detection is provided in each frequency band. In the second iteration, smoothing excludes relatively strong speech components, which makes the minimum tracking robust during speech activity. Performance was evaluated for white Gaussian noise (WGN), car noise, and F16 cockpit noise from Noisex92 database and speech signal obtained by concatenating six sentences (three male, three female) from TIMIT database. The noises

were added with SNR's from -5 dB to 15 dB. The proposed method was found to track the actual noise better than the minimum statistics method. The segmental relative estimator for various types of noises at different SNR's was calculated and the new method had significantly lower estimator error than the minimum statistics. The noise estimators were combined with optimally-modified log-spectral amplitude estimator and evaluated objectively using the improvement in segmental SNR measure and subjectively using informal listening. The new method resulted in higher segmental SNR improvement than the minimum statistics consistently in all environmental conditions.

2.2.3 Histogram-based techniques

In histogram-based techniques [24], [27] noise is estimated based on the histogram of the power spectra of the past frames. For each incoming frame, a histogram for past frames is updated, and the value corresponding to the maximum in the histogram at each frequency bin is considered as an estimate of noise spectrum. Appropriate bin width for histogram at each frequency needs to be used. Too narrow a bin width results in a high variability while too wide a bin width leads to very coarse estimation.

Hirsch and Ehrlicher [24] used a noise estimation algorithm based on histogram obtained for noisy speech over past 400 ms duration. In this method, the noise level is estimated as maximum of the distribution in each sub-band. To avoid spikes, the estimated values are smoothed along time index. The performance of the algorithm was evaluated and compared with the weighted spectral average method [24] using an objective evaluation based on relative error, calculated as the ratio of mean square error between average of added noise and average of estimated noise to the mean square of average of added noise. Relative error was low for histogram method than that for the weighted spectral average method. Both the techniques were combined with non-linear spectral subtraction method and informal listening showed noise to have been well suppressed. Recognition experiment was carried for evaluating both the techniques using the isolated words of Noisex92 and ten digits spoken 100 times separately for training and testing. The noise was added at different SNRs. With an HMM based word recognizer the enhanced signals using weighted average noise estimate

had slightly higher recognition rate than the enhanced signal using histogram based noise estimate.

2.2.4 Quantile-based noise estimation

Quantile-based noise estimation (QBNE) [26] is based on the observation that the speech signal energy in a particular band is very low in most of the frames and high in only 10-20% of the frames containing voiced speech segments. Therefore it may be possible to estimate the noise spectrum by selecting a certain quantile value from the previous frames of the noisy speech spectrum. Several frequency and SNR-dependent methods for quantile selection have been used, but a median based noise estimation (MBNE) has been reported to work well and in a robust manner [26]. The QBNE was compared with the noise estimation based on recursive averaging in pause detection. The material used in the experiments consisted of 6034 utterances of German digits and digit strings by 770 speakers in 10 different cars. QBNE gave significantly higher pause detection scores than the other methods. The method is unsuitable for real-time operation, because sorting of the past frames is computation intensive and also has a large memory requirement.

2.2.5 Cascaded-median based noise estimation

A cascaded-median [28], [29] can be used as an approximation to median, with a significantly reduced computation and memory requirement. In a p-frame q-stage cascaded-median, as shown in Fig. 2.2, each stage has a first-in-first-out buffer holding p magnitude spectra. The first stage receives the input-frame spectrum. After every p inputs, an ensemble median is calculated and given as input to the next stage. The same process is followed in all the stages and the output of the last stage is taken as an approximation of the ensemble median of the spectra over p^q past frames. Let us compare the number of sorting operations and storage per frequency bin, assuming that the noise spectrum is estimated every M frames from the previous M frames. True-median requires M-sample array for buffering and M-sample array for sorting. For arranging the samples in ascending order, it requires a total of $M(M–1)/2$ sorting operations, i.e. $(M\text{-}1)/2$ operations per frame. With $M = p^q$, the cascaded-median requires q p-sample arrays. It results in a storage saving ratio of $2M/(pq)$, and $q \approx$

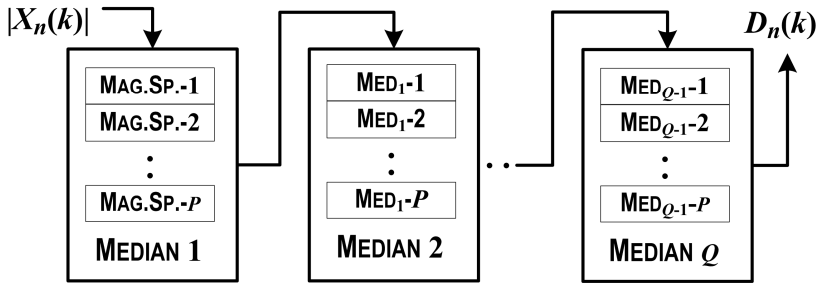


Fig. 2.2 A *p*-point *q*-stage cascaded-median based noise estimation [29]

ln(*M*) gives the highest saving. For uniformity in the number of computational operations across frames, median is calculated in only one stage at each frame position, giving priority to the higher stage. In this method, some frames do not contribute to the median calculation, but this fact does not significantly affect the noise estimation. In this case, $p(p\text{-}1)/2$ sorting operations are needed per frame. Thus the saving ratio for sorting operation per frame is $(M\text{-}1)/p(p\text{-}1)$. A lower p results in lesser computation and $p = 3$ simplifies the programming for sorting operations.

2.3 Comparison of enhancement techniques

A comparison of 19 speech enhancement techniques was carried out using implementations available on CD accompanying [10]. The techniques belong to spectral-subtractive, statistical-model, and subspace based algorithms. In spectral-subtractive algorithm the estimate of clean signal is obtained by subtracting an estimate of noise spectrum from the noisy speech spectrum. In statistical-model based algorithms, a non-linear estimator of the parameter of interest needs to be found using the given set of measurements. For noise suppression, the measurements are the noisy speech spectra and the parameters of interest are the estimates of clean speech spectra. Subspace algorithms are based on the assumption that the clean signal is confined to a subspace of the noisy Euclidean space. For noise suppression, the noisy speech signal is decomposed into subspaces which are primarily occupied by clean signal and noise. The clean signal is re-synthesized after the noise subspace in noisy vectors is nullified.

The evaluation involved using informal listening and an objective evaluation using perceptual evaluation of speech quality (PESQ) measure [10], [30]. Investigations were carried out on speech materials taken from the NOIZEUS database [31], consisting of 30 IEEE sentences recorded from 3 male and 3 female speakers with 25 kHz sampling and down-sampled to 8 kHz. For testing, six sentences from the database were concatenated and up sampled to 10 kHz. The concatenated material is "The birch canoe slid on the smooth planks. He knew the skill of the great young actress. Her purse was full of useless trash. Read verse out loud for pleasure. Wipe the grease off his dirty face. Men strive but seldom get rich". Informal listening showed that quality degradation is generally more noticeable during vowel segments while intelligibility degradation is more noticeable during consonantal segments. For a quick comparison of effects of different processing steps and parameters, we have used speech material recorded in our lab. The material consisted of three isolated vowels, a Hindi sentence, and an English sentence (-/a/-/i/-/u/– "aayiye aap kaa naam kyaa hai?" – "Where were you a year ago?") from a male speaker. It was recorded with sampling frequency of 11.025 kHz and converted to 10 kHz. A longer test sequence was generated by speech-speech-silence-speech concatenation of the recording. This material is referred to "vowel, Hindi sentence, English sentence" or "VHSES". The test materials NOIZEUS and VHSES are of 14 s and 25 s duration respectively. While NOIZEUS is rich in consonants, VHSES is dominated by vowels and vowel-like segments. Speech was mixed with different types of noises at different SNR values and processed by the speech enhancement techniques.

PESQ scores were obtained for the processed output using unprocessed clean speech as reference. The scores for SNR of 0 dB are given in Table 2.1 and Table 2.2 for VHSES and NOIZEUS speech materials. For NOIZEUS, the unprocessed speech has scores of 1.55 – 2.05. In terms of decreasing the scores for unprocessed speech, the noises are ranked as train (2.05), street (1.83), babble (1.73), car (1.72), pink (1.60), and white (1.55) for NOIZEUS. The scores for noisy VHSES followed the same ranking, and were slightly smaller, by 0.01 – 0.05. In comparison with other noises, processing of speech corrupted by white noise showed maximum improvement. The observation is valid for most of the processing methods. As expected, the improvements are lowest for babble noise. Effectiveness of the techniques in improving the scores is different across noises and also for the two speech materials. The

Table 2.1 PESQ scores for enhanced speech using various algorithms. Speech material: VHSES, SNR: 0 dB. Improvements over the scores for unprocessed noisy speech are given in brackets.

Enhancement method	Noise type					
	White	Babble	Street	Pink	Train	Car
un proc.	1.54	1.73	1.78	1.59	2.00	1.67
specsub	1.78 (0.24)	1.74 (0.01)	1.85 (0.07)	2.01 (0.42)	2.33 (0.33)	1.91 (0.24)
mband	1.43 (-0.11)	1.88 (0.15)	2.06 (0.28)	1.72 (0.13)	2.61 (0.61)	2.09 (0.42)
ga	1.82 (0.28)	2.05 (0.32)	2.42 (0.64)	2.11 (0.52)	2.67 (0.67)	2.26 (0.59)
wiener_iter	1.96 (0.42)	1.71 (-0.02)	1.56 (-0.22)	1.80 (0.21)	1.78 (-0.22)	1.81 (0.14)
wiener_as	1.81 (0.27)	1.82 (0.09)	1.93 (0.15)	2.02 (0.43)	2.57 (0.57)	2.07 (0.40)
wiener_wt	1.73 (0.19)	1.69 (-0.04)	1.73 (-0.05)	1.84 (0.25)	2.18 (0.18)	1.67 (0)
mt_mask	1.36 (0.18)	1.39 (-0.34)	1.87 (0.09)	1.54 (-0.05)	2.11 (0.11)	1.52 (-0.15)
audnoise	1.88 (0.34)	1.64 (-0.09)	1.74 (-0.04)	1.74 (0.15)	2.32 (0.32)	1.69 (0.02)
mmse	1.95 (0.41)	1.86 (0.13)	1.99 (0.21)	2.22 (0.63)	2.69 (0.69)	2.24 (0.57)
logmmse	1.98 (0.44)	1.85 (0.12)	2.02 (0.24)	2.24 (0.65)	2.72 (0.72)	2.30 (0.63)
logmmse_spu1	1.94 (0.40)	1.92 (0.19)	1.97 (0.19)	2.16 (0.57)	2.58 (0.58)	2.15 (0.48)
logmmse_spu2	1.90 (0.36)	1.90 (0.17)	1.95 (0.17)	2.14 (0.55)	2.55 (0.55)	2.13 (0.46)
logmmse_spu3	2.11 (0.57)	1.93 (0.20)	1.94 (0.16)	2.25 (0.66)	2.62 (0.62)	2.16 (0.49)
logmmse_spu4	1.71 (0.17)	1.80 (0.07)	1.81 (0.03)	1.87 (0.28)	2.46 (0.46)	1.88 (0.21)
stsa_weuclid	2.01 (0.47)	1.85 (0.12)	1.99 (0.21)	2.26 (0.67)	2.67 (0.67)	2.24 (0.57)
stsa_wcosh	2.11 (0.57)	1.90 (0.17)	2.20 (0.42)	2.31 (0.72)	2.65 (0.65)	2.14 (0.47)
stsa_mis	1.75 (0.21)	1.81 (0.08)	1.89 (0.11)	1.96 (0.37)	2.23 (0.23)	1.86 (0.19)
klt	2.19 (0.65)	1.84 (0.11)	2.00 (0.22)	2.38 (0.79)	2.60 (0.60)	2.07 (0.40)
pklt	2.51 (0.97)	1.89 (0.16)	1.84 (0.06)	2.27 (0.68)	2.32 (0.32)	1.91 (0.24)

improvements are generally highest with stsa_wcosh (Bayesian measure based on cosh distortion measure) [19], logmmse (log MMSE algorithm) [22], mmse (MMSE algorithm with speech presence uncertainty) [17], ga (geometric approach to spectral subtraction) [13], stsa_weuclid (Bayesian measure based on Euclidean distortion measure) [19], and log MMSE algorithm incorporating speech presence uncertainty (logmmse_spu3) [18].

Figures 2.3 – 2.14 show the PESQ score versus SNR plot of unprocessed and processed signals for different noises and processing with the best six techniques. SNR

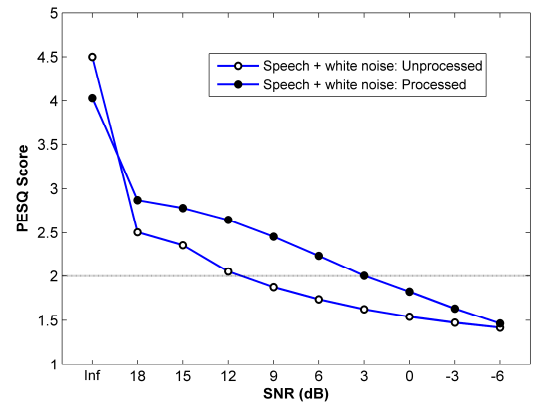
advantage was calculated using PESQ score vs. SNR plots for all the enhancement methods at a PESQ score of 2.0, which is generally considered as lowest score for acceptable speech. Table 2.3 shows the SNR advantages of different enhancement techniques for different noises. The results can be summarized in terms of ranges of SNR advantage as the following

stsa_wcosh:	4 – 13 dB for VHSES, 2 – 9 dB for NOIZEUS
logmmse:	3 – 11 dB for VHSES, 2 – 7.5 dB for NOIZEUS
mmse:	3 – 11 dB for VHSES, 3 – 6.5 dB for NOIZEUS
ga:	6 – 10.5 dB for VHSES, 1 – 5.5 dB for NOIZEUS
stsa_weuclid:	3 – 11 dB for VHSES, 1.5 – 7 dB for NOIZEUS
logmmse_spu3:	2.5 – 13.5 dB for VHSES, 0.5 – 7 dB for NOIZEUS

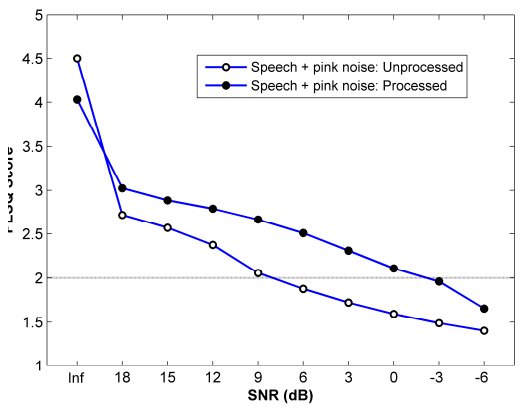
In all the cases, the improvements were generally highest for white noise and lowest for babble and street noise.

Table 2.2 PESQ scores for enhanced speech using various algorithms. Speech material: NOIZEUS, SNR: 0 dB. Improvements over the scores for unprocessed noisy speech are given in brackets.

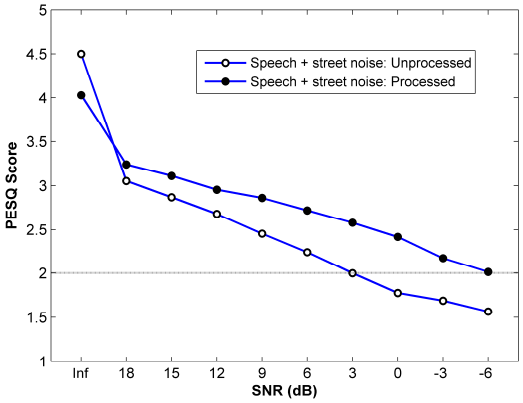
Enhancement method	Noise type					
	White	Babble	Street	Pink	Train	Car
un proc.	1.55	1.75	1.83	1.60	2.05	1.72
specsub	1.63 (0.08)	1.58 (-0.17)	1.81 (-0.02)	1.78 (0.18)	2.04 (-0.01)	1.77 (0.05)
mband	1.59 (0.04)	1.84 (0.09)	2.02 (0.19)	1.73 (0.13)	2.30 (0.25)	1.91 (0.19)
ga	1.61 (0.06)	1.82 (0.07)	2.07 (0.24)	1.86 (0.26)	2.35 (0.3)	1.96 (0.24)
wiener_iter	1.56 (0.01)	1.29 (-0.46)	1.17 (-0.66)	1.31 (-0.29)	1.31 (-0.74)	1.36 (-0.36)
wiener_as	1.77 (0.22)	1.80 (0.05)	1.98 (0.15)	1.90 (0.30)	2.34 (0.29)	1.92 (0.20)
wiener_wt	1.61 (0.06)	1.41 (-0.34)	1.68 (-0.15)	1.61 (0.01)	1.93 (-0.12)	1.36 (-0.36)
mt_mask	1.16 (-0.39)	1.12 (-0.63)	1.65 (-0.18)	1.28 (-0.32)	2.05 (0)	1.43 (-0.29)
audnoise	1.58 (0.03)	1.19 (-0.56)	1.44 (-0.39)	1.46 (-0.14)	1.86 (-0.19)	1.28 (-0.44)
mmse	1.82 (0.27)	1.78 (0.03)	2.02 (0.19)	2.01 (0.41)	2.43 (0.38)	1.99 (0.27)
logmmse	1.80 (0.25)	1.81 (0.06)	2.04 (0.21)	2.00 (0.40)	2.46 (0.41)	2.02 (0.3)
logmmse_spu1	1.68 (0.13)	1.58 (-0.17)	1.86 (0.03)	1.84 (0.24)	2.24 (0.19)	1.79 (0.07)
logmmse_spu2	1.69 (0.14)	1.53 (-0.22)	1.85 (0.02)	1.85 (0.25)	2.19 (0.14)	1.79 (0.07)
logmmse_spu3	1.80 (0.25)	1.66 (-0.09)	1.91 (0.08)	1.99 (0.39)	2.35 (0.30)	1.92 (0.20)
logmmse_spu4	1.34 (-0.21)	1.44 (-0.31)	1.62 (-0.21)	1.53 (-0.07)	1.93 (-0.12)	1.51 (-0.21)
stsa_weuclid	1.78 (0.23)	1.81 (0.06)	1.99 (0.16)	1.96 (0.36)	2.40 (0.35)	1.98 (0.26)
stsa_wcosh	1.95 (0.34)	1.81 (0.06)	2.04 (0.21)	2.07 (0.47)	2.41 (0.36)	1.91 (0.19)
stsa_mis	1.78 (0.23)	1.78 (0.03)	1.81 (-0.02)	1.90 (0.30)	2.15 (0.10)	1.82 (0.10)
klt	1.89 (0.34)	1.71 (-0.04)	1.87 (0.04)	1.97 (0.37)	2.07 (0.02)	1.78 (0.06)
pklt	1.78 (0.23)	1.50 (-0.25)	1.62 (-0.21)	1.70 (0.10)	1.71 (-0.34)	1.44 (-0.28)



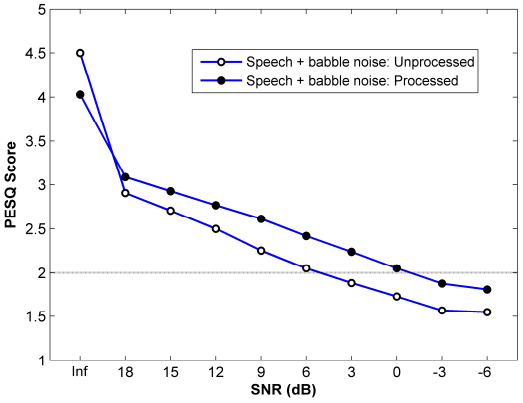
(a) Noise: white



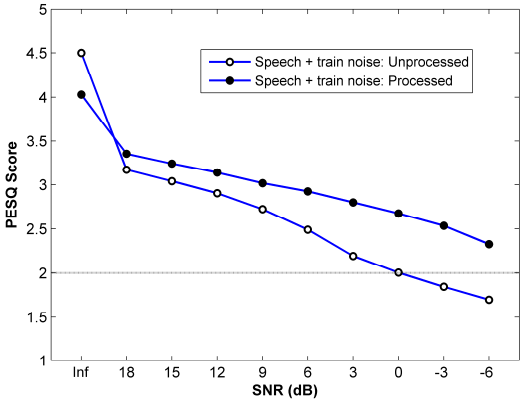
(b) Noise: pink



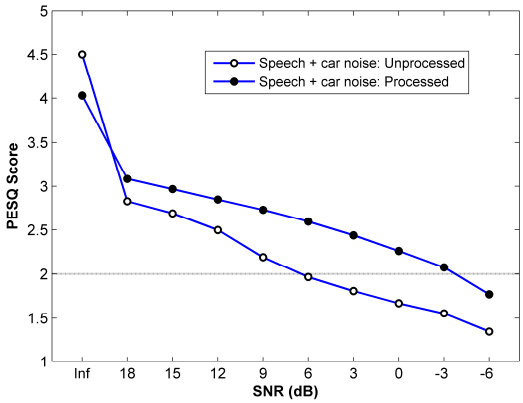
(c) Noise: street



(d) Noise: babble

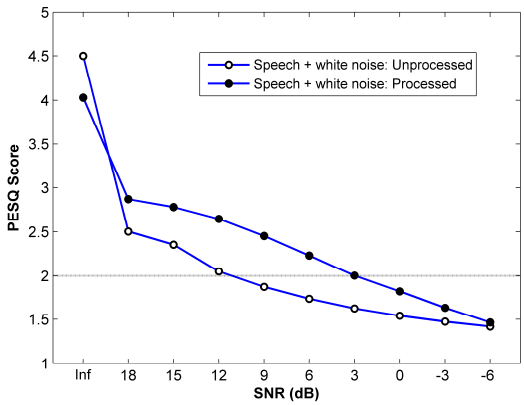


(e) Noise: train

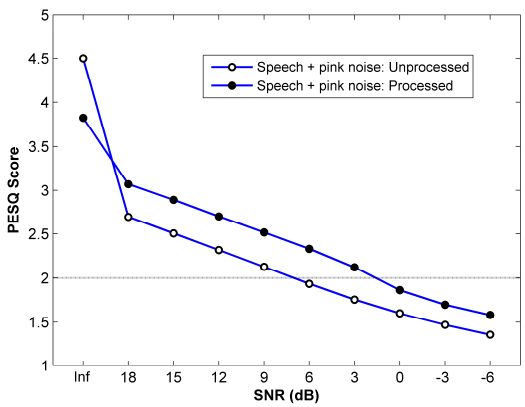


(f) Noise: car

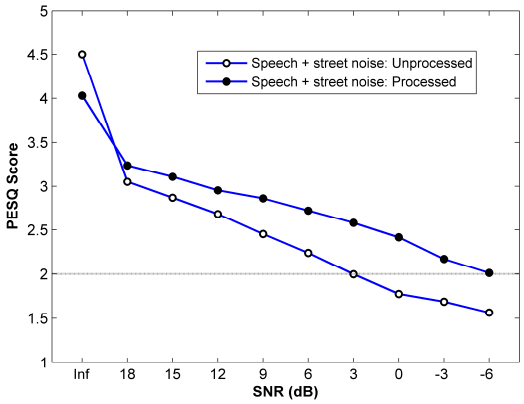
Fig. 2.3 PESQ score vs SNR for noisy and enhanced speech using geometric approach to spectral subtraction, speech material: VHSES.



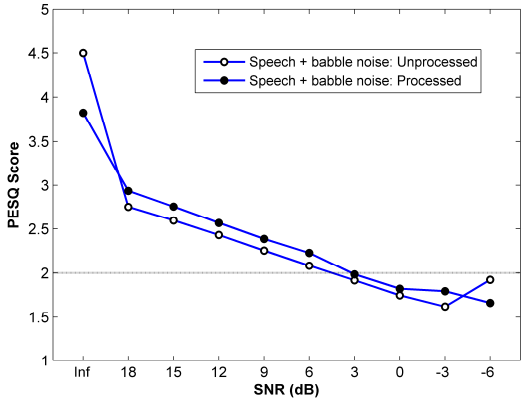
(a) Noise: white



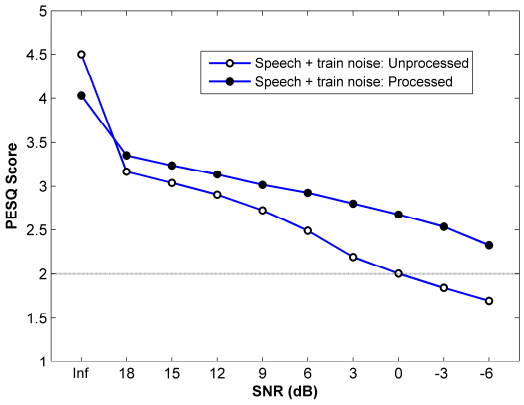
(b) Noise: pink



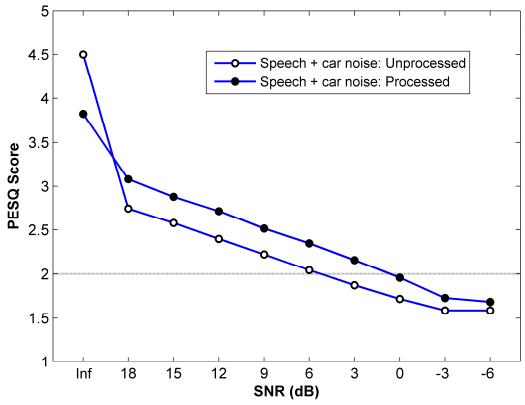
(c) Noise: street



(d) Noise: babble

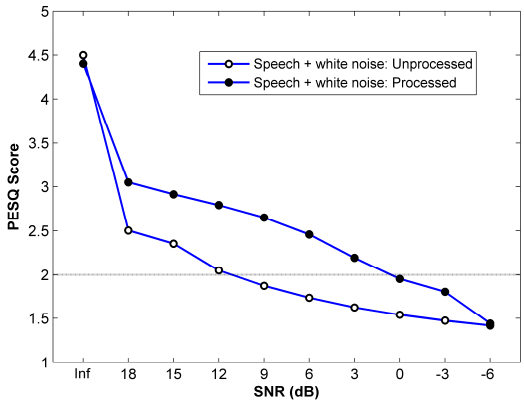


(e) Noise: train

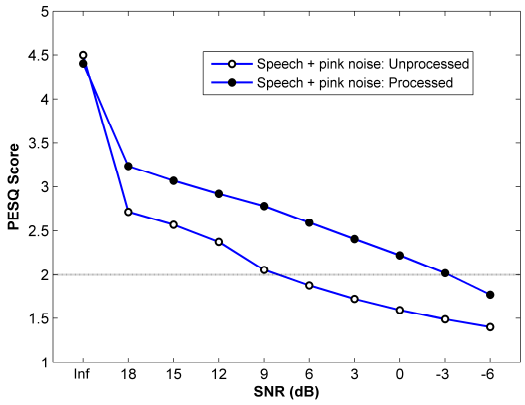


(f) Noise: car

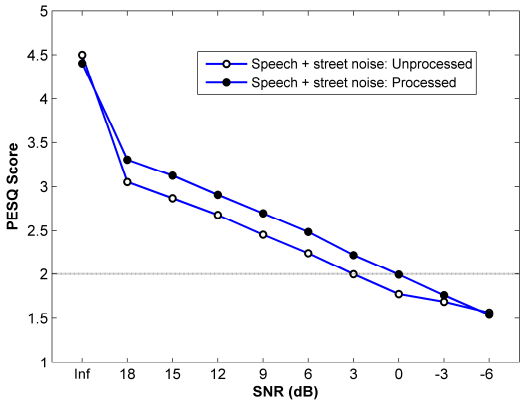
Fig. 2.4 PESQ score vs SNR for noisy and enhanced speech using geometric approach to spectral subtraction, speech material: NOIZEUS.



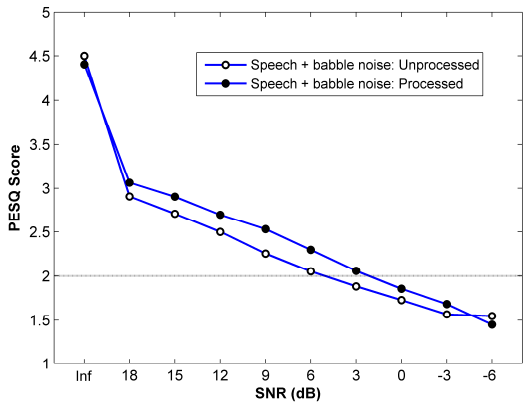
(a) Noise: white



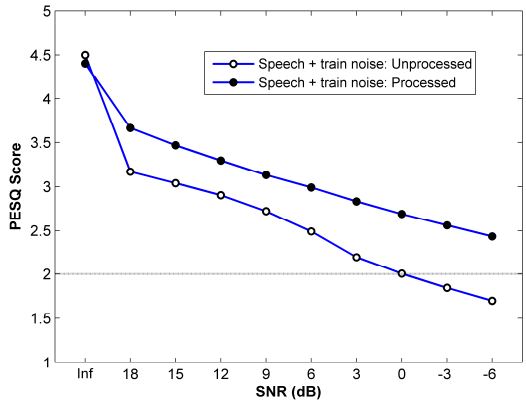
(b) Noise: pink



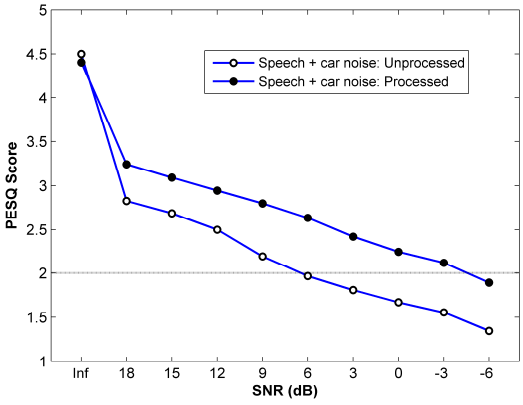
(c) Noise: street



(d) Noise: babble

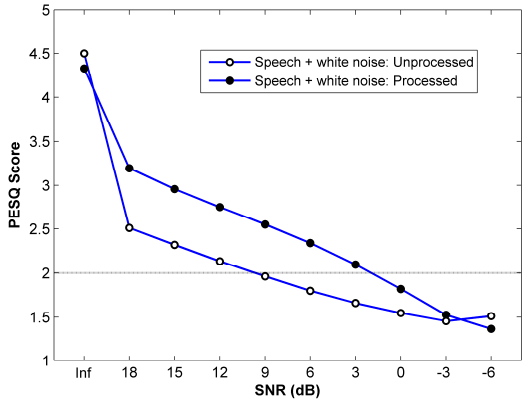


(e) Noise: train

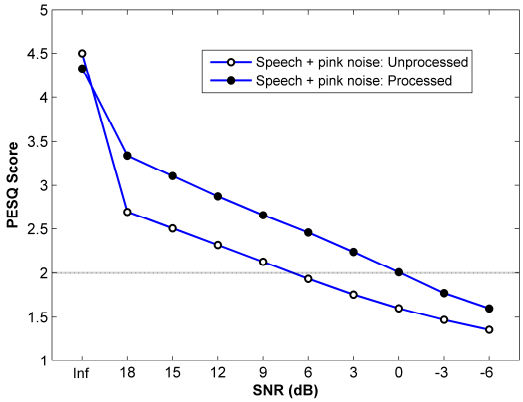


(f) Noise: car

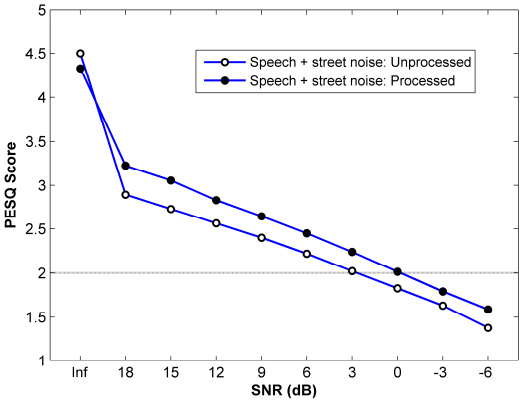
Fig. 2.5 PESQ score vs SNR for noisy and enhanced speech using MMSE algorithm with speech presence uncertainty, speech material: VHSES.



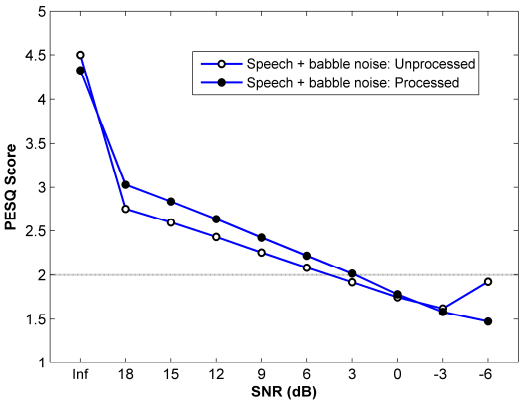
(a) Noise: white



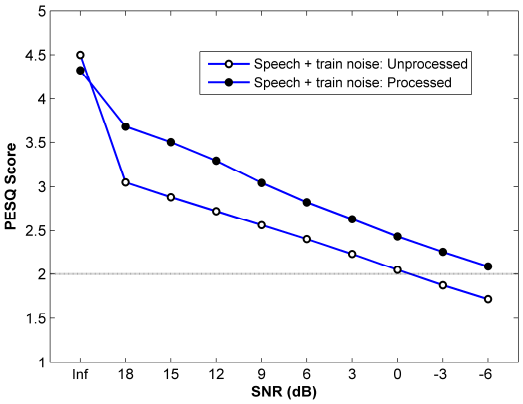
(b) Noise: pink



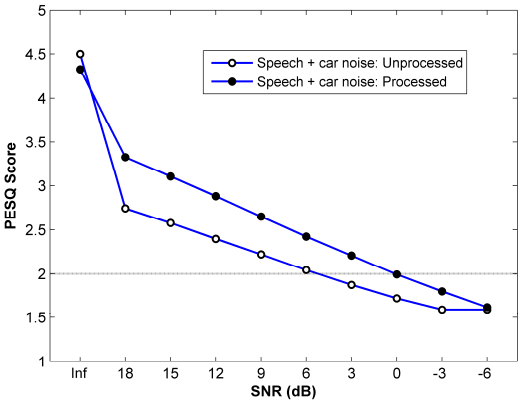
(c) Noise: street



(d) Noise: babble

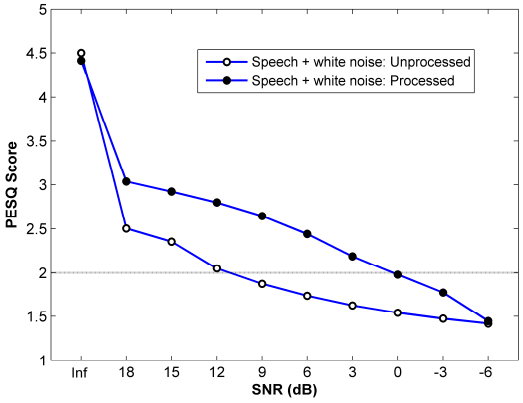


(e) Noise: train

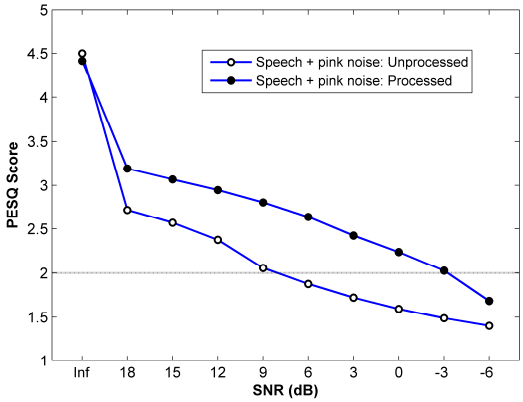


(f) Noise: car

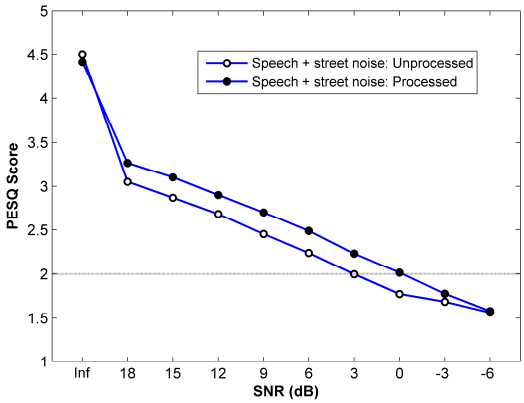
Fig. 2.6 PESQ score vs SNR for noisy and enhanced speech using MMSE algorithm with speech presence uncertainty, speech material: NOIZEUS.



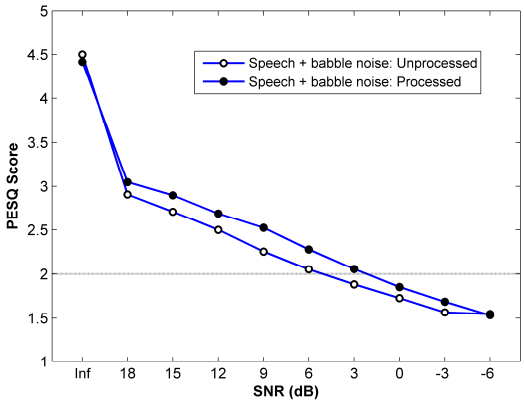
(a) Noise: white



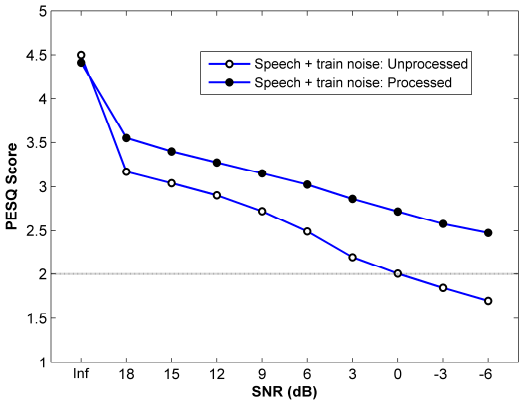
(b) Noise: pink



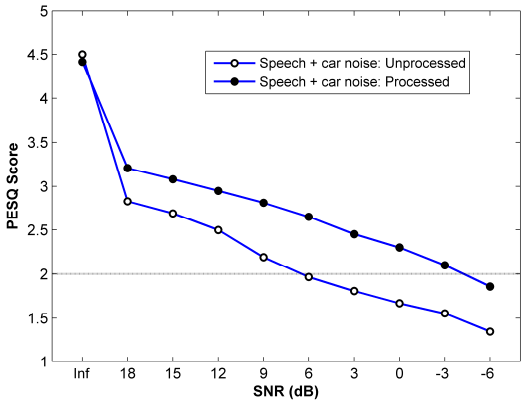
(c) Noise: street



(d) Noise: babble

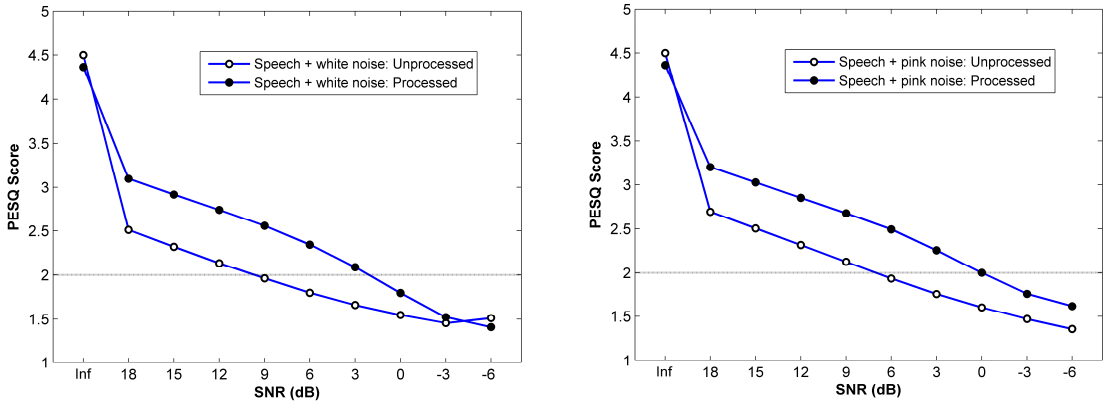


(e) Noise: train



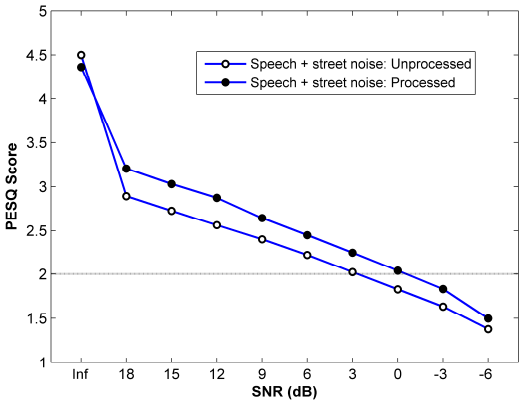
(f) Noise: car

Fig. 2.7 PESQ score vs SNR for noisy and enhanced speech using log MMSE algorithm, speech material: VHSES.

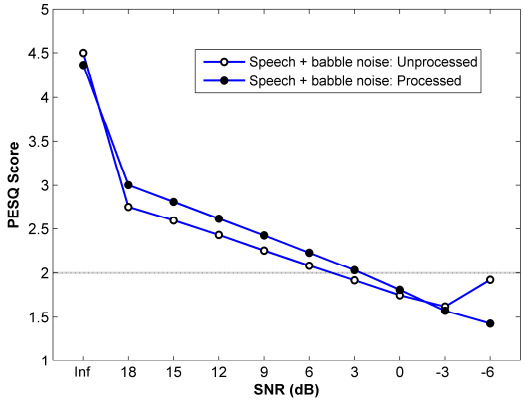


(a) Noise: white

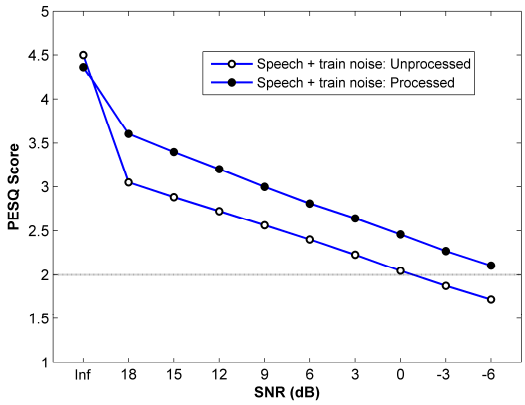
(b) Noise: pink



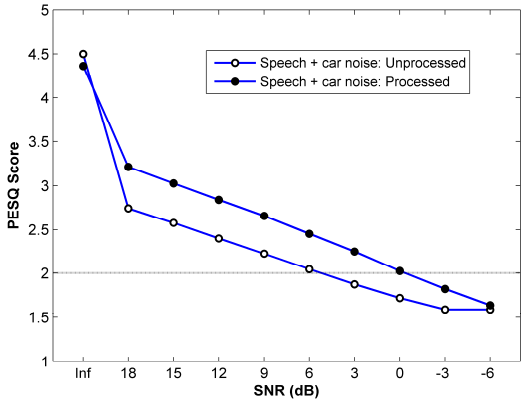
(c) Noise: street



(d) Noise: babble

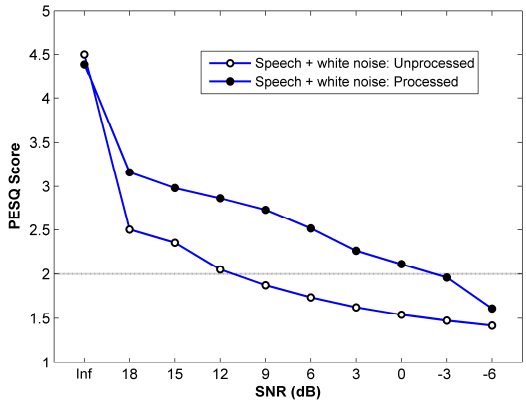


(e) Noise: train

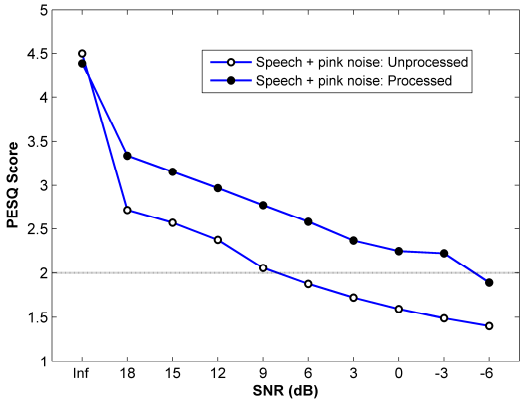


(f) Noise: car

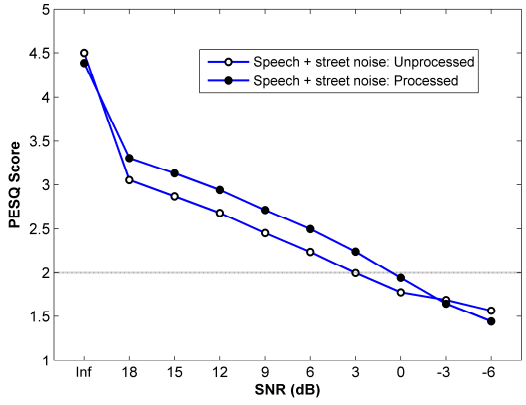
Fig. 2.8 PESQ score vs SNR for noisy and enhanced speech using log MMSE algorithm, speech material: NOIZEUS.



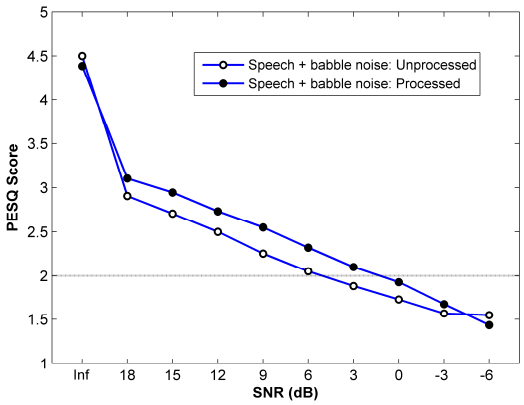
(a) Noise: white



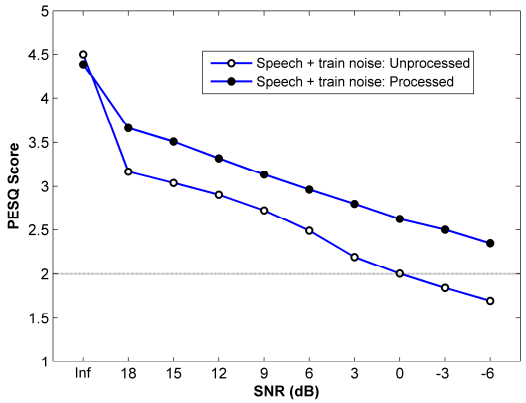
(b) Noise: pink



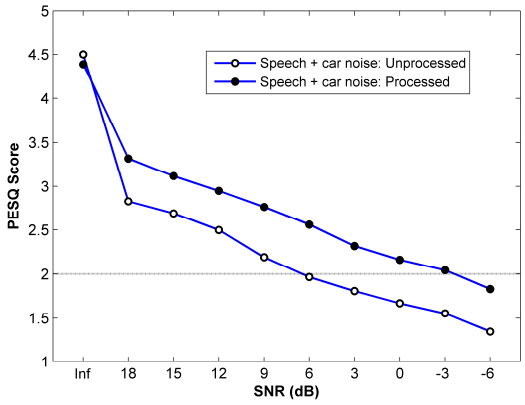
(c) Noise: street



(d) Noise: babble

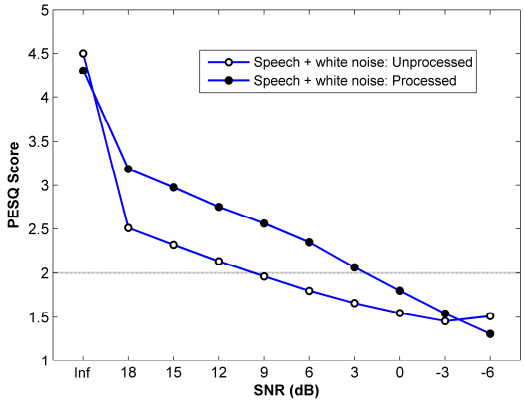


(e) Noise: train

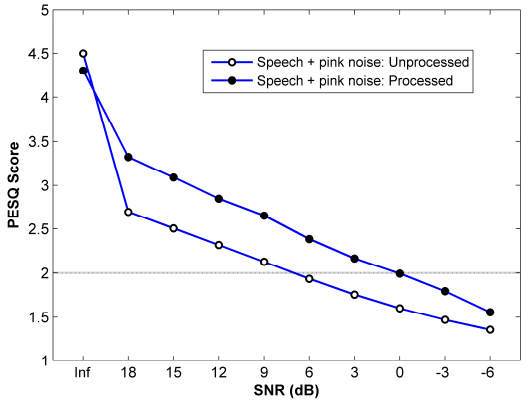


(f) Noise: car

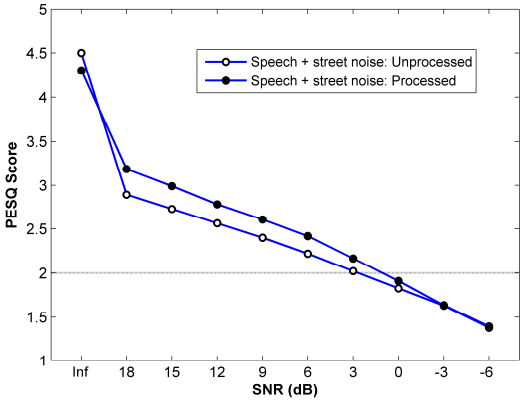
Fig. 2.9 PESQ score vs SNR for noisy and enhanced speech using log MMSE algorithm incorporating speech presence uncertainty, speech material: VHSES.



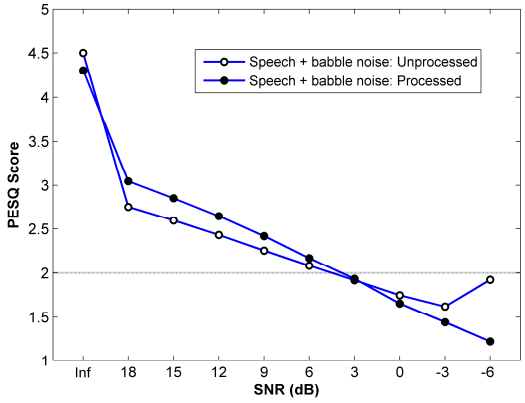
(a) Noise: white



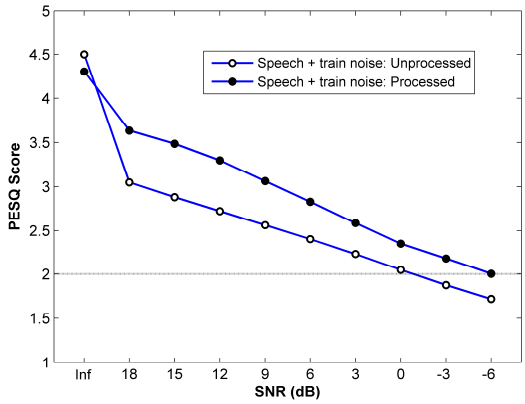
(b) Noise: pink



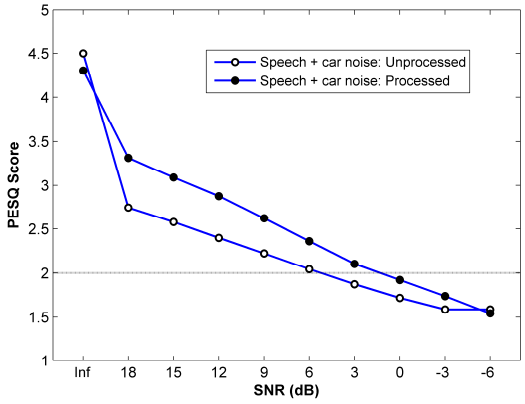
(c) Noise: street



(d) Noise: babble

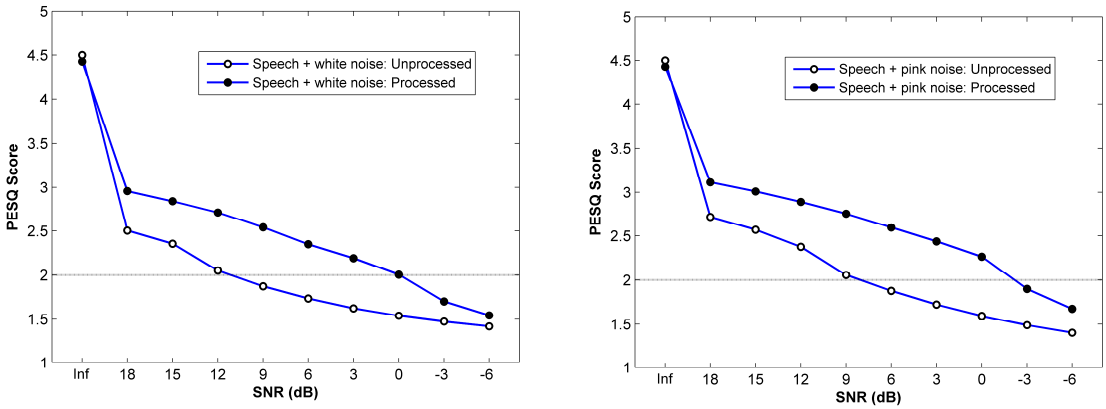


(e) Noise: train



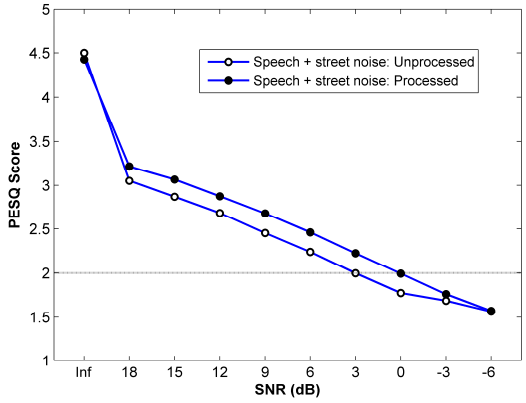
(f) Noise: car

Fig. 2.10 PESQ score vs SNR for noisy and enhanced speech using log MMSE algorithm incorporating speech presence uncertainty, speech material: NOIZEUS.

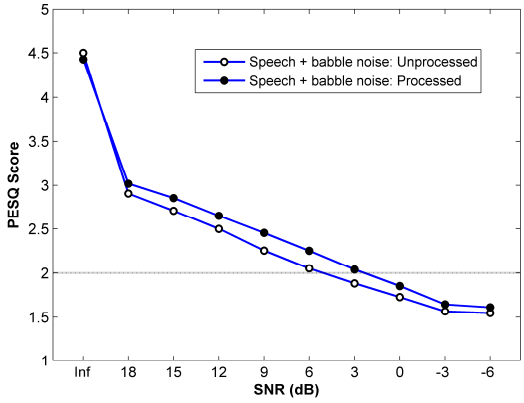


(a) Noise: white

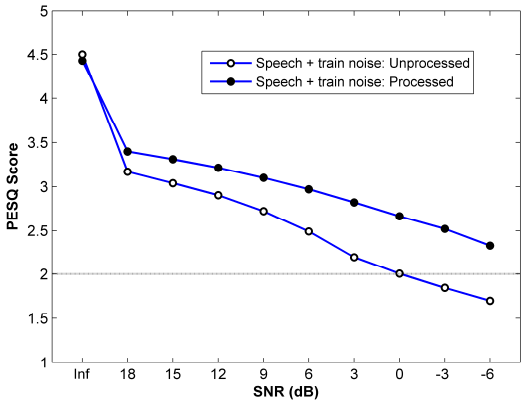
(b) Noise: pink



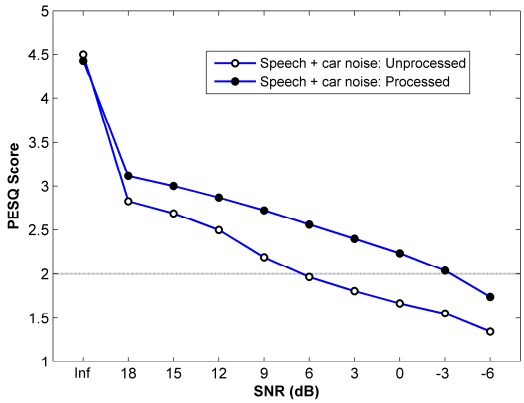
(c) Noise: street



(d) Noise: babble

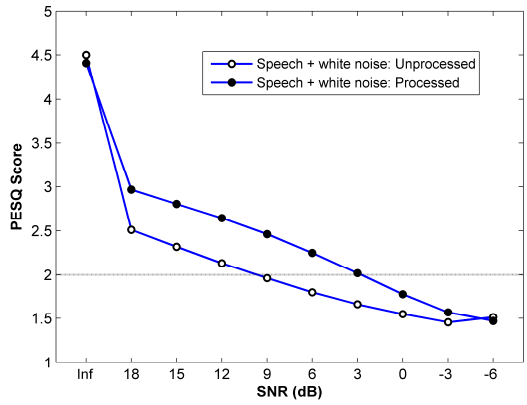


(e) Noise: train

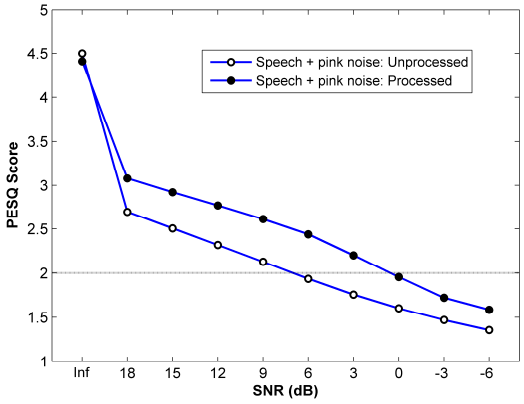


(f) Noise: car

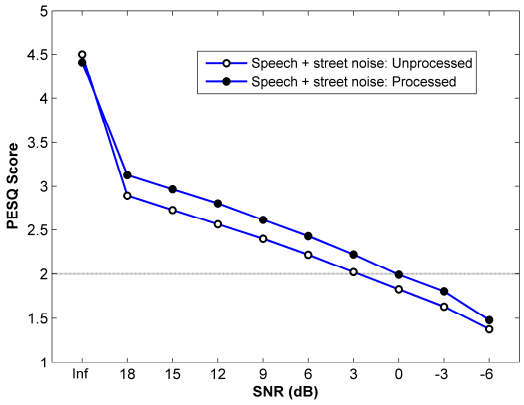
Fig. 2.11 PESQ score vs SNR for noisy and enhanced speech using Bayesian measure based on Euclidean distortion measure, speech material: VHSES.



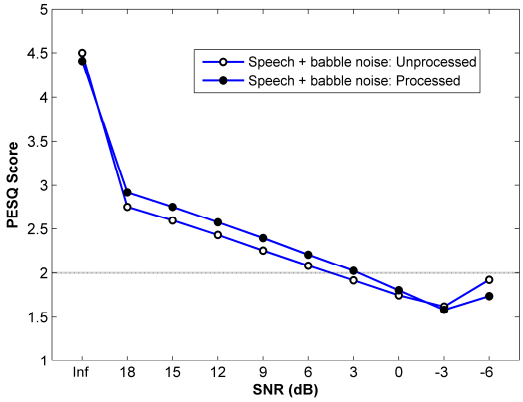
(a) Noise: white



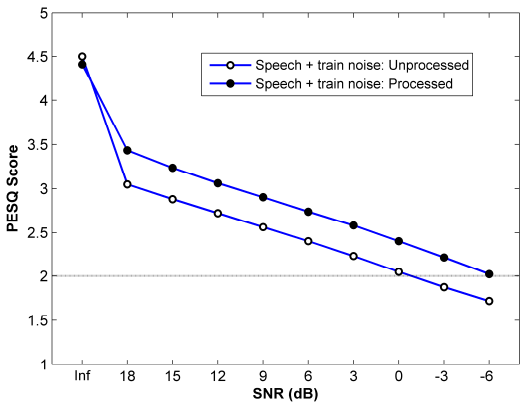
(b) Noise: pink



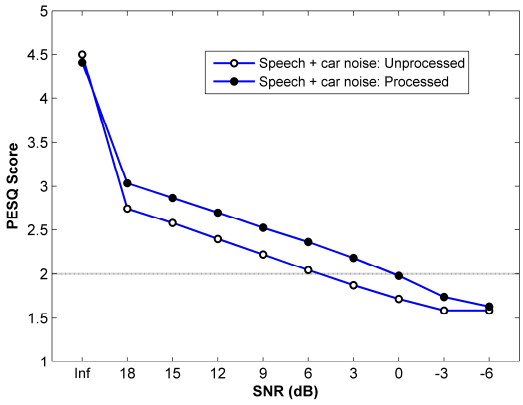
(c) Noise: street



(d) Noise: babble

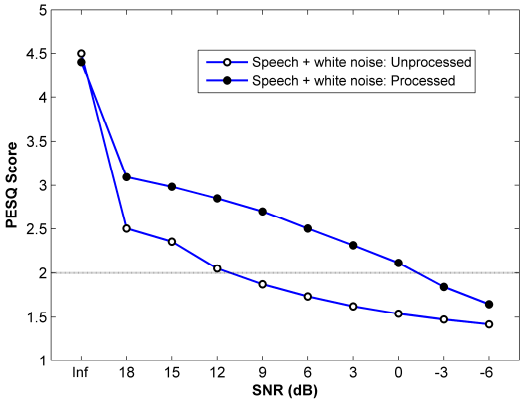


(e) Noise: train

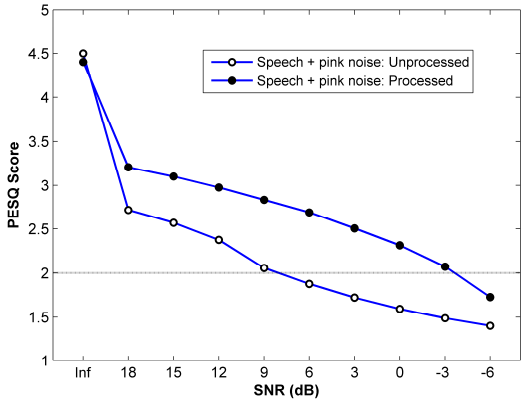


(f) Noise: car

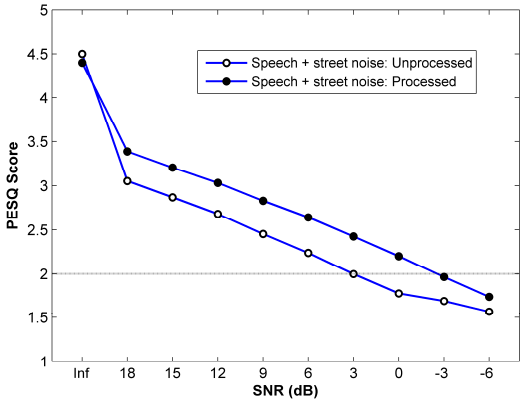
Fig. 2.12 PESQ score vs SNR for noisy and enhanced speech using Bayesian measure based on Euclidean distortion measure, speech material: NOIZEUS.



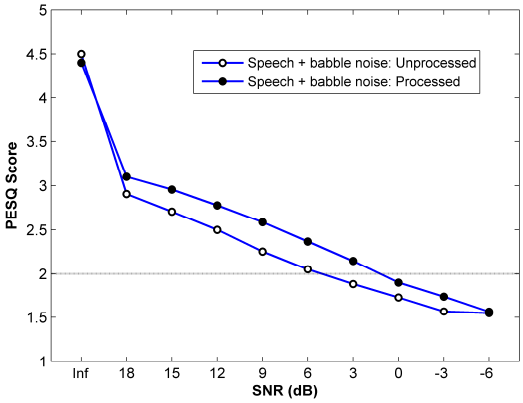
(a) Noise: white



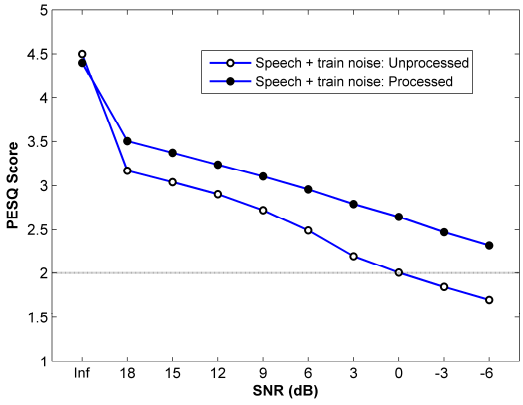
(b) Noise: pink



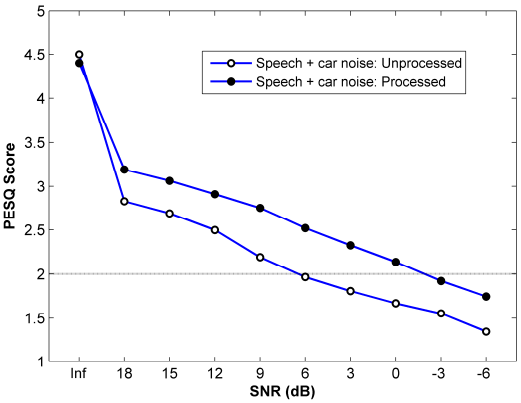
(c) Noise: street



(d) Noise: babble

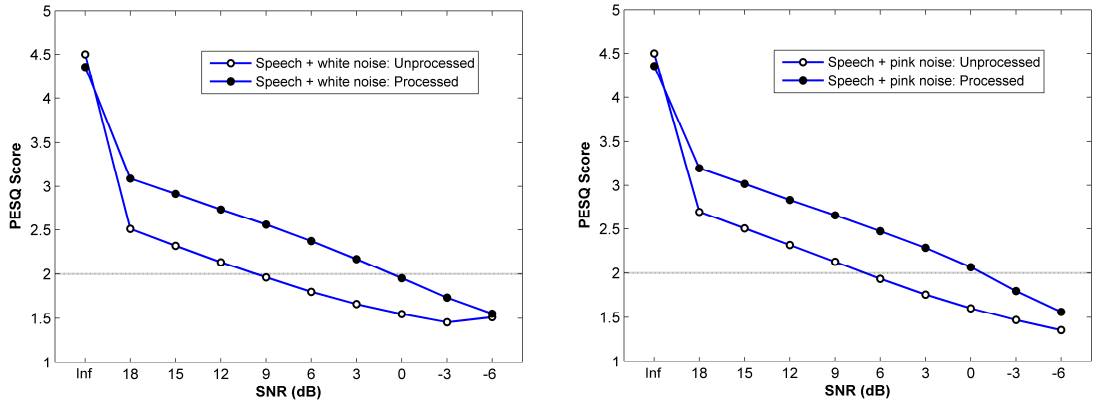


(e) Noise: train



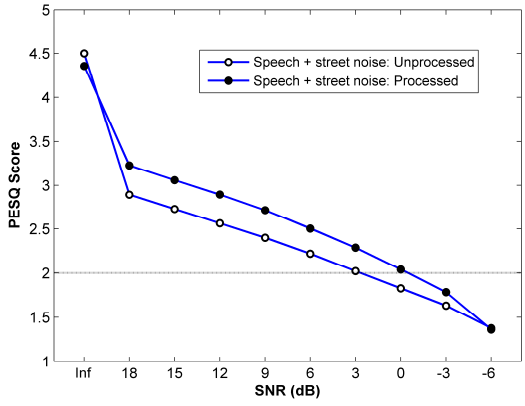
(f) Noise: car

Fig. 2.13 PESQ score vs SNR for noisy and enhanced speech using Bayesian measure based on cosh distortion measure, speech material: VHSES.

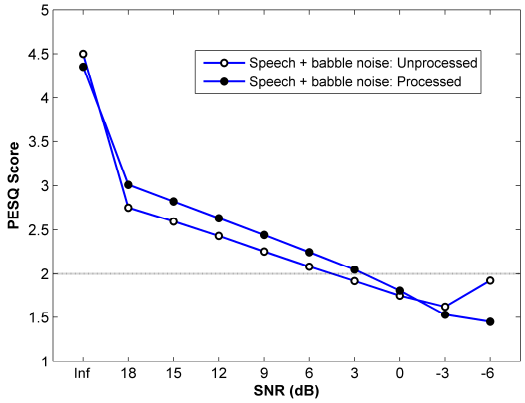


(a) Noise: white

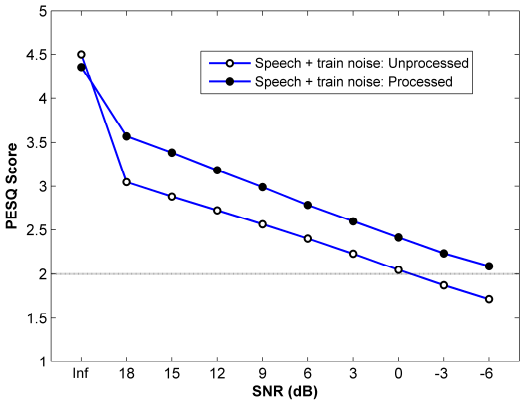
(b) Noise: pink



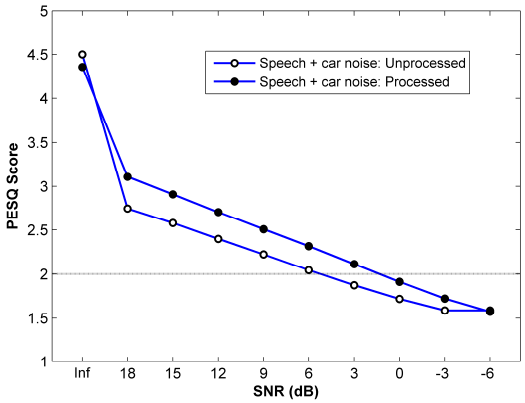
(c) Noise: street



(d) Noise: babble



(e) Noise: train



(f) Noise: car

Fig. 2.14 PESQ score vs SNR for noisy and enhanced speech using Bayesian measure based on cosh distortion measure, speech material: NOIZEUS.

Table 2.3 SNR advantage (in dB) obtained using six techniques

Enhancement technique	Noise type	SNR advantage (dB) for VHSES	SNR advantage (dB) for NOIZEUS
stsa_wcosh	White	13.0	9.0
	Babble	4.0	2.0
	Street	5.5	3.0
	Pink	12.0	8.0
	Train	10.0	6.0
	Car	7.5	4.0
logmmse	White	11.0	7.5
	Babble	3.0	2.0
	Street	3.0	3.0
	Pink	11.0	7.0
	Train	11.0	7.0
	Car	10.5	5.5
mmse	White	10.0	8.0
	Babble	3.0	1.5
	Street	3.0	3.0
	Pink	11.0	7.0
	Train	10.0	6.5
	Car	11.0	5.5
ga	White	8.0	5.0
	Babble	6.0	1.0
	Street	9.0	4.5
	Pink	10.5	5.5
	Train	9.0	5.0
	Car	8.5	5.0
stsa_weuclid	White	11.0	7.0
	Babble	3.0	1.5
	Street	3.0	3.0
	Pink	10.0	6.0
	Train	9.0	5.5
	Car	9.0	5.5
logmmse_spu3	White	13.5	7.0
	Babble	3.5	0.5
	Street	2.5	2.0
	Pink	13.0	7.0
	Train	10.0	5.0
	Car	10.0	4.0

2.4 Proposed investigation

For implementing using a low-power processor in a sensory aid, the noise suppression technique should have low computational complexity and signal delay (sum of algorithmic delays, computational delay, and I/O related delay) should be acceptable for face-to-face communication. On the basis of these considerations, generalized spectral subtraction along with cascaded-median based noise estimation is selected for real-time processing using a fixed-point processor. To select the optimal set of steps and parameters in the processing, detailed investigations are carried out using offline processing as presented in the next chapter. Based on these results obtained, implementation for real-time processing is carried out, as presented in the fourth chapter.

Chapter 3

INVESTIGATIONS USING OFFLINE PROCESSING

Detailed investigations were carried out using offline processing to select the optimal set of steps and parameters for real-time noise suppression. The processing steps in speech enhancement as shown in Fig. 2.1 are windowing, FFT calculation, noise spectrum estimation, magnitude spectral subtraction, complex spectrum calculation with noisy phase, and re-synthesis using IFFT with overlap-add. Implementation using 50% and 75% overlap resulted in similar enhanced output and hence 50% overlap was selected for investigations. It was observed that for different noises and SNR values, appropriate selection of subtraction factor α and floor factor β resulted in almost similar results for magnitude subtraction (exponent factor $\gamma = 1$) and power subtraction ($\gamma = 2$). The results of magnitude subtraction showed higher tolerances to variation in the values of α and β, and hence only magnitude subtraction was used. Investigations involved studying the effect of (a) noise estimation, (b) analysis window length and noise estimation duration, (c) parameters of generalized spectral subtraction, (d) phase estimation. Finally the performance of the method selected for real-time processing was compared with some of the methods reported in the literature. The evaluation method used and the investigations are presented in the subsequent sections. The results are discussed in the last section.

3.1 Evaluation method

Implementation of signal processing for speech enhancement was carried out using Matlab for investigating the effects of various steps and parameters. The evaluation involved using informal listening and an objective evaluation using perceptual evaluation of speech quality (PESQ) measure [10], [30]. This objective measure is a prediction of the subjective mean opinion score (MOS) of the degraded speech and is calculated from the difference between the loudness spectra of level-equalized and time aligned original and degraded signals.

Investigations were carried on speech materials taken from the NOIZEUS and VHSES as described earlier in section 2.3. Noisy speech was generated by adding white, babble, street, pink, car and train noises at SNR of 18, 15, 12, 9, 6, 3, 0, -3, and -6 dB. The babble, street, train, and car noises were taken from AURORA database [32].

For SNR calculation during noise addition RMS value of strong speech (vowel or vowel like) segments of the speech material was considered. The speech signal segmented using 20 ms rectangular window with 75% overlap. Speech energy in all frames was calculated to find the peak energy E_p. A threshold value, E_{th}, was selected to be 20 dB below E_p. The speech segments with energy in the range [E_{th} E_p] were used to calculate the RMS. Same method was used to calculate the RMS value of the noise. Based on the calculated speech and noise RMS values a scaling factor for noise was determined to get an appropriate SNR. The scaled noise was then added to clean speech signal to get the desired noisy speech signal. Then the peak RMS of the generated noisy speech signal is normalized to 0.25.

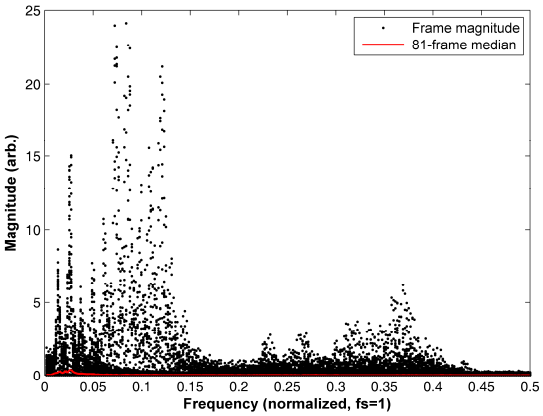
3.2 Investigation on noise estimation

Investigations were carried out to examine the distribution of magnitude spectrum of noise, speech, and noisy speech and to compare the noise estimations using mean, median and minimum statistics. Figure 3.1 and Figure 3.2 show the scatter plot of magnitude spectra superimposed with median for speech, noise and noisy speech, for two types of speech material and different noises. Plots were obtained for the magnitude of initial frames 20 to 101. Processing was carried with window length of 30 ms, 50% overlap and 512-point FFT. It is seen in Figure 3.1 (a) and Figure 3.2 (a) that the spectral magnitude of clean speech is high only for a few frames and hence the median value is very low. For the noisy speech signal, the medians increase because of added noise. Scatter plots of the magnitude spectrum of noisy speech signals with different noises and with different SNR's are given in Appendix A, and these plots show the same pattern. In addition to the median, mean and minimum were also calculated and plotted. From Figure 3.3, it is seen that median and minimum of noisy speech track the noise median and minimum respectively at almost all the frequencies, while mean of noisy speech tracks the noise mean at higher frequencies and clean speech mean at lower frequencies. More plots of scatter, mean, median, and minimum of different

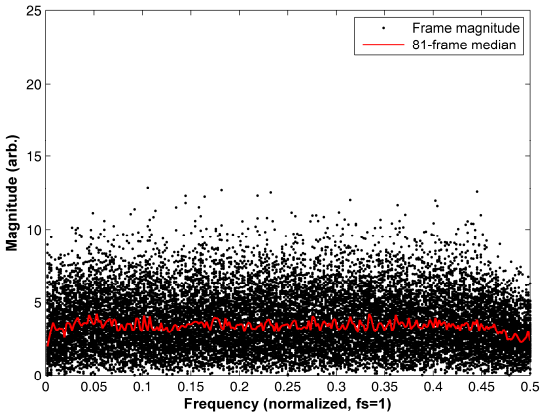
noises with 0 dB SNR are given in Appendix A and these show the same pattern. Although minimum tracks the noise, the minimum needs to be multiplied by a factor to get the correct noise magnitude. Errors in estimating the factor may lead to over-subtraction and killing of speech in weaker segments. Figure 3.4 shows the relative RMS error. It is calculated as a dB value of the RMS spectral samples of the estimation error with reference to the RMS of the spectral samples of the noise, i.e.

$$\varsigma_{\mathrm{dB}} = 10\log\frac{\sum[\hat{D}(k) - D(k)]^2}{\sum[D(k)]^2} \tag{3.1}$$

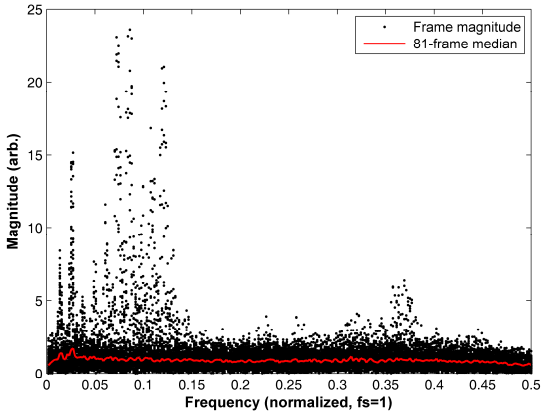
where $D(k)$ is the estimate from the noise spectrum and $\hat{D}(k)$ is the estimation obtained from the noisy speech. The RMS error in dB decreases as the SNR decreases for both the test materials. At 0 dB SNR, the values of RMS error in estimating the noise from noisy speech were -18.4, -12.8, -18.2, -18.4, -22.1, and -19.6 dB, respectively, for VHSES. The corresponding values for NOIZEUS were -17.4, -12.1, -16.1, -15.9, -18.4, and -16.4 dB. As these errors are small, median of the noisy speech may be considered as a suitable estimate of noise.



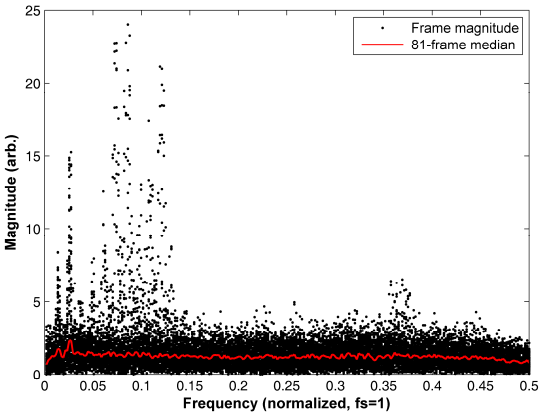
(a) Clean speech signal



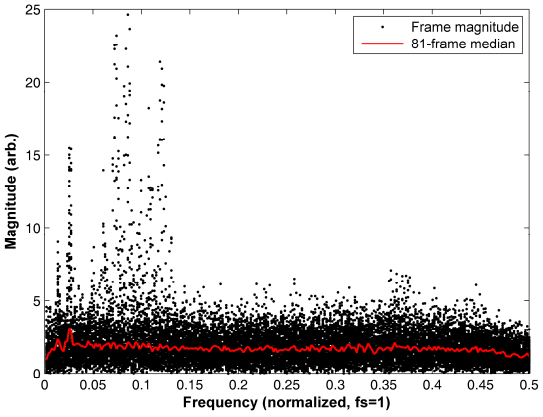
(b) White noise



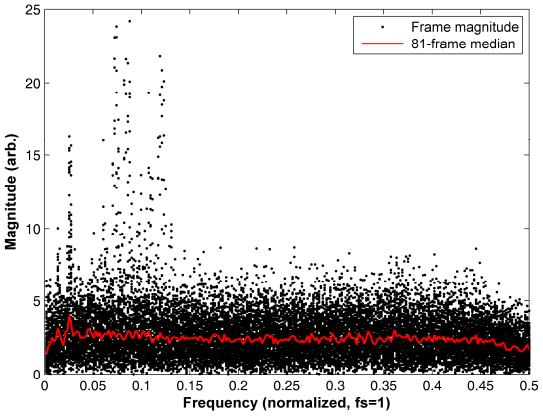
(c) Noisy speech: white noise, 6 dB SNR



(d) Noisy speech: white noise, 3 dB SNR

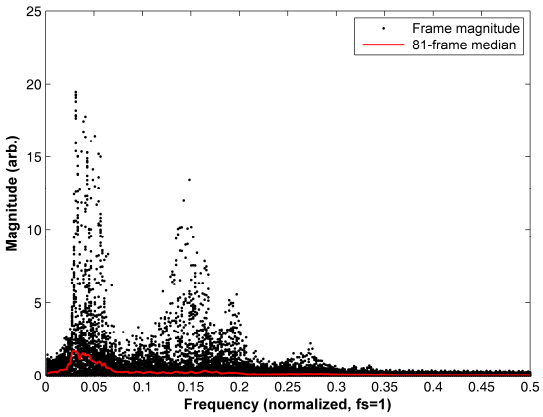


(e) Noisy speech: white noise, 0 dB SNR

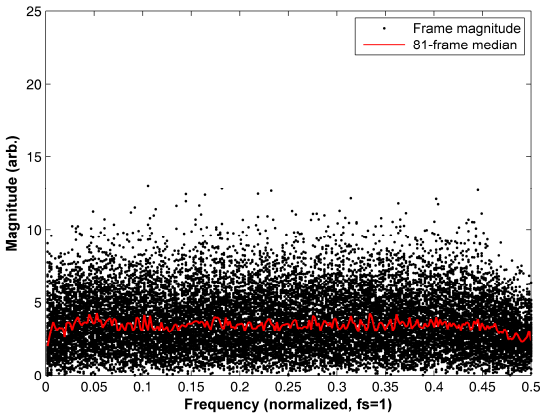


(f) Noisy speech: white noise, -3 dB SNR

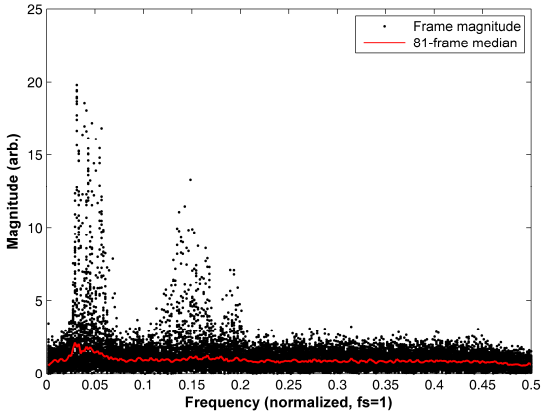
Fig. 3.1 Scatter plots of magnitude spectra of clean speech signal, noise, and noisy (white) speech signals. Speech material: VHSES



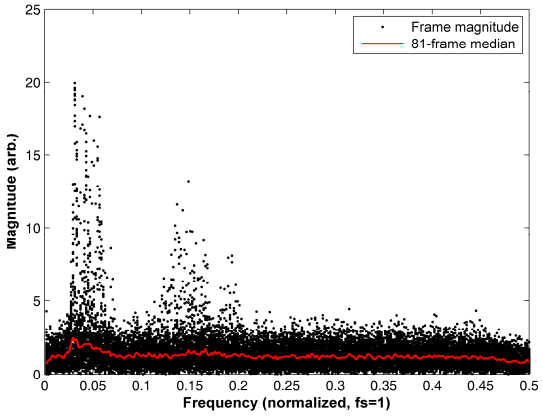
(a) Clean speech signal



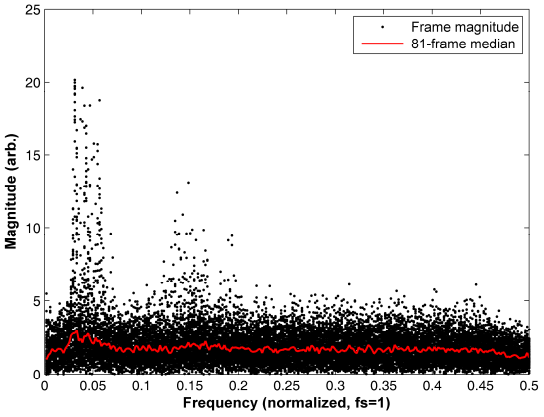
(b) White noise



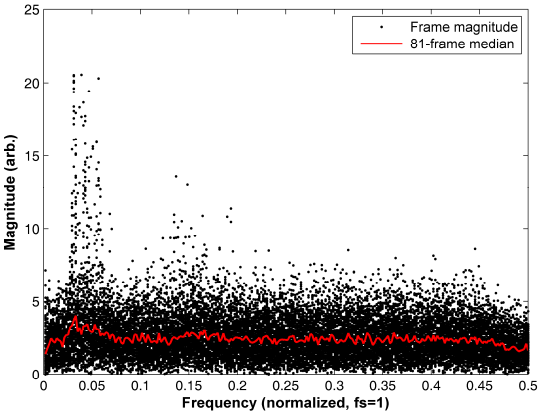
(c) Noisy speech: white noise, 6 dB SNR



(d) Noisy speech: white noise, 3 dB SNR

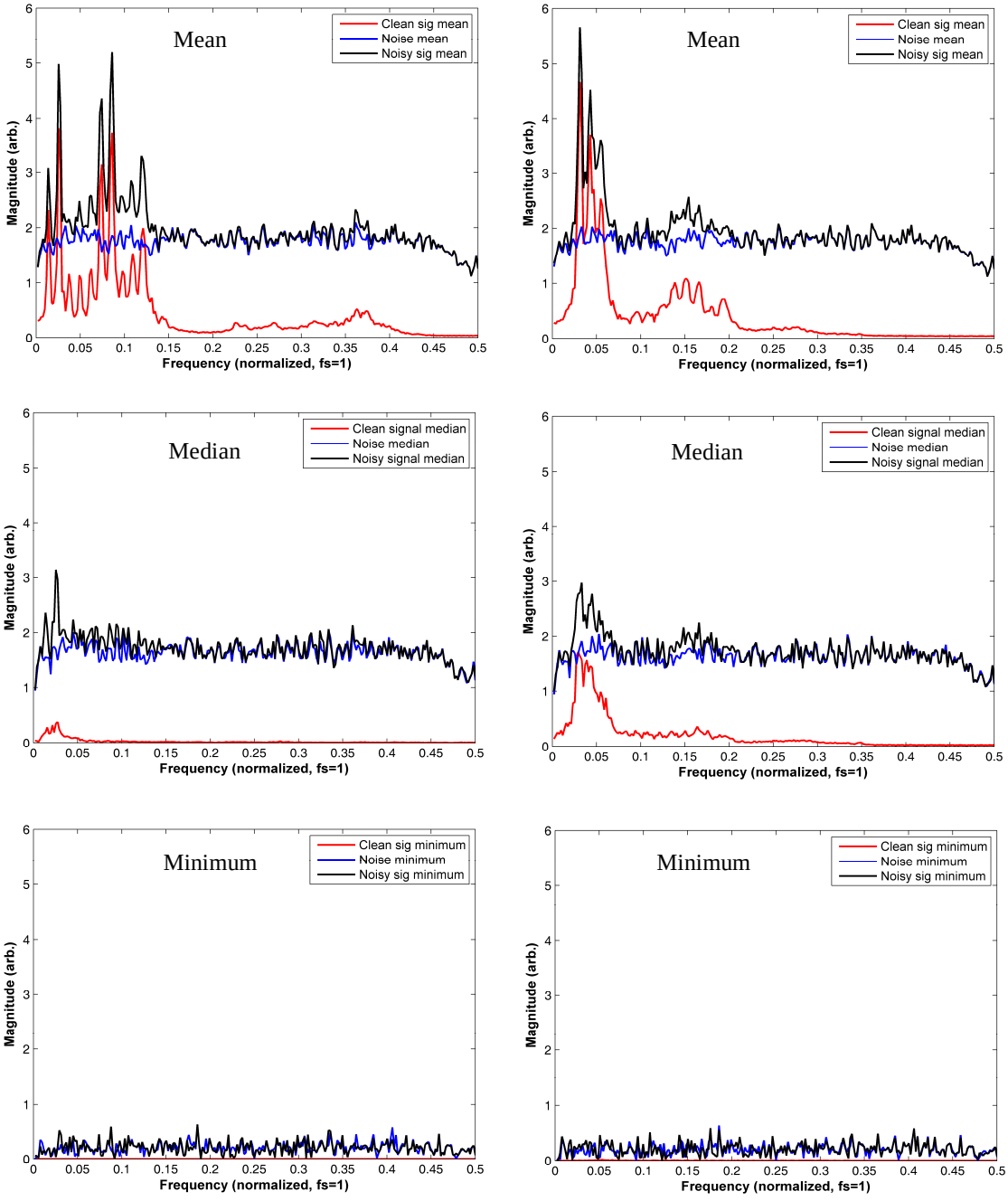


(e) Noisy speech: white noise, 0 dB SNR



(f) Noisy speech: white noise, -3 dB SNR

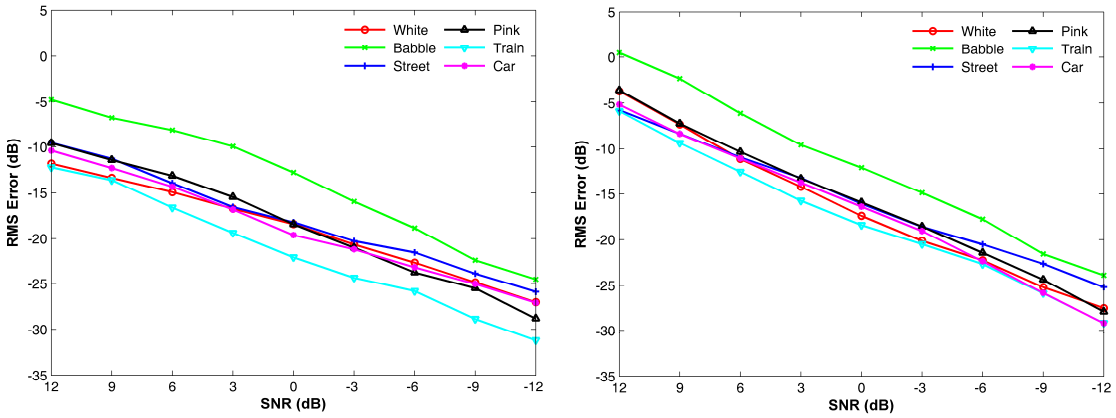
Fig. 3.2 Scatter plots of magnitude spectra of clean speech signal, noise, and noisy (white) speech signals. Speech material: NOIZEUS



(a) Speech material: VHSES

(a) Speech material: NOIZEUS

Fig. 3.3 Mean, median, and minimum of magnitude spectra of clean speech signal, noise and noisy speech (white, SNR: 0 dB).

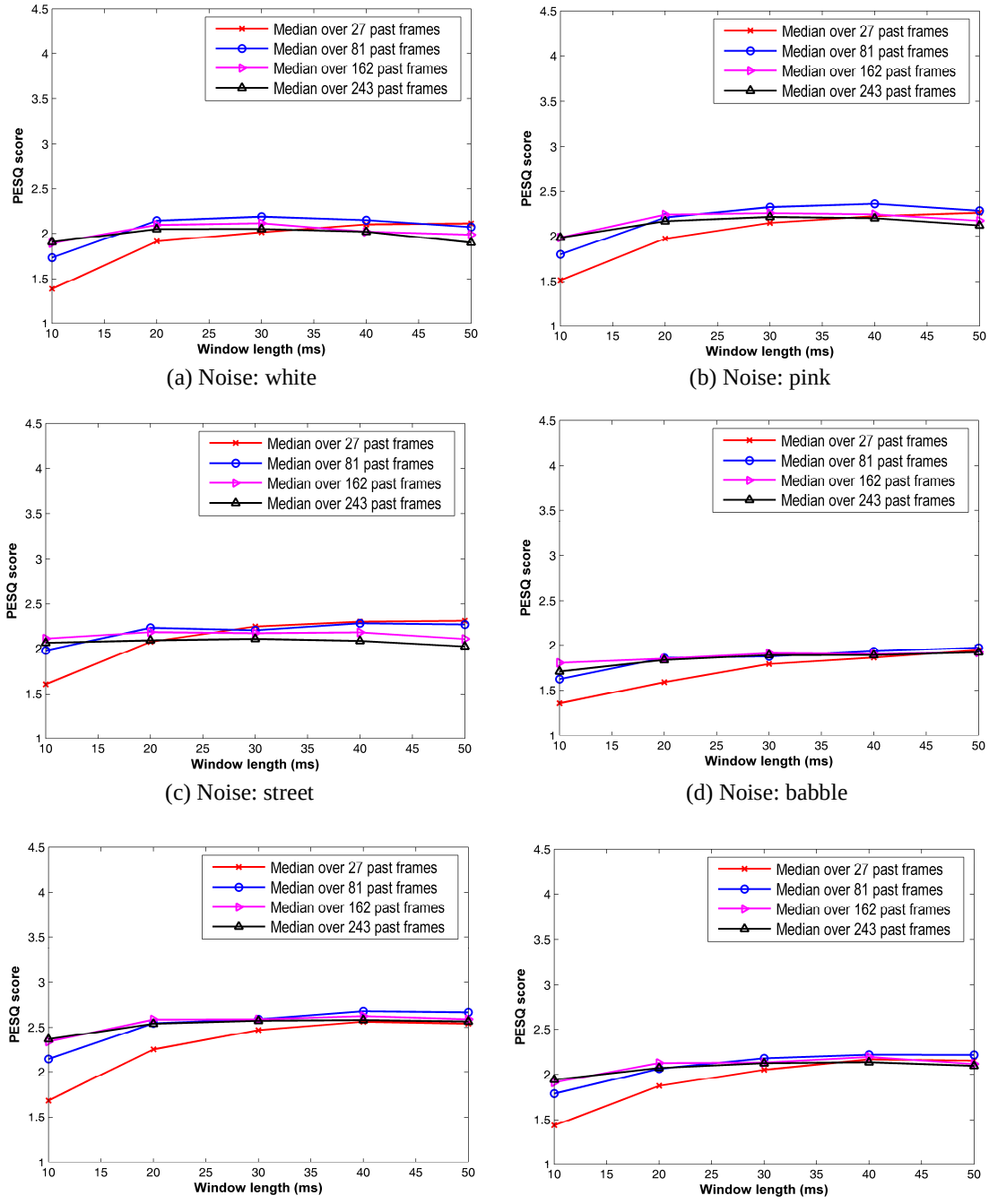


(a) Speech material: VHSES (b) Speech material: NOIZEUS

Fig. 3.4 Relative RMS error (dB) for different noises for two speech materials

3.3 Effect of window length and noise estimation duration

Investigations were carried out to study the effects of noise estimation duration and window length on noise suppression. Analysis was carried out using rectangular window with 50% overlap and MBNE as noise estimate. Figure 3.5 and Figure 3.6 show the PESQ score for the enhanced speech as function of window length, for different types of noises at SNR 0 dB. The plots are shown for noise estimation over 27, 81, 162, and 243 frames. The PESQ score was high in most of the cases for the noise estimation using 81 past frames and window length of 20 – 40 ms. Hence further investigations are carried out using 30 ms window length and noise is estimated using 81 past frames. It corresponds to noise estimation duration of approximately 1.2 s. Since MBNE has large memory requirement and sorting of past frames is computation intensive, an alternative method involving 3-frame 4-stage cascaded median as an approximation to median (described in section 2.2.5) is used for estimating the noise [28], [29]. Informal listening showed that the enhanced speech signals using spectral subtraction with MBNE and CMBNE for various noise types sounded almost the same. Table 3.1 and Table 3.2 show the PESQ scores of enhanced speech using spectral subtraction with median based noise estimate and cascaded-median based noise estimate for the two speech materials. The PESQ scores for the QBNE and CMBNE are almost the same, with a maximum difference of 0.06. Based on the results, it may be concluded that CMBNE method can be used as computationally efficient substitute for MBNE.



(a) Noise: white

(b) Noise: pink

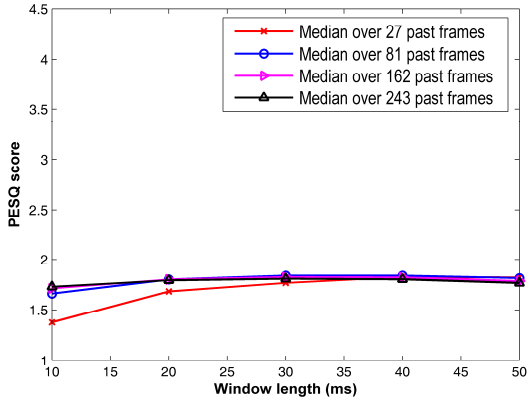
(c) Noise: street

(d) Noise: babble

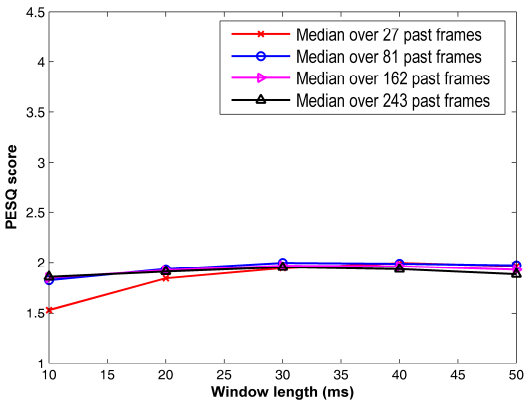
(e) Noise: train

(f) Noise: car

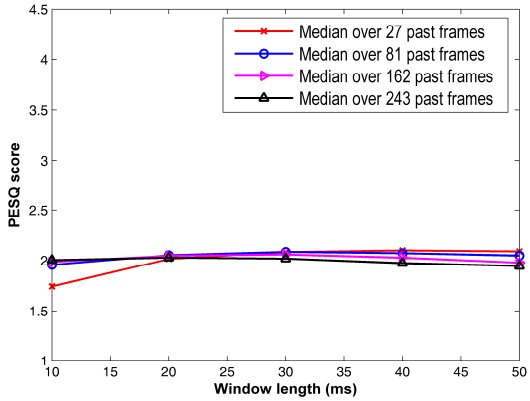
Fig. 3.5 PESQ score for the enhanced noisy speech, Speech: VHSES, SNR: 0 dB. α = 2.5 (white, pink) & 2 (babble, street, train, car), β = 0.001, and γ = 1.



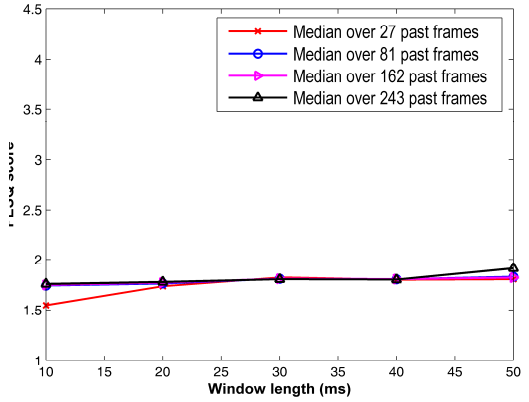
(a) Noise: white



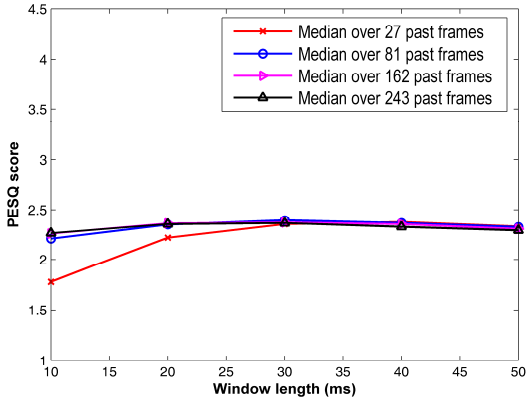
(b) Noise: pink



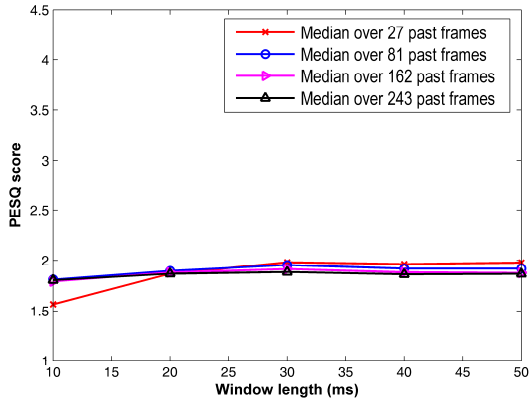
(c) Noise: street



(d) Noise: babble



(e) Noise: train



(f) Noise: car

Fig. 3.6 PESQ score for the enhanced noisy speech, Speech: NOIZEUS, SNR: 0 dB. α = 1.6 (white, pink, train, car) &1. 2 (babble, street), β = 0.01, and γ = 1.

Table 3.1 Comparison of enhanced speech signal using MBNE and CMBNE using PESQ score. Speech: VHSES, SNR: 0 dB, α = 2.5 (white, pink) & 2 (babble, street, train, car), β = 0.001, and γ = 1.

Noise type	PESQ score with clean speech as reference		
	Un proc.	Proc. using MBNE	Proc. using CMBNE
White	1.54	2.19	2.13
Babble	1.73	1.87	1.91
Street	1.78	2.20	2.14
Pink	1.59	2.34	2.30
Train	2.00	2.61	2.62
Car	1.67	2.18	2.14

Table 3.2 Comparison of enhanced speech signal using MBNE and CMBNE using PESQ score. Speech: NOIZEUS, SNR: 0 dB, α = 1.6 (white, pink, train, car) & 1.2 (babble, street), β = 0.01, and γ = 1.

Noise type	PESQ score with clean speech as reference		
	Un proc.	Proc. using MBNE	Proc. using CMBNE
White	1.55	1.84	1.84
Babble	1.75	1.80	1.81
Street	1.83	2.08	2.04
Pink	1.60	2.00	1.98
Train	2.05	2.40	2.35
Car	1.72	1.95	1.89

3.4 Effect of spectral subtraction parameters

Investigations were carried out for studying the effect of varying the values of α and β in the generalized spectral subtraction as given by [11], with γ = 1. Analysis was carried out using 50% overlap, rectangular window of length 30 ms and 3-point 4-stage cascaded-median based noise estimation over 81 frames i.e. for estimation duration of 1.215 s. Table 3.3 and Table 3.4 show the PESQ score for the enhanced speech using different sets of α and β. For speech material VHSES, the best PESQ scores are obtained for α as 2 – 2.5 and β = 0.01. For speech material NOIZEUS, the scores were best for α as 1.2 – 1.6 and β = 0.01. The processed output had perceptible amount of musical noise. To further improve the quality and intelligibility of the processed output, investigations need to be carried out with multi-band spectral subtraction [12] and geometric approach to spectral subtraction [13].

Table 3.3 PESQ score for the enhanced speech using spectral subtraction, for different types of noise. Material: VHSES, SNR: 0 dB, $\gamma = 1$, noise estimation: CMBNE.

(a) Noise: white

α	β			
	0	0.001	0.01	0.1
0	1.54	1.54	1.54	1.54
1.0	1.74	1.74	1.74	1.74
1.5	2.01	2.01	2.01	1.94
2.0	2.12	2.12	2.14	2.07
2.5	2.11	2.11	2.13	2.02
3.0	2.04	2.04	1.97	1.96
3.5	2.02	2.03	1.80	1.91

(b) Noise: pink

α	β			
	0	0.001	0.01	0.1
0	1.59	1.59	1.59	1.59
1.0	1.91	1.91	1.91	1.88
1.5	2.22	2.22	2.22	2.18
2.0	2.30	2.29	2.31	2.22
2.5	2.27	2.28	2.31	2.22
3.0	2.21	2.23	2.18	2.10
3.5	2.17	2.20	2.04	2.03

(c) Noise: street

α	β			
	0	0.001	0.01	0.1
0	1.77	1.77	1.77	1.77
1.0	2.10	2.10	2.11	2.13
1.5	2.15	2.15	2.17	2.23
2.0	2.14	2.14	2.18	2.27
2.5	2.09	2.11	2.16	2.29
3.0	2.06	2.08	2.12	2.28
3.5	2.03	2.05	2.09	2.27

(d) Noise: babble

α	β			
	0	0.001	0.01	0.1
0	1.73	1.73	1.73	1.73
1.0	1.87	1.87	1.87	1.88
1.5	1.91	1.91	1.91	1.95
2.0	1.91	1.91	1.93	2.01
2.5	1.86	1.86	1.92	2.04
3.0	1.80	1.81	1.89	2.06
3.5	1.71	1.73	1.84	2.07

(e) Noise: train

α	β			
	0	0.001	0.01	0.1
0	2.00	2.00	2.00	2.00
1.0	2.54	2.54	2.54	2.52
1.5	2.63	2.63	2.63	2.62
2.0	2.62	2.62	2.63	2.63
2.5	2.54	2.55	2.58	2.58
3.0	2.51	2.52	2.52	2.52
3.5	2.46	2.49	2.45	2.48

(f) Noise: car

α	β			
	0	0.001	0.01	0.1
0	1.67	1.67	1.67	1.67
1.0	1.96	1.96	1.96	1.94
1.5	2.13	2.13	2.14	2.13
2.0	2.13	2.13	2.15	2.18
2.5	2.05	2.06	2.11	2.14
3.0	1.99	2.00	2.03	2.11
3.5	1.93	1.95	1.95	2.08

Table 3.4 PESQ score for the enhanced speech using spectral subtraction, for different types of noise. Material: NOIZEUS, SNR: 0 dB, $\gamma = 1$, noise estimation: CMBNE.

(a) Noise: white

α	β			
	0	0.001	0.01	0.1
0	1.55	1.55	1.55	1.55
1.0	1.76	1.76	1.76	1.75
1.2	1.80	1.80	1.80	1.79
1.4	1.83	1.83	1.83	1.83
1.6	1.82	1.82	1.83	1.84
1.8	1.79	1.80	1.81	1.85
2.0	1.75	1.75	1.78	1.84

(b) Noise: pink

α	β			
	0	0.001	0.01	0.1
0	1.60	1.60	1.60	1.60
1.0	1.85	1.85	1.85	1.84
1.2	1.90	1.90	1.90	1.89
1.4	1.95	1.95	1.95	1.93
1.6	1.97	1.97	1.98	1.96
1.8	1.96	1.96	1.98	1.97
2.0	1.94	1.94	1.97	1.96

(c) Noise: street

α	β			
	0	0.001	0.01	0.1
0	1.83	1.83	1.83	1.83
1.0	2.01	2.01	2.02	2.04
1.2	2.02	2.02	2.04	2.06
1.4	2.01	2.01	2.03	2.07
1.6	1.99	1.99	2.02	2.08
1.8	1.94	1.95	1.99	2.08
2.0	1.91	1.92	1.97	2.07

(d) Noise: babble

α	β			
	0	0.001	0.01	0.1
0	1.75	1.75	1.75	1.75
1.0	1.82	1.82	1.82	1.85
1.2	1.80	1.80	1.81	1.85
1.4	1.78	1.78	1.79	1.86
1.6	1.75	1.75	1.76	1.86
1.8	1.72	1.72	1.73	1.86
2.0	1.68	1.68	1.70	1.86

(e) Noise: train

α	β			
	0	0.001	0.01	0.1
0	2.05	2.05	2.05	2.05
1.0	2.29	2.29	2.29	2.29
1.2	2.31	2.31	2.31	2.33
1.4	2.33	2.33	2.33	2.35
1.6	2.33	2.33	2.34	2.37
1.8	2.33	2.33	2.34	2.37
2.0	2.31	2.32	2.34	2.37

(f) Noise: car

α	β			
	0	0.001	0.01	0.1
0	1.72	1.72	1.72	1.72
1.0	1.89	1.89	1.89	1.90
1.2	1.89	1.89	1.90	1.93
1.4	1.89	1.89	1.90	1.95
1.6	1.88	1.88	1.89	1.96
1.8	1.85	1.85	1.88	1.96
2.0	1.82	1.82	1.86	1.95

3.5 Phase estimation for spectral subtraction

The magnitude spectrum of the noisy speech and the estimated magnitude spectrum of the noise are used to get the clean magnitude spectrum using (2.1). The resulting magnitude spectrum $|Y_n(k)|$ was combined with the original noisy phase to get the complex spectrum to be used for estimation of the enhanced speech signal. The IFFT of the complex spectrum was used with overlap-add for signal resynthesis. It may be expected that quality will improve if the phase spectrum was also noise free. Several methods have been proposed for phase reconstruction with a minimum phase assumption: iterative method of signal estimation from its magnitude spectrum [33], [34] and non-iterative methods using cepstral analysis [35]. The minimum phase assumption may be considered to be valid for speech signal produced with non-time varying vocal tract configuration and glottal excitation. However, this assumption may not be valid for signal segments produced with time-varying vocal tract configuration or when the source of excitation is present within the vocal tract.

3.5.1 Phase estimation methods

For minimum phase sequence, Quatieri and Oppenheim [33] reported an iterative algorithm, for estimating the phase spectrum of a sequence from the magnitude spectrum and one known sample. The method can be used for spectrum calculated using N-point DFT with N greater than twice the sequence length. The algorithm begins with the magnitude spectrum $|Y(k)|$ and initial phase estimate $\theta_0(k)$ to estimate the desired phase spectrum $\theta(k)$. Let the phase spectrum after jth iteration be $\theta_j(k)$. This is associated with the original magnitude spectrum to obtain the complex spectrum

$$Y_j(k) = |Y(k)|\mathrm{e}^{\mathrm{j}\theta_j(k)} \tag{2.4}$$

and the sequence is estimated as

$$y_j(n) = \mathrm{IDFT}\left[Y_j(k)\right] \tag{2.5}$$

For next iteration, the sequence $\tilde{y}_{j+1}(n)$ is calculated by imposing causality condition and the first known sample $y(0)$.

$$\tilde{y}_{j+1}(n) = \begin{cases} y(0), & n = 0 \\ y_j(n), & 1 \le n \le N/2 \\ 0 & N/2+1 \le n \le N-1 \end{cases} \tag{2.6}$$

The modified sequence is used to calculate complex spectrum

$$\tilde{Y}_{j+1}(k) = \text{DFT}\left[\tilde{y}_{j+1}(n)\right] \tag{2.7}$$

The revised phase spectrum is obtained as

$$\theta_{j+1}(k) = \angle \tilde{Y}_{j+1}(k) \tag{2.8}$$

If the mean square error between $|\tilde{Y}_{j+1}(k)|$ and $|Y(k)|$ is less than a threshold then the iteration will be halted and $\theta_{j+1}(k)$ is taken as the desired phase spectrum $\theta(k)$. Otherwise, the iteration is repeated. For reconstruction of signal from its short-time magnitude spectra with overlapped frames, the first sample from each frame is taken from the corresponding sample of the previous frame.

Nawab *et al.* proposed an iterative technique [34], for extrapolating a finite-length signal from its first known M samples and short-time magnitude spectra of the signal. The algorithm begins with an initial estimate of unknown samples as the following

$$\begin{matrix} y_0(n) = y(n), & \text{for } 0 \le n \le M-1 \\ 0 & \text{otherwise} \end{matrix} \tag{2.9}$$

This initial signal estimate and the magnitude spectrum $|Y(k)|$ are used to estimate the signal using an iterative process. Let the signal after jth iteration be $y_j(n)$. We calculate the complex spectrum as

$$Y_j(k) = \text{DFT}\left[y_j(n)\right] \tag{2.10}$$

Its phase spectrum is associated with $|Y(k)|$, to get the next iteration complex spectrum

$$\tilde{Y}_{j+1}(k) = \left|Y(k)\right| \mathrm{e}^{j\angle Y_j(k)} \tag{2.11}$$

and is used to calculate

$$\tilde{y}_{j+1}(n) = \text{IDFT}\left[\tilde{Y}_{j+1}(k)\right] \tag{2.12}$$

Signal estimate, $y_{j+1}(n)$ is updated as

$$y_{j+1}(n) = \begin{cases} y(n), & 0 \le n \le M-1 \\ \tilde{y}_{j+1}(n), & M \le n \le L-1 \\ 0 & L \le n \le N-1 \end{cases} \tag{2.13}$$

If the mean square error between $|\tilde{Y}_{j+1}(k)|$ and $|Y(k)|$ is less than a threshold then the iteration is halted and y_{j+1}(n) is taken as the estimated signal. Otherwise, the iteration is repeated. If the known sequence samples are equal to or greater than half the length of the sequence, then the sequence can be uniquely reconstructed. The method can be used for reconstructing a long sequence from its overlapped short-time magnitude spectra by taking the initial *M* samples in each frame from the corresponding samples of the preceding overlapped frame.

Phase reconstruction for the minimum phase sequence from the magnitude has been described by Rabiner and Schafer [35]. For minimum phase signals, the complex cepstrum can be obtained from the log of the magnitude of the discrete Fourier transform. Cepstrum coefficients are calculated from the magnitude spectrum as

$$c(n) = \text{IDFT}\left[\log\left(|Y(k)|\right)\right] \tag{2.14}$$

For a minimum phase sequence, the complex cepstrum coefficients can be calculated as

$$\hat{y}(n) = \begin{cases} c(0), & n = 0 \\ 2c(n), & 1 \le n \le N/2 \\ 0 & N/2+1 \le n < N-1 \end{cases} \tag{2.15}$$

From the complex cepstrum, complex spectrum is calculated as

$$Y(k) = \exp\left(\text{DFT}\left[\hat{y}(n)\right]\right) \tag{2.16}$$

Cepstrum computation using DFT suffers from circular aliasing, and to reduce these errors, the DFT size *N* should be much longer than the expected length of the cepstrum. We generally use $N \approx 3L$ where *L* was the sequence length [35]. We have also investigated use of non-iterative method to find an initial guess for the iterative method.

3.5.2 Results of different methods of phase estimation

Investigations were carried for resynthesis using, zero phase, phase estimating by Quatieri and Oppenheim iterative method, Nawab *et al.* iterative method, and cepstrum based non-iterative method. Analysis-synthesis was carried using (a) 50% overlap rectangular window,

and (b) 75% overlap rectangular window, and (c) Griffin-Lim method [36] of signal estimation from modified short-time Fourier transform (STFT). Griffin-Lim method [36] is based on least square error estimation (LSEE), i.e. minimizing the mean squared error between the STFT of the estimated signal and the modified STFT. The output signal is resynthesized by overlap-add of the segments obtained as IDFT of the modified complex spectra after multiplication with the analysis window. The window used should meet the requirement that sum of the squares of all the windows is unity, i.e.

$$\sum_{m=-\infty}^{\infty} w^2(mS-n) = 1 \tag{3.1}$$

For window length L and window shift $S = L/4$ corresponding to 75% overlap, this requirement is met by modified Hamming window, $w(n)$ with length L and $w(n)$ is given as

$$w(n) = \left[1/(4p^2+2q^2)^{0.5}\right]\left[p + q\cos\left(2\pi(n+0.5)/L\right)\right] \tag{3.2}$$

where, $p = 0.54$ and $q = -0.46$. Except for multiplication by the modified Hamming window and 75% overlap-add, the method does not involve any other computational complexities and hence it is suitable for real-time implementation.

Effect of phase estimation was first investigated on clean speech. Table 3.5 shows the PESQ score for the synthesized speech using various phase estimation methods with clean magnitude spectrum. For zero phase, it has been observed that reconstructed speech was poor in quality with 50% and 75% overlap using rectangular window. Use of Griffin-Lim method resulted in an increase in the quality. For the iterative method of Quatieri and Oppenheim [33], zero phase is assumed as the initial phase estimate and the first sample is obtained from the previous overlapped frame. After 20 iterations there was no significant improvement in the signal reconstruction. The reconstructed signal is better in quality as compared to reconstructed signal with zero phase. Nawab *et al.* [34] have proposed a similar iterative method to [33] by increasing the number of known samples. This method is implemented assuming the initial frame as zero valued samples and overlapped samples are assumed as known samples. The reconstructed signal does not improve much in quality compared to reconstructed signal using Quatieri and Oppenheim method. Using non-iterative method [35] investigations were carried. It has been found that minimum FFT length to be used for non-

Table 3.5 PESQ score for different phase estimation techniques for clean magnitude spectrum.

(A) Speech material: VHSES

Phase estimation	Signal estimation		
	Rect. win.		Griffin-Lim method
	50% overlap	75% overlap	
Original phase	4.50	4.50	4.50
Zero phase	2.16	2.15	2.74
Cepstrum method	2.76	2.62	2.97
Quatieri-Oppenheim, zero initial phase	2.77	2.57	2.98
Nawab *et al.*	2.76	2.93	2.79
Quatieri-Oppenheim, cepstrum initial phase	2.77	2.65	2.98

(B) Speech material: NOIZEUS

Phase estimation	Signal estimation		
	Rect. win.		Griffin-Lim method
	50% overlap	75% overlap	
Original phase	4.50	4.50	4.50
Zero phase	1.90	1.90	2.44
Cepstrum method	2.50	2.31	2.70
Quatieri-Oppenheim, zero initial phase	2.49	2.37	2.66
Nawab *et al.*	2.55	2.81	2.41
Quatieri-Oppenheim, cepstrum initial phase	2.52	2.40	2.68

iterative method should be greater than the twice of window length to avoid circular aliasing effect. Using non-iterative method, the phase was obtained and given as the initial phase estimate to the iterative method of Quatieri and Oppenheim [33]. The reconstructed signal was perceptually similar and has similar PESQ score to that of reconstructed signal using zero phase as initial phase.

Next effect of phase estimation was investigated during speech enhancement by spectral subtraction. PESQ score of the reconstructed speech using various phase estimation methods with enhanced magnitude spectrum are shown in Table 3.6. Enhanced magnitude

Table 3.6 PESQ score for different phase estimation techniques for enhanced magnitude spectrum.

(A) Speech material: VHSES, noise: white, SNR: 0 dB, $\alpha = 2$, $\beta = 0.01$, and $\gamma = 1$

Phase estimation	Signal estimation		
	Rect. win.		Griffin-Lim method
	50% overlap	75% overlap	
Original phase	2.14	2.11	1.96
Zero phase	2.05	2.01	1.85
Cepstrum method	2.13	2.10	1.79
Quatieri-Oppenheim, zero initial phase	2.12	2.04	1.82
Nawab *et al.*	2.12	2.01	1.76
Quatieri-Oppenheim, cepstrum initial phase	2.13	2.06	1.82

(B) Speech material: NOIZEUS, noise: white, SNR: 0 dB, $\alpha = 1.4$, $\beta = 0.01$, and $\gamma = 1$

Phase estimation	Signal estimation		
	Rect. win.		Griffin-Lim method
	50% overlap	75% overlap	
Original phase	1.83	1.83	1.74
Zero phase	1.68	1.67	1.60
Cepstrum method	1.77	1.73	1.56
Quatieri-Oppenheim, zero initial phase	1.74	1.73	1.51
Nawab *et al.*	1.65	1.72	1.54
Quatieri-Oppenheim, cepstrum initial phase	1.74	1.66	1.57

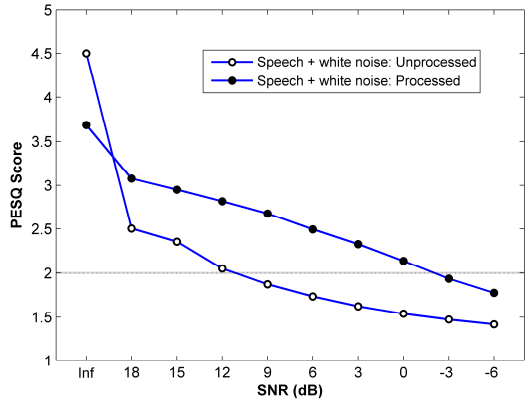
spectrum was obtained using spectral subtraction with cascaded-median based noise estimation. Subtraction parameters $\beta = 0.01$ and $\alpha = 2$ for "VHSES" and $\alpha = 1.4$ for " NOIZEUS " were used. Signal estimation is done using different phase estimation techniques as discussed in earlier section. The results showed that phase estimation by the different methods did not result in an improvement over use of noisy phase, although all of them gave better scores than use of zero phase. As the phase estimation methods involve additional computing, it may be concluded that use of noisy phase as used conventionally is the most suitable method for real-time processing.

Table 3.7 SNR advantage obtained using spectral subtraction with 3-point 4-stage cascaded median based noise estimation. α: as given below, $\beta = 0.01$, $\gamma = 1$.

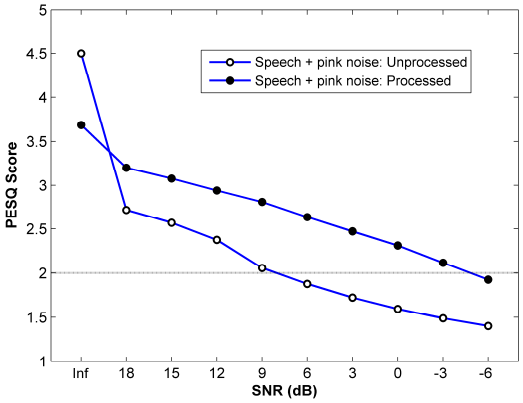
Noise type	Material: VHSES		Material: NOIZEUS	
	SNR advantage	Optimal α	SNR advantage	Optimal α
White	13.0	2.0	7.0	1.4
Babble	4.0	2.0	1.5	1.2
Street	5.0	2.5	3.5	1.4
Pink	12.5	2.0	6.5	1.4
Train	8.0	2.0	4.0	1.4
Car	9.0	2.0	4.0	1.4

3.6 Discussion

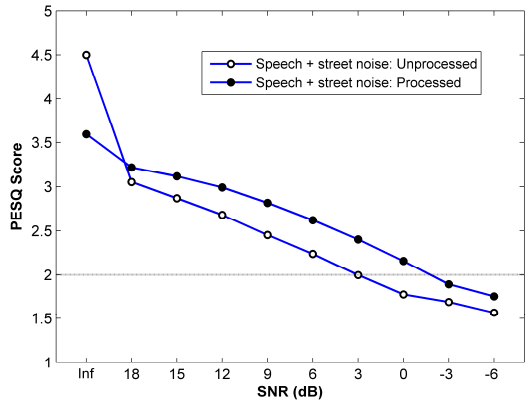
Processing was carried out with sampling frequency of 10 kHz and 30 ms frames (frame length $L = 300$ samples). As the processed outputs with FFT length $N = 512$ and higher were indistinguishable, $N = 512$ has been used in real-time processing. Informal listening showed that the processing significantly enhanced the speech for all noises with different SNR's and there was no audible roughness. While spectral floor factor $\beta = 0.01$ was found to be appropriate in all cases, most appropriate value of subtraction factor α varied over 2 – 2.5 for "VHSES" and 1.2 – 1.4 for "NOIZEUS". An objective evaluation was carried out using PESQ measure for different types of noise and SNR conditions. Figure 3.7 and Figure 3.8 show the PESQ score vs. SNR plot of unprocessed and processed signals for noisy speech signals. For unprocessed speech, the score decreased progressively with decrease in SNR. While processing of noise-free speech decreased the score from 4.5 to 3.7. SNR advantage was calculated using PESQ score vs. SNR plots at a score of 2.0, which is generally considered as lowest score for acceptable speech. Table 3.7 shows the SNR advantage for different types of noises and the optimal α used. It resulted in SNR advantage of approximately 4 – 13 dB for "VSHES" and 1.5 – 7 dB for "NOIZEUS" speech material. Thus the results show that the SNR advantage obtained with the proposed method is comparable to the best enhancement methods as evaluated earlier in section 2.3.



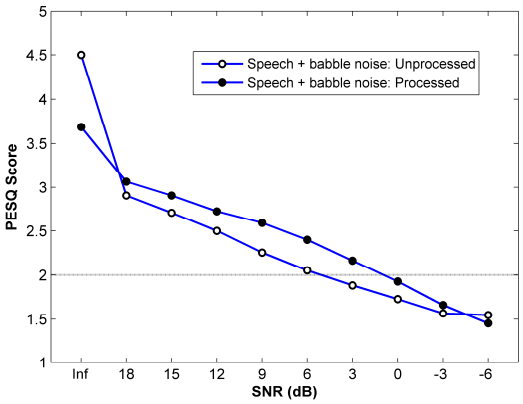
(a) Noise: white



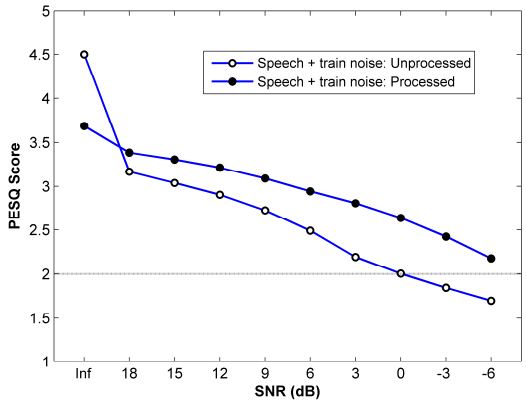
(b) Noise: pink



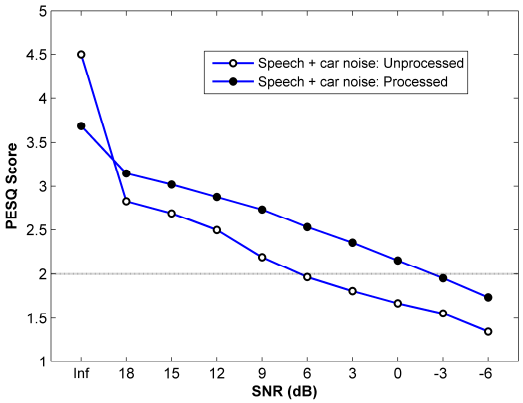
(c) Noise: street



(d) Noise: babble

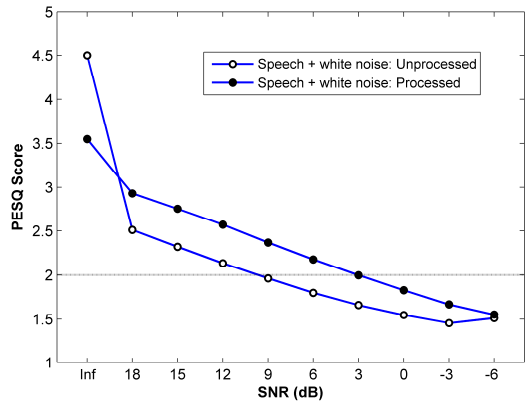


(e) Noise: train

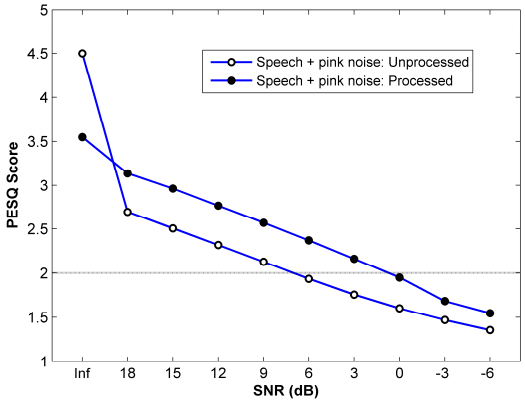


(f) Noise: car

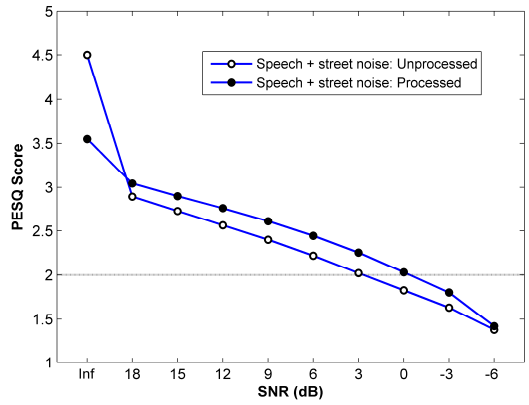
Fig. 3.7 PESQ score vs SNR for noisy and enhanced speech using spectral subtraction with 3-point 4-stage cascaded median based noise estimation, speech material: VHSES.



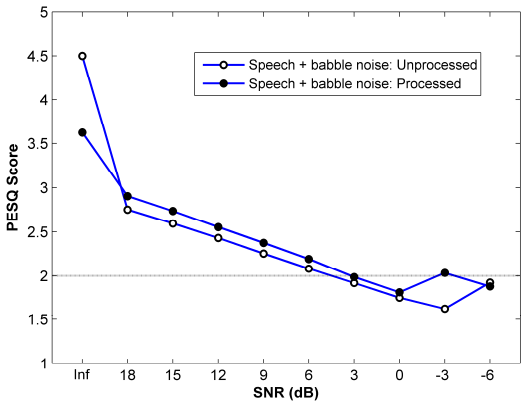
(a) Noise: white



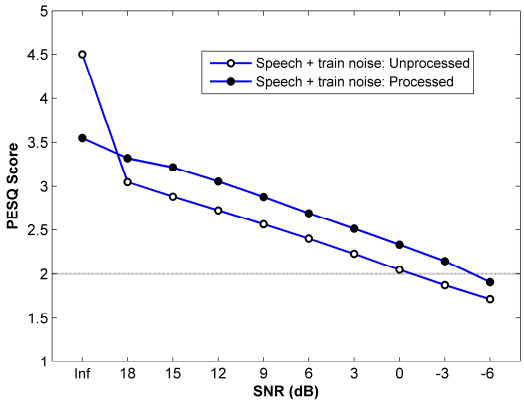
(b) Noise: pink



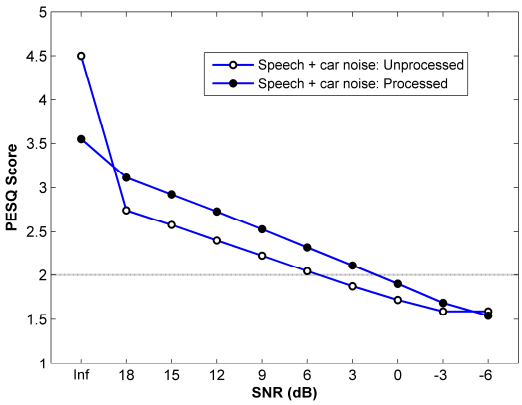
(c) Noise: street



(d) Noise: babble



(e) Noise: train



(f) Noise: car

Fig. 3.8 PESQ score vs SNR for noisy and enhanced speech using spectral subtraction with 3-point 4-stage cascaded median based noise estimation, speech material: NOIZEUS

Chapter 4

IMPLEMENTATION FOR REAL-TIME PROCESSING

In order to use it in sensory aids for the hearing impaired [1], [3], the noise suppression and spectral subtraction technique was implemented on a DSP board “Spectrum Digital eZdsp USB Stick” [37], based on 16-bit fixed point processor "TI TMS320C5515" [38] for real-time operation. The block diagram of the DSP Board is shown in Fig. 4.1. The board has embedded JTAG emulator XDS100 for source level debugging and 4 MB flash for user program. Line in, and headphone out connectors on the board may be used to give stereo input and recording the stereo output. The board has programmable codec "TI TLV320AIC3204" [39] with stereo ADC and DAC with 16/20/24/32-bit quantization and 8 – 192 kHz sampling. The program was written in C, using "TI CCStudio, ver. 4.0" as the development environment. Internal bus structure of the processor supports one 32-bit data read bus, two 16-bit data read buses, two 16-bit data write buses, one program bus, and peripheral DMA buses with capability of handling up to four 16-bit data reads and two 16-bit data writes in a cycle. The processor has a unified memory space of 16 MB with 320 KB on-chip RAM (including 64 KB dual access RAM), 128 KB on-chip ROM. It also has four 4-channel DMA controllers, three 32-bit timers, tightly coupled FFT hardware accelerator for efficiently computing 8–1024-point FFT. A complex number is stored as 4-byte word, with 16-bit real and 16-bit imaginary parts. The processor can be operated at a clock frequency of up to 120 MHz.

4.1 Implementation

The implementation uses one channel of the codec, with 16-bit quantization and 12 kHz sampling. A block diagram of the spectral subtraction for noisy speech in real-time is shown in Fig. 4.2. At the set sampling frequency, DMA channel-2 reads the ADC values into the input cyclic buffer and channel-0 writes the output cyclic buffer values to DAC. The input

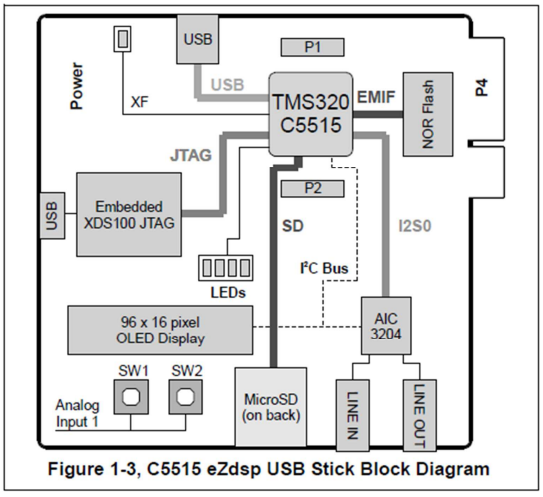


Fig. 4.1 Block diagram of TMS320C5515 eZdsp USB Stick [37]

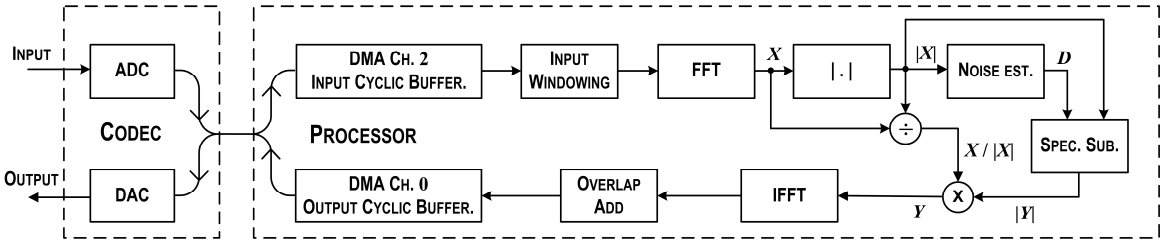


Fig. 4.2 Implementation of spectral subtraction on the DSP board [29]

samples, spectral values, and the processed samples are all stored as 4-byte words with 16-bit real and 16-bit imaginary parts, in order to reduce the conversion overheads. The input, output, data transfer, and buffering operations are devised for an efficient realization of the processing with 50% overlap and zero padding. As shown in Fig. 4.3, the input samples are acquired using a 3-block input cyclic buffer and the processed samples are output using a 2-block cyclic buffer, with *S*-word blocks and $S = L/2$. Pointers with cyclic values (.., 1, 2, 3, 1, ..) are used to track the current input and just-filled input blocks. They are initialized to 1 and 3, respectively. The current output and write-to output blocks are tracked by pointers with toggling values of 1 and 2, and initialized as 1 and 2, respectively. A DMA interrupt is generated when the current input block gets filled. All pointers are incremented cyclically. The DMA-mediated reading from ADC and writing to DAC are continued. The samples of the just-filled and the previous blocks are copied to the input data buffer, and are padded with

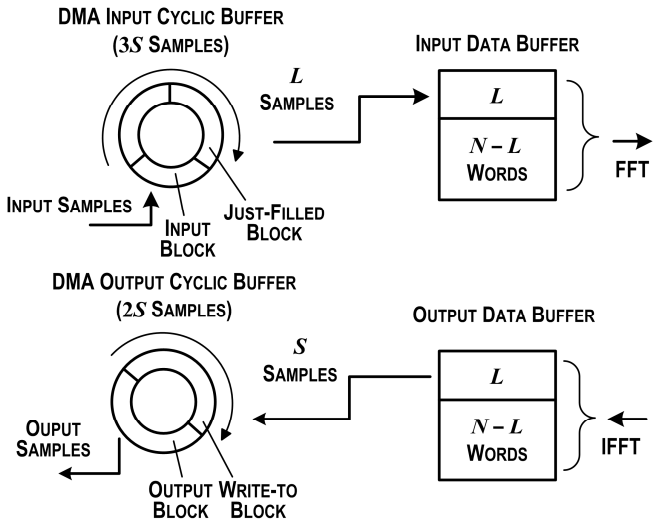


Fig. 4.3 Data transfer and buffering operations ($S = L/2$) [29]

$N-L$ zero-valued samples to serve as input to N-point FFT. The processing for noise estimation, spectral subtraction, and re-synthesis of output signal were implemented as discussed earlier with due care to avoid overflows.

The time domain segment is obtained as the first L samples of the real part of the N-point IFFT of the modified complex spectrum and stored in the output data buffer. A buffer of S samples is used for overlap-add operations. The first S samples of the output data buffer are added to the first S samples of the overlap buffer containing the partial results from the previous operation. The resulting S samples are written to the write-to output block. The last S samples of the output data buffer are copied to the overlap buffer.

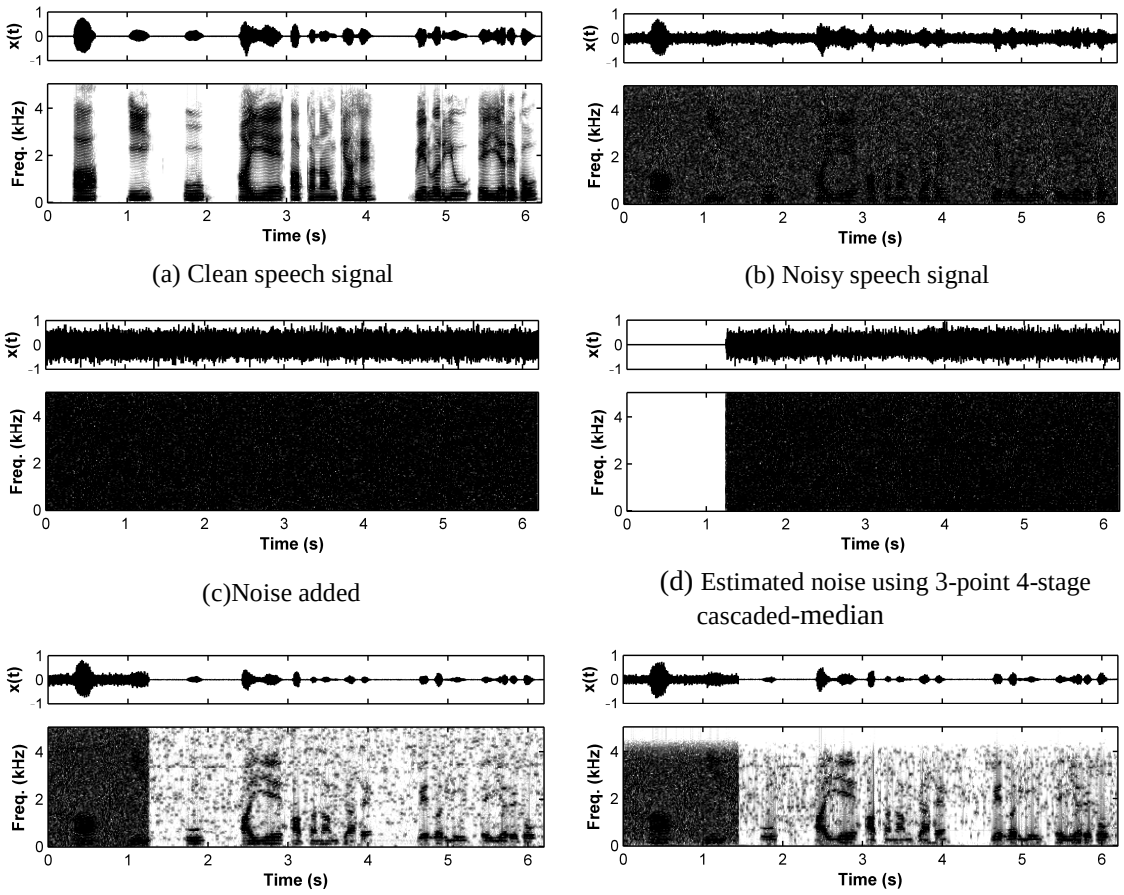
Based on the offline investigations, real-time processing was implemented using magnitude spectral subtraction with 3-point 4-stage cascaded-median based noise estimation, analysis-synthesis using 30 ms window with 50% overlap, and synthesis using phase spectrum of the noisy speech signal along with the enhanced magnitude spectrum. For testing of the program, value of β is kept as 0.01. The optimum value of α is selected as obtained from offline processing and given in Table 3.7. It is 1.2 – 1.4 for speech material with consonant-rich sentences (NOIZEUS) and as 2 – 2.5 for vowels and voiced consonant rich material (VHSES). In real-time processing code, the two factors are defined as macros, with

names "ALPHA" for subtraction factor α and "BETA" for spectral floor factor β. They should be assigned the values as ALPHA = round (8α), BETA = $1/\beta$.

4.2 Results

The real-time processing was carried out with sampling frequency of 10 kHz, L = 300, and N = 512. Speech enhancement with real-time processing was tested for speech signals mixed with different noises at different SNRs. For these tests, the noise added signal from the PC sound card was given to the codec input of the DSP board and the output from its codec was acquired through the sound card. As an example of processing, the noise-free speech, noisy speech with white noise at 3 dB SNR, noise added at 3 dB SNR, estimated noise using 3-point 4-stage cascaded-median, output from offline processing, and output from real-time processing are shown in Figure 4.4. Informal listening showed that the processed output from the DSP board was perceptually similar to the corresponding output from the offline implementation.

For an objective evaluation, PESQ score for the processed output with the noise-free speech as the reference signal was calculated. The scores were calculated for (a) unprocessed noisy speech (Unproc.), (b) offline processed output (Proc. Matlab), (c) unprocessed DSP output (Unproc. real-time), and (d) real-time processed DSP output (Proc. real-time). The plots of PESQ score vs. SNR are shown Figure 4.5 and Figure 4.6 for the two speech materials. For noise-free speech, passing the speech signal without any processing decreases the score from 4.5 to 3.3. This score serves as a reference for examining the scores after speech enhancement. Application of the processing on noise-free speech decreases the score from 4.5 to 3.7 for offline processing and from 3.3 to 2.9 for real-time processing. For unprocessed speech, decrease in SNR lowers PESQ score, from 2.5 at 18 dB to 1.5 at -6 dB, and the scores are almost similar for the two materials. Also for SNR of 18 dB and below, passing the signal through the DSP board does not contribute to any significant degradation of score. For both materials, both types of processing result in increase in scores. The offline processing introduces an improvement in the scores of approximately 0.57 – 0.80 for VHSES



(a) Clean speech signal

(b) Noisy speech signal

(c)Noise added

(d) Estimated noise using 3-point 4-stage cascaded-median

(e) Offline processed output signal

(f) Real-time processed output signal

Fig. 4.4 Processing of (-/a/-/i/-/u/– "aayiye aap kaa naam kyaa hai?" – "Where were you a year ago?", from a male speaker, with white noise at 3 dB SNR.

and 0.28 – 0.44 for NOIZEUS, for SNR from 18 dB to 0 dB. Real-time processing results in a slightly less improvement: 0.39 – 0.71 for VHSES and 0.22 – 0.32 for NOIZEUS.

The processor has maximum clock frequency of 120 MHz and the speech enhancement was found to be satisfactory for a clock frequency down to 16.4 MHz, indicating that the technique needed approximately 14% of the processing capacity at the clock frequency of 120 MHz and the rest could be used in implementing other processing as needed for a sensory aid. A comparison of input and output using a DSO showed a processing delay of 48 ms and it was found to be independent of the clock speed. The speech enhancement method has 1.5 frame algorithmic delay. Hence the frame size of 30 ms accounts for delay of 45 ms. The processing configuration does not contribute any additional

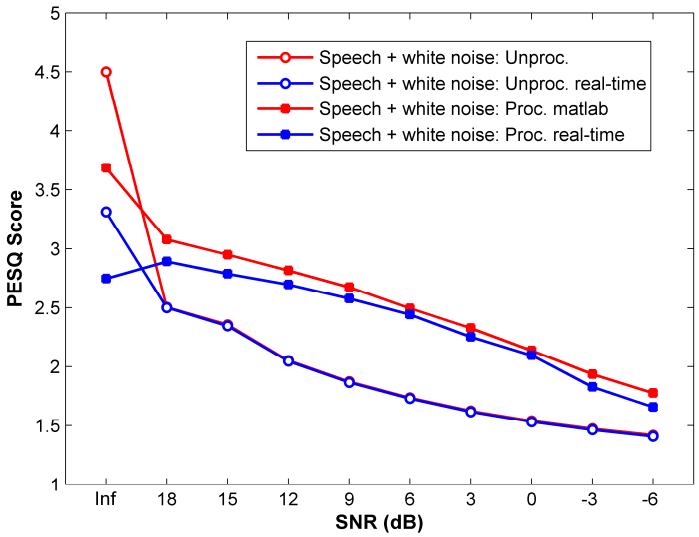


Fig. 4.5 PESQ Score vs SNR for noisy and enhanced speech using offline and real-time processing. Speech: VHSES, noise: white

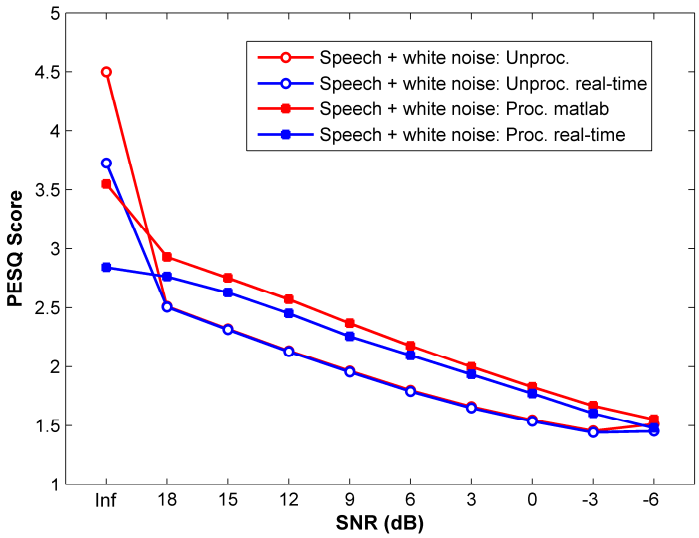


Fig. 4.6 PESQ Score vs SNR for noisy and enhanced speech using offline and real-time processing. Speech: NOIZEUS, noise: white

computation delay beyond the algorithmic delay. The additional delay of 3 ms may be due to the DMA mediated I/O operations.

Chapter 5

SUMMARY AND CONCLUSIONS

Noisy environments have an adverse effect on speech perception by normal as well as hearing impaired person. Presence of noise increases the hearing threshold, and increases spectral masking, leading to degraded speech perception. Further, for persons with sensorineural hearing loss with associated widened auditory filters, and elevated hearing thresholds, speech perception becomes very difficult in noisy environment. For enhancement of noisy speech, a spectral subtraction algorithm was implemented for offline processing and investigations were performed with different types and SNRs of noise. Processing techniques were evaluated using informal listening and objective evaluation using PESQ score.

The investigations were carried out to compare the noise estimation using mean, median and minimum statistics. It was found that median-based noise estimation tracks the noise spectrum well at all the frequencies with different noises at different SNR values. Investigations were also carried out with different window lengths and noise estimation durations. It was found that window length of 20 – 40 ms and noise estimation using the past 81 frames (corresponding to approximately 1.2 s for 30 ms window length) resulted in good PESQ score. Median-based noise estimation (MBNE) is computationally expensive and has large memory requirement. Hence a cascaded-median based noise estimation (CMBNE), was used to get an approximate median, as it involves much less computation and memory. The enhanced noisy speech obtained using MBNE and 3-point 4-stage CMBNE were perceptually similar and have no significant difference in PESQ score. Further, investigations were carried out for studying the effect of the spectral subtraction parameters. The PESQ score was high with α in the range 2 – 2.5 and β = 0.01 for speech materials rich in vowels and voiced sounds and with α in the range 1.2 – 1.4 and β = 0.01 for sentences rich in consonants. To test the hypothesis that the PESQ score may improve if the phase spectrum used for resynthesis is noise free, investigations were carried out using different phase

estimation methods. It was found that use of phase estimated from the enhanced magnitude spectrum did not improve scores as compared with those obtained with the use of noisy phase.

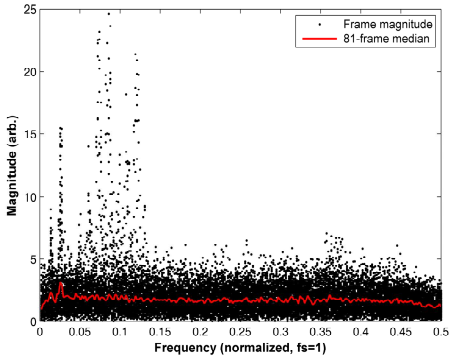
Based on the above investigations, magnitude spectral subtraction with 3-point 4-stage cascaded-median based noise estimation, analysis-synthesis using 30 ms window with 50% overlap, and synthesis using phase spectrum of the noisy speech signal along with the enhanced magnitude spectrum was used. Speech enhancement with our proposed method for noisy speech of 6 dB SNR with different types of additive stationary and non-stationary noises and speech material from NOIZEUS database resulted in improvement of 0.11 – 0.43 in PESQ scores. An examination of the improvements in the scores for noisy speech with SNR of 18 dB down to -6 dB showed that the processing resulted in SNR advantage of 1.5 – 7 dB. The SNR advantage was comparable to or better than those obtained using the enhancement methods available in [10] and tested on the same speech material and types of noise.

For real-time operation, the processing technique was implemented on a DSP board based on 16-bit fixed point processor TMS320C5515 with on-chip FFT hardware. The data transfer and buffering operations were devised for an efficient realization with 50% overlap. Informal listening showed that the processed output from the DSP board was perceptually similar to the corresponding output from the offline implementation. The technique used about one-seventh of computing capacity of the processor and resulted in a signal delay of approximately 48 ms.

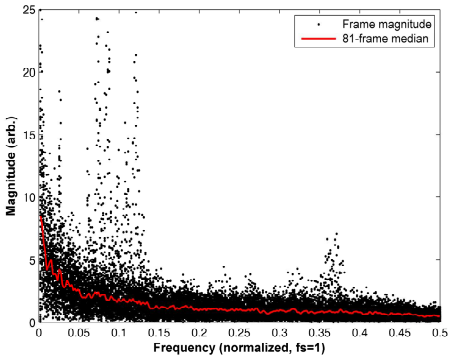
For further improving the performance of the proposed method, use of subtraction and spectral floor factors dependent on frequency and *a posteriori* SNR estimate need to be examined. Also its feasibility for real-time processing needs to be checked. The proposed speech enhancement technique may be combined with other signal processing techniques used in the sensory aids and may be tested for improving perception of different speech materials by the hearing-impaired listeners. The implementation using other processors may also be investigated. Subjective evaluation of intelligibility and quality of enhanced speech needs to be carried out. For subjective evaluation of intelligibility test a GUI has been developed using Matlab and instructions for the test are given in Appendix B.

Appendix A

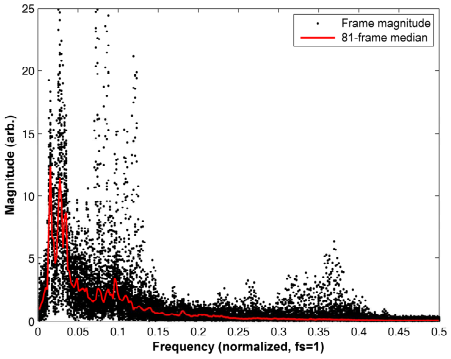
INVESTIGATION ON NOISE ESTIMATION



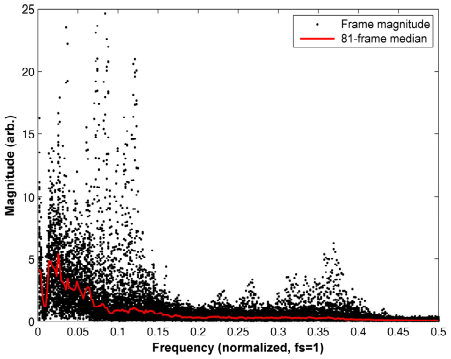
(a) Noise: white



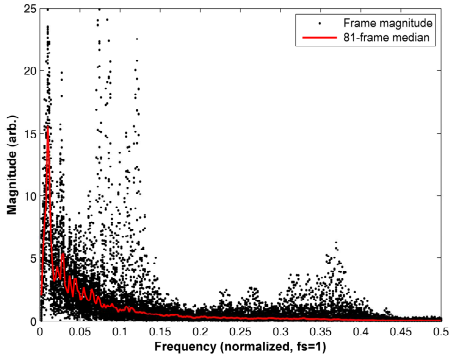
(b) Noise: pink



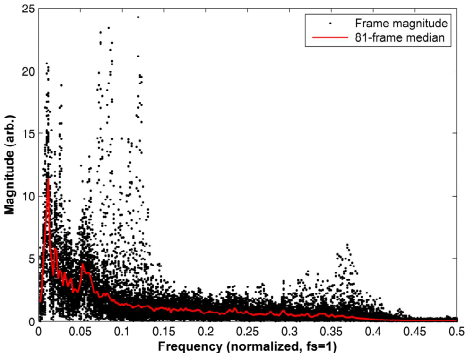
(c) Noise: street



(d) Noise: babble

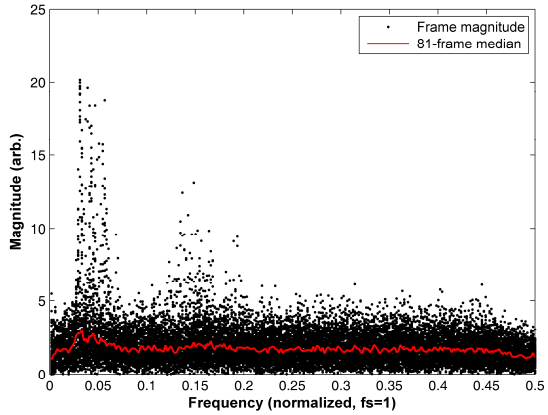


(e) Noise: train

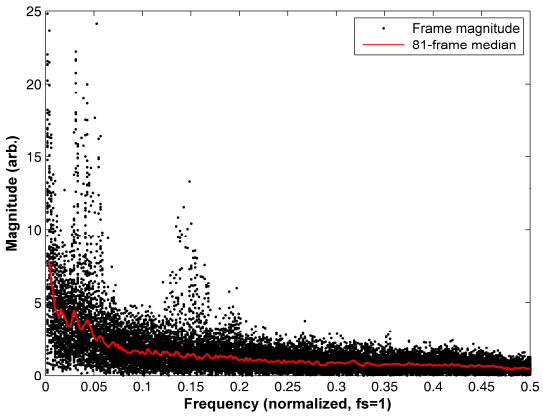


(f) Noise: car

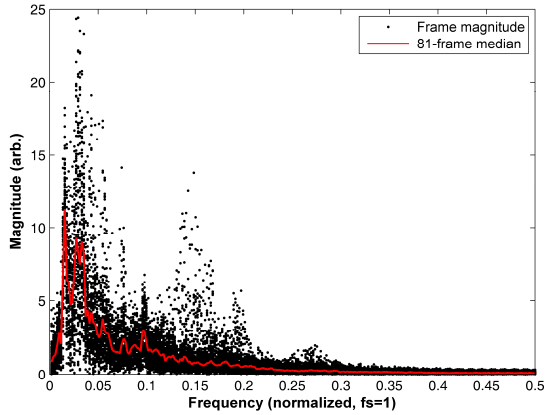
Fig. A.1 Scatter plots of magnitude spectra of noisy speech signals, speech material: VHSES.



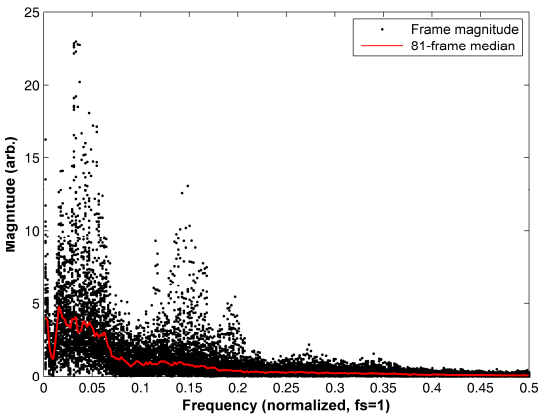
(a) Noise: white



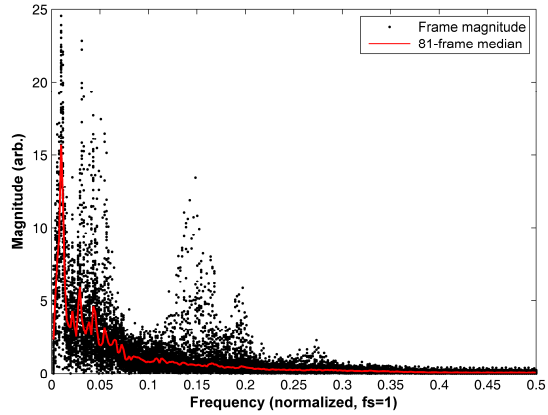
(b) Noise: babble



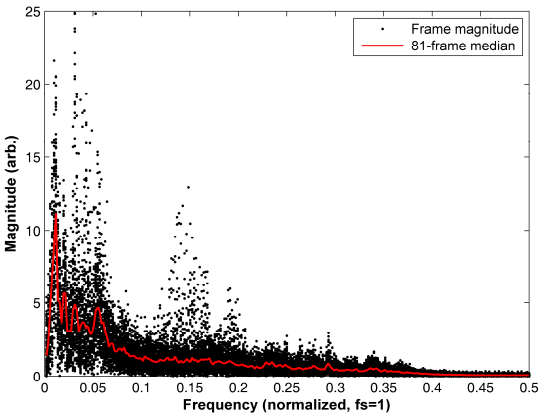
(c) Noise: street



(d) Noise: babble

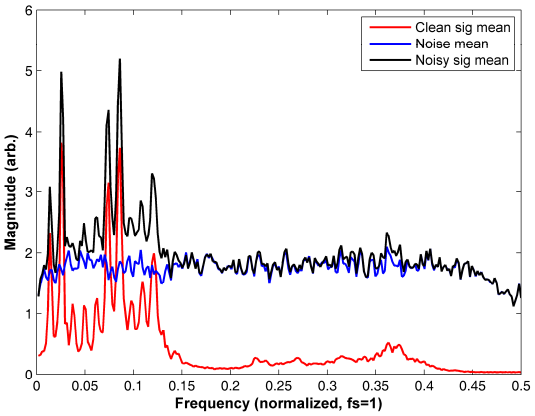


(e) Noise: train

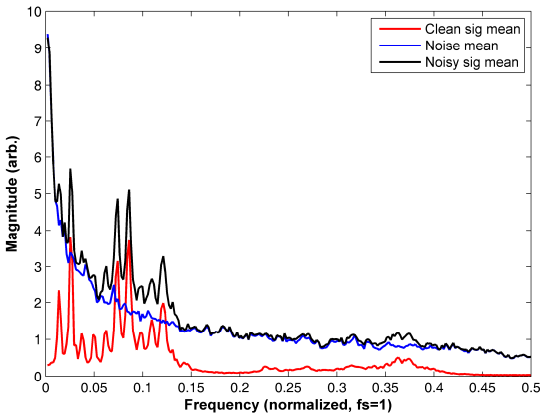


(f) Noise: car

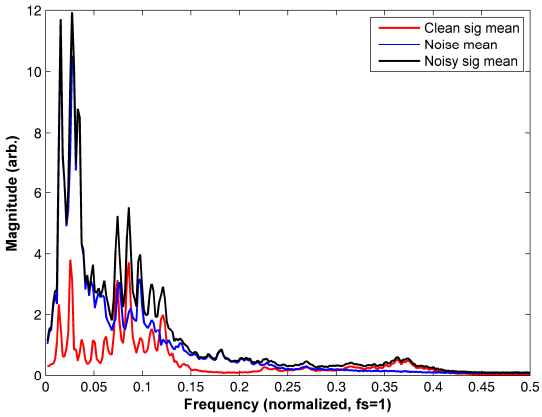
Fig. A.2 Scatter plots of magnitude spectra of noisy speech signals, speech material: NOIZEUS.



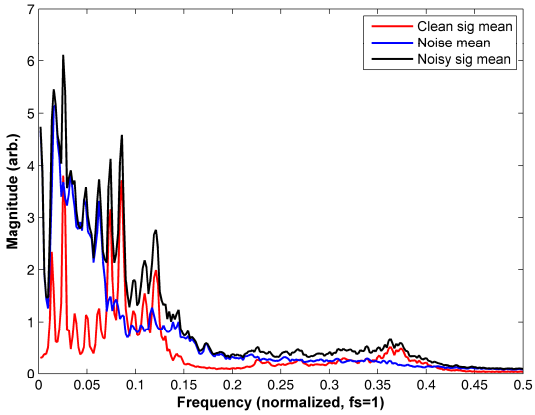
(a) Noise: white



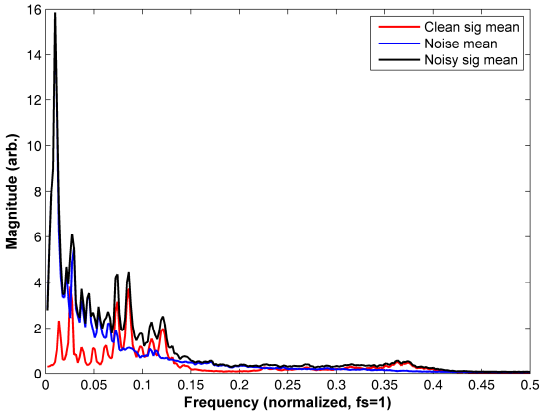
(b) Noise: pink



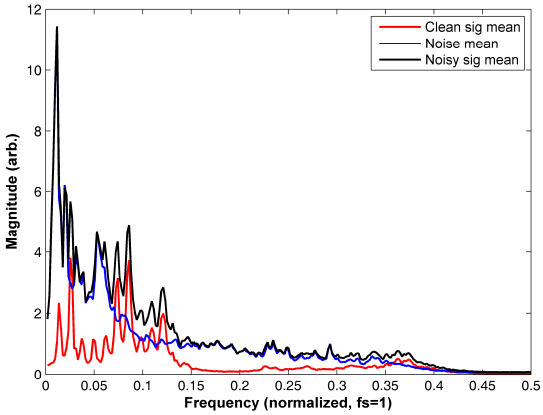
(c) Noise: street



(d) Noise: babble

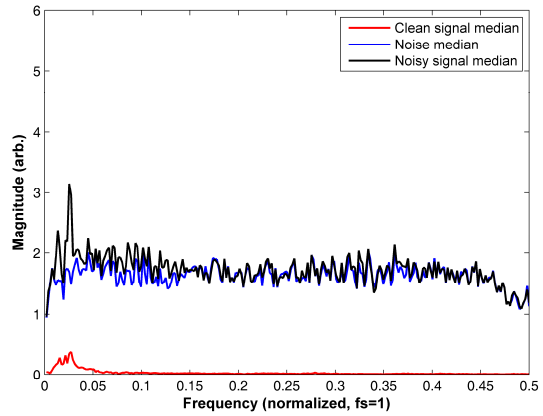


(e) Train noise

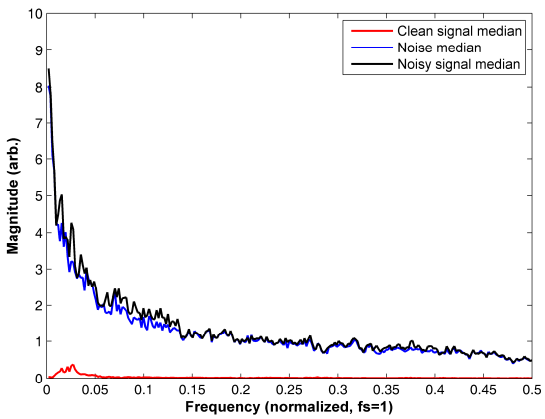


(f) Car noise

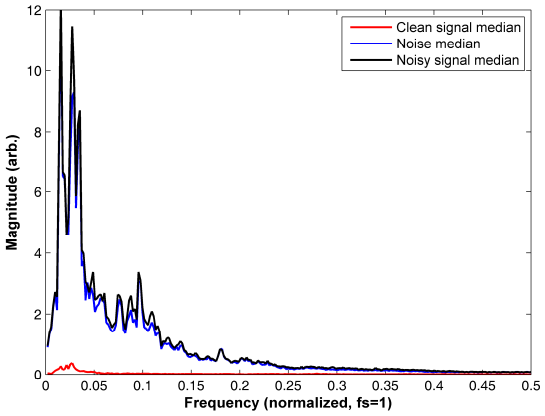
Fig. A.3 Mean of magnitude spectra of clean speech signal, noise and noisy speech (SNR: 0 dB), speech material: VHSES



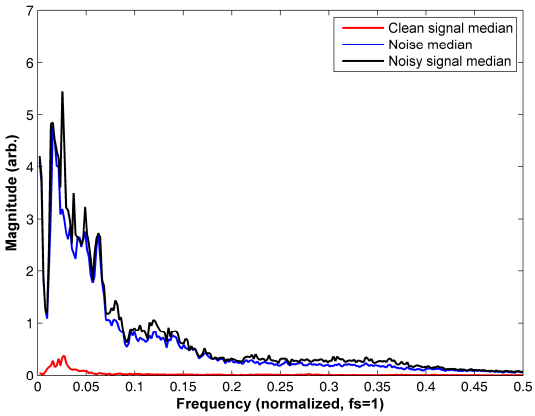
(a) Noise: white



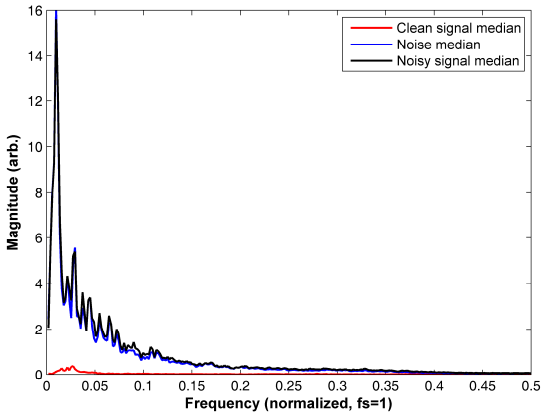
(b) Noise: pink



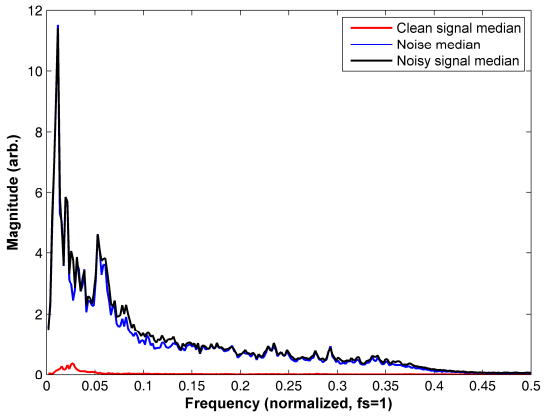
(c) Noise: street



(d) Noise: babble

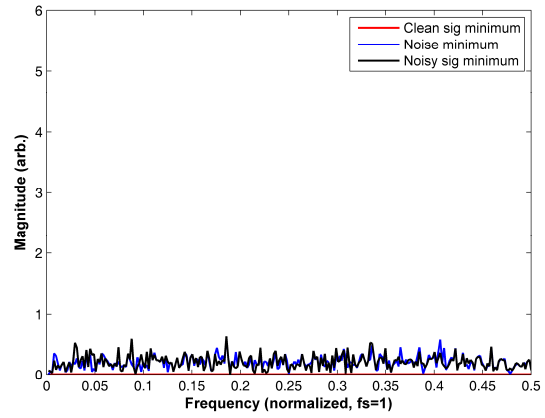


(e) Noise: train

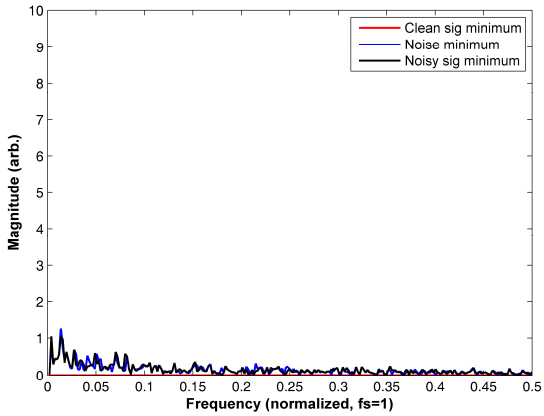


(f) Noise: car

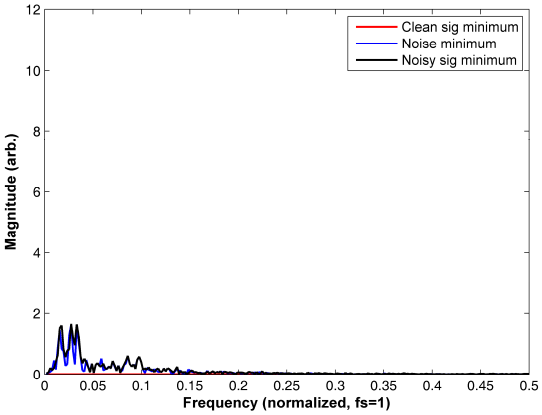
Fig. A.4 Median of magnitude spectra of clean speech signal, noise and noisy speech (SNR: 0 dB), speech material: VHSES



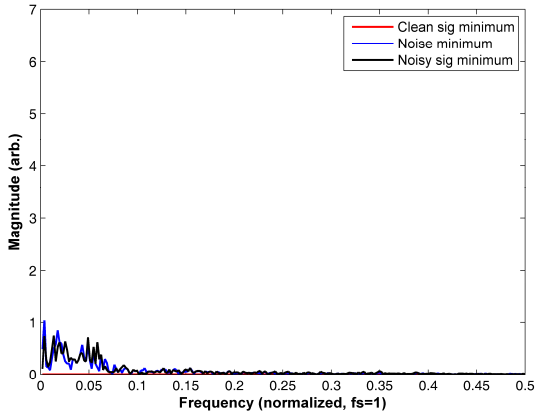
(a) Noise: white



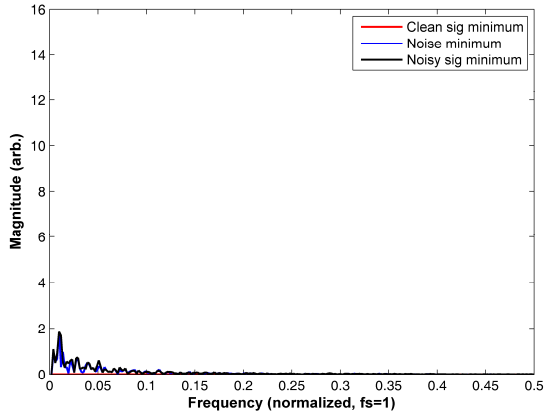
(b) Noise: pink



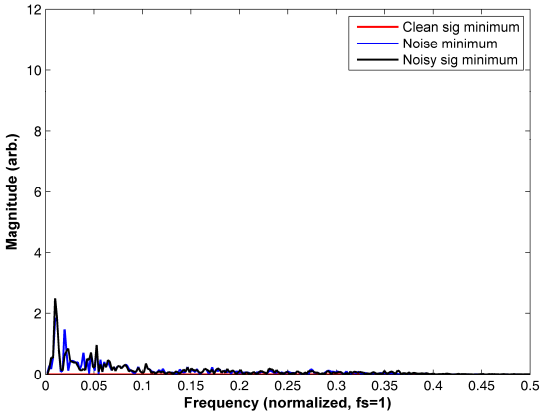
(c) Noise: street



(d) Noise: babble

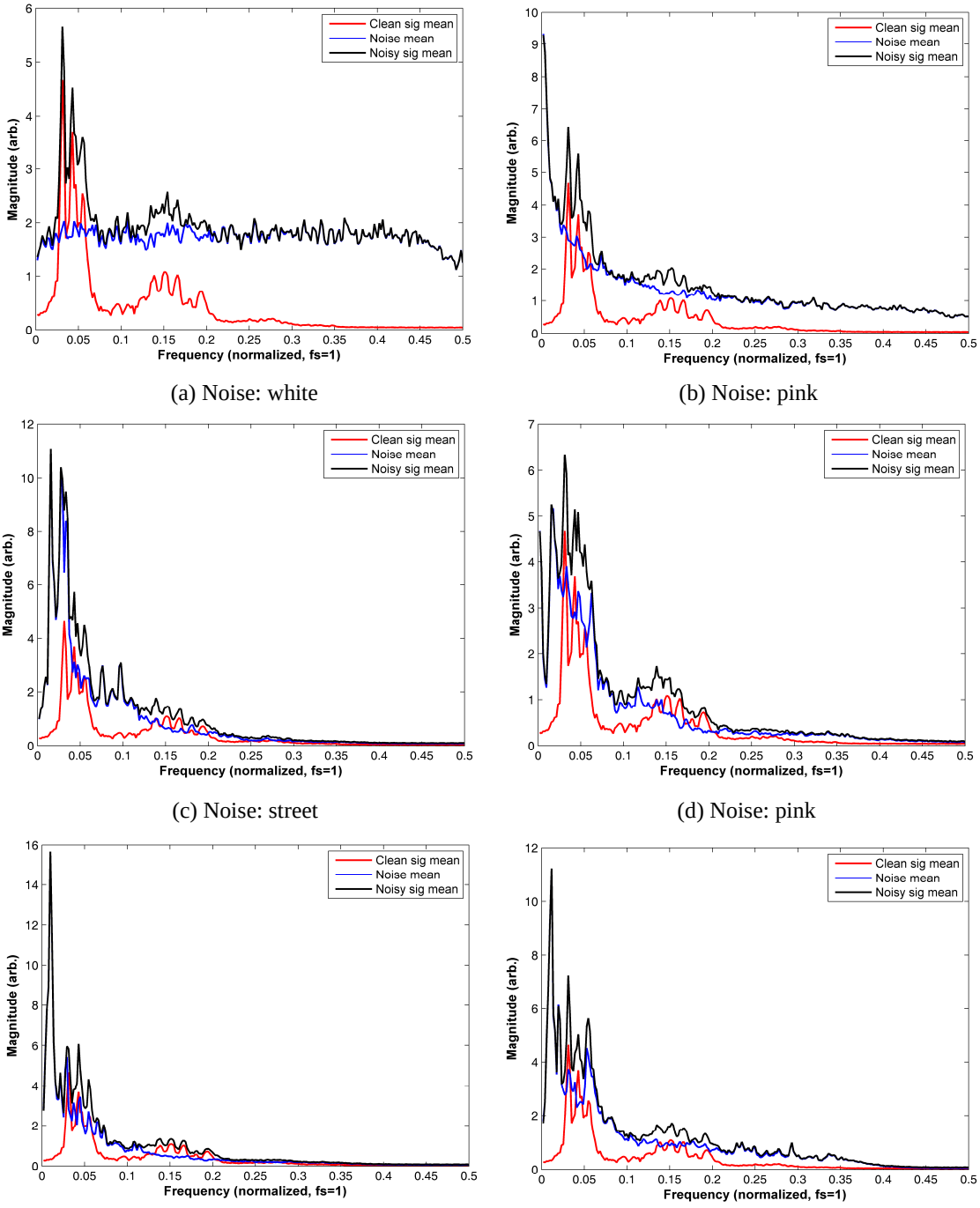


(e) Noise: train



(f) Noise: car

Fig. A.5 Minimum of magnitude spectra of clean speech signal, noise and noisy speech (SNR: 0 dB), speech material: VHSES



(a) Noise: white

(b) Noise: pink

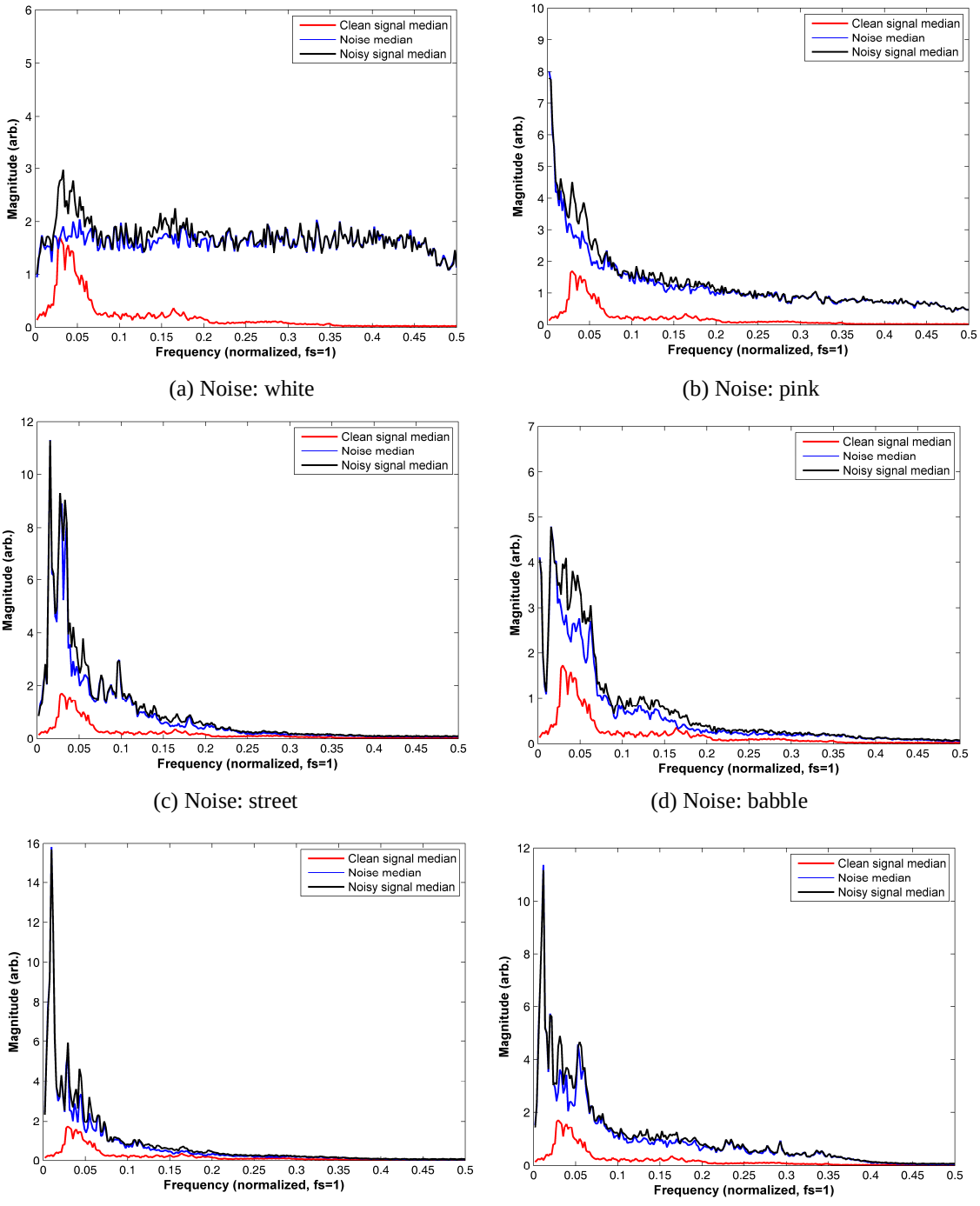
(c) Noise: street

(d) Noise: pink

(e) Noise: train

(f) Noise: car

Fig. A.6 Mean of magnitude spectra of clean speech signal, noise and noisy speech (SNR: 0 dB), speech material: NOIZEUS



(a) Noise: white

(b) Noise: pink

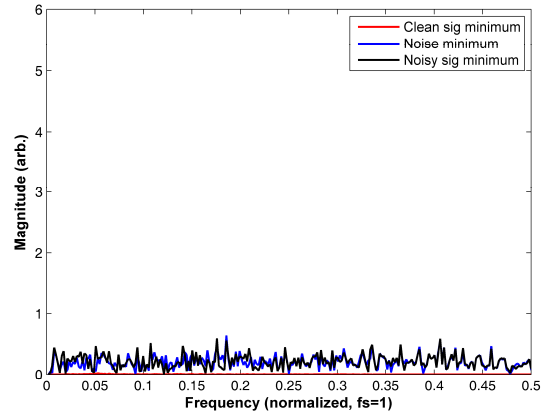
(c) Noise: street

(d) Noise: babble

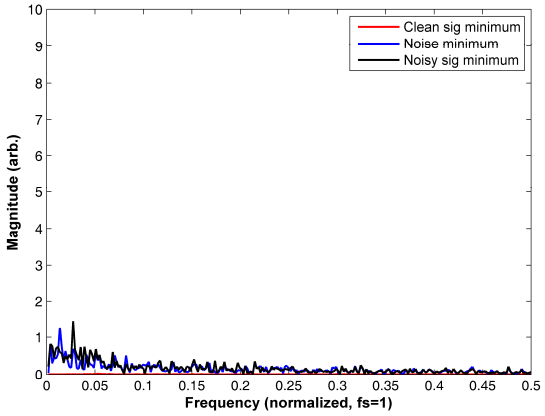
(e) Noise: train

(f) Noise: car

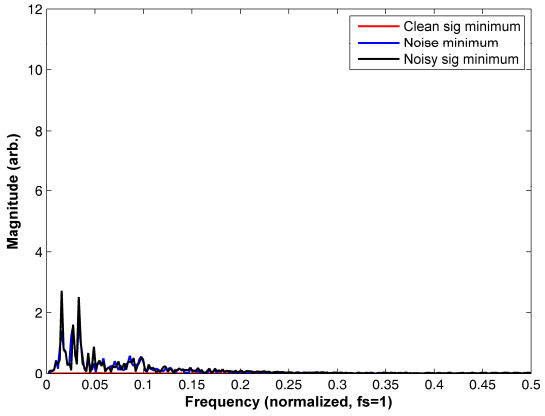
Fig. A.7 Median of magnitude spectra of clean speech signal, noise and noisy speech (SNR: 0 dB), speech material: NOIZEUS



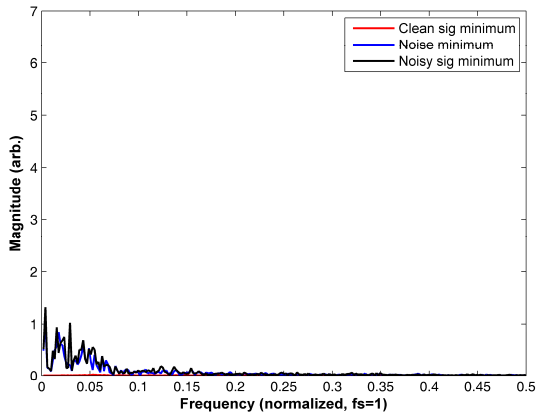
(a) Noise: white



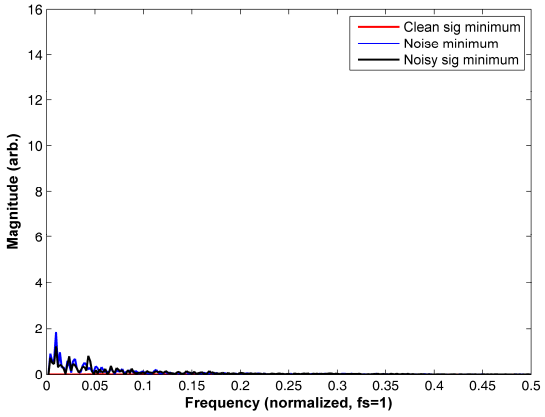
(b) Noise: pink



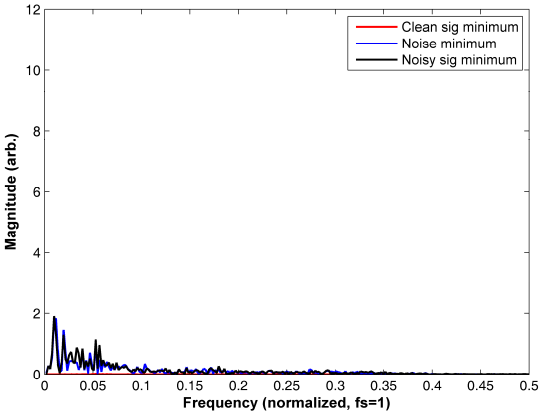
(c) Noise: street



(d) Noise: babble



(e) Noise: train



(f) Noise: car

Fig. A.8 Minimum of magnitude spectra of clean speech signal, noise and noisy speech (SNR: 0 dB), speech material: NOIZEUS

Appendix B

INTELLIGIBILITY TEST

This is a listening test involving presentation of speech sounds. For each sound presented, you need to identify the corresponding word from a list of six words. You will be seated in front of a computer monitor and the test will be conducted by presenting the sounds and recording your responses. The sounds will be presented using a speaker or a pair of headphones, with the volume level of the sounds adjusted to the most comfortable level for you. The test starts with opening of a window on the screen. Once the window is open, you will be able to access the 'Test id' pop-up menu, 'Listening test', 'Practice' and 'Close' buttons.

Practice session is to familiarize you to the words of the test. Clicking on 'Close' will terminate the practice session. On clicking 'Practice', the screen will show 30 words in a panel, a vertical scrollbar and a 'Close' button. Using scrollbar you can see all the words. On clicking on any word you will be able to listen it.

The experiment consists of a set of six tests. In each test, there will be 25 presentations. Before starting the test you need to select the test id from the 'Test id' pop-up menu in MRT window. On clicking 'Listening test', test will be started. During the test, the screen shows the current presentation number, the total number of presentations and test id. There are nine buttons marked as 'Play', 'Next', six response buttons and 'Close' button. After 'Play' is clicked, "Would you write <test word>", sentence will be played. You need to select the most appropriate word from the six words shown in the panel. The response buttons appear inactive until the sounds have been presented. You can indicate your response by clicking on one of the six responses depending on which one is perceived to be more appropriate, or you may listen to the sound again by clicking on 'Play'. After the response, 'Next' and 'Close' buttons become active. Once you are sure of your response, click on 'Next' for the next presentation. Clicking on 'Close' will terminate the test. The sequence of

presentations will be continued until the display show “Test <test_id> is over”. Next test can be continued by selecting test id in MRT window. Clicking on ‘Listening test’, will start the test.

REFERENCES

[1] H. Levitt, J. M. Pickett, and R. A. Houde, Eds., *Senosry Aids for the Hearing Impaired.* New York: IEEE Press, 1980, pp. 3–10.

[2] J. M. Pickett, *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology*. Boston, Mass.: Allyn Bacon, 1999, pp. 289–323.

[3] H. Dillon, *Hearing Aids*. New York: Thieme Medical, 2001.

[4] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, London, UK: Academic, 1997, pp 66–107.

[5] T. Lunner, S. Arlinger, and J. Hellgren, "8-channel digital filter bank for hearing aid use: preliminary results in monaural, diotic, and dichotic modes," *Scand. Audiol. Suppl.*, vol. 38, pp. 75–81, 1993.

[6] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, "Binaural dichotic presentation to reduce the effects of spectral masking in moderate bilateral sensorineural hearing loss," *Int. J. Audiol.*, vol. 51, no. 4, pp. 334–344, 2012.

[7] J. Yang, F. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Commun.*, vol. 39, no. 1–2, pp. 33–46, 2003.

[8] T. Arai, K. Yasu, and N. Hodoshima, "Effective speech processing for various impaired listeners," in *Proc. 18th Int. Cong. Acoust. (ICA 2004)*, Kyoto, Japan, 2004 pp. 1389–1392.

[9] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, "Multi-band frequency compression for improving speech perception by listeners with moderate sensorineural hearing loss," *Speech Commun.*, vol. 54, no. 3, pp. 341–350, 2012.

[10] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York: CRC, 2007.

[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP* 1979, Washington, DC, pp. 208–211.

[12] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing

speech corrupted by colored noise," in *Proc. IEEE ICASSP*, 2002, Orlando, Florida, vol. 4, pp. IV–4164.

[13] Y. Lu and P. C. Loizou, " A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453–466, 2008.

[14] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[15] 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.

[16] L. Lin, E. Ambikairajah and W. H. Holmes, "Auditory filter bank design using masking curves," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 411–414.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp.113–116, 2002.

[19] P. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum," *IEEE Trans. Speech Audio Process*., vol. 13, no. 5, pp. 857–869, 2005.

[20] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Process. Conf.*, 1994, pp. 1182-1185.

[21] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. 4th Eur. Conf. Speech Commun. and Technology (EUROSPEECH'95)*, Madrid, Spain, 1995, pp. 1513–1516.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.

33, no. 2, pp. 443-445, 1985

[23] L. Lin, W.H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, no. 9, pp.754–755, 2003.

[24] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE ICASSP*, 1995, Detroit, MI, pp. 153–156.

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

[26] V. Stahl, A. Fisher, and R. Bipus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE ICASSP*, 2000, Istanbul, Turkey, pp. 1875–1878.

[27] C. Ris and S. Dupont, "Assessing local noise level estimation methods: application to noise robust ASR," *Speech Commun.*, vol. 34, no. 1-2, pp. 141–158, 2001.

[28] S. K. Basha and P. C. Pandey, "Real-time enhancement of electrolaryngeal speech by spectral subtraction," in *Proc. Nat. Conf. on Commun. (NCC 2012)*, Kharagpur, India, 2012, pp. 516–520.

[29] 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.

[30] 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.

[31] Y. Hu and P. C. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[32] 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. (ICSLP 2000)*, Beijing, China, 2000, pp. 29–32.

[33] T. F. Quatieri, and A. V. Oppenheim, “Iterative techniques for minimum phase signal reconstruction from phase or magnitude,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1187–1193, 1981.

[34] S. H. Nawab, T. F. Quatieri, and J. S. Lim, “Signal-reconstruction from short time Fourier transform magnitude,” *IEEE Trans. Acoust., Speech Signal Process.*, vol. 31, no. 4, pp. 986–998, 1983.

[35] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, New Jersey: Prentice Hall, 1978, pp. 356–362.

[36] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[37] Spectrum Digital, Inc. (2010) TMS320C5515 eZdsp USB Stick Technical Reference. [online]. Available: support.spectrumdigital.com/boards/usbstk5515/reva/files/usbstk5515_TechRef_RevA.pdf

[38] Texas Instruments, Inc. (2011) TMS320C5515 Fixed-Point Digital Signal Processor. [online]. Available: focus.ti.com/lit/ds/symlink/tms320c5515.pdf.

[39] Texas Instruments, Inc. (2008) TLV320AIC3204 Ultra Low Power Stereo Audio Codec. [online]. Available: focus.ti.com/lit/ds/symlink/tlv320aic3204.pdf.

Acknowledgements

I express my sincere gratitude towards my respected guide Prof. P. C. Pandey, for his invaluable guidance and support he gave during this project. He introduced me to the world of speech processing, encouraged me to take up new challenges and inspired me all through my stay here.

I am thankful to Khadar Basha and Nitya for guiding me in project implementation and sharing interesting discussions with me. I would like to thank Vidyadhar Kamble for helping me in all the lab related issues, and Lehana, Nataraj, Rajath, and Jagbandhu for their helpful advices in the different issues of the project. I am thankful to Dr. Panduranga Kulkarni and Jayan, for their help in reviewing my paper draft for NCC2013. I am also thankful to all my friends especially in the SPI lab and EI Lab for their whole-hearted support during the tenure of this project.

Finally, I am thankful to my parents and family members for their unconditional love, encouragement and support.

Santosh Kumar Waddi

June 2013

Author’s Resume

Santosh Kumar Waddi: The author received the B. Tech. degree in electronics and communication engineering from the Gayatri Vidya Parishad College of Engineering, Visakhapatnam (Andhra Pradesh), affiliated to JNTU Kakinada in 2010. Presently, he is pursuing the M. Tech. degree in electrical engineering at the Indian Institute of Technology Bombay. His research interests include speech processing, digital signal processing, and embedded system design.

Thesis related publication

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.