DCT-Based Spectral Subtraction for Speech Enhancement

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

Master of Technology
in
Communication and Signal Processing

by

Guntaka Madhu Lekha
(11D070063)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology Bombay
June 2016

This page is intentionally left blank

Indian Institute of Technology Bombay

M. Tech. Dissertation Approval

This dissertation entitled **"DCT-Based Spectral Subtraction for Speech Enhancement"** by **Guntaka Madhu Lekha** (Roll No. **11D070063**) is approved, after the successful completion of *viva voce* examination, for the award of the degree of **Master of Technology** in **Communication and Signal Processing.**

Supervisor ..... (Prof. P. C. Pandey)

Examiners ..... (Prof. M. S. Shah)

..... (Prof. Preeti Rao)

Chairperson ..... (Prof. Preeti Rao)

Date: 30th June 2016

Place: Mumbai

This page is intentionally left blank

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.

_____

(**Guntaka Madhu Lekha**)

Date: 30 June 2016

Place: Mumbai

This page is intentionally left blank

G. Madhu Lekha / Prof. P. C. Pandey (Supervisor): “DCT-Based Spectral Subtraction for Speech Enhancement”, *M.Tech. Dissertation*, Department of Electrical Engineering, Indian Institute of Technology Bombay, June 2016.

ABSTRACT

Spectral subtraction for speech enhancement has been widely investigated, using analysis and synthesis based on discrete Fourier transform and is reported to perform well for suppression of stationary noises. This thesis presents investigations on implementation of spectral subtraction using the discrete cosine transform (DCT) and its real-time implementation. DCT is considered suitable for this application because of its superior energy compaction and binary phase representation. The noise is estimated using a 3-frame 4-stage cascaded-median without involving a voice activity detector. The qualitative results are compared to that of the same technique used with DFT. Real-time implementation is done on a DSP board with the 16-bit fixed point processor TMS320C5515. The on-board codec is used to continuously acquire the input signal and output the processed signal at a sampling rate of 10kHz. DMA is used to facilitate the input and output buffering. A fast cosine transform implementation of DCT is realized and the on-chip FFT hardware is used for performing forward and inverse transformations. The real-time processing is implemented with a 300-point analysis-synthesis window and 512-point FFT. The implementation uses about 1/6th of the computing capacity, and the processing delay is approximately 49 ms, making it acceptable for hearing aid applications.

This page is intentionally left blank

CONTENTS

Abstract i

List of abbreviations and symbols v

List of figures vii

List of tables ix

Chapters

1. Introduction 1

1.1 Problem overview 1

1.2 Project objective 2

1.3 Dissertation outline 2

2. DCT basics 3

2.1 Introduction 3

2.2 Fast cosine transform 3

2.3 Comparison between DCT and DFT 5

3. Spectral subtraction for speech enhancement 7

3.1 Generalized spectral subtraction 7

3.2 Noise estimation for spectral subtraction 9

3.2.1 Histogram-based techniques 9

3.2.2 Quantile-based noise estimation (QBNE) 9

3.2.3 Cascaded-median based noise estimation (CMBNE) 10

3.2.4 Dynamic quantile tracking based noise estimation (DQTBNE) 10

4. Investigations using offline implementation 11

4.1 Introduction 11

4.2 Evaluation method and test materials 11

4.3 Investigations on noise estimation 12

4.4 Effect of noise estimation duration and window length 21

4.5 Effect of window type and overlap 21

4.6 Effect of over-subtraction factor 26

4.7 Phase estimation for spectral subtraction 26

4.8 Comparison between DCT and DFT 27

4.9 Conclusion 27

5. Real-time implementation 35

5.1 Introduction 35

5.2 Implementation details 35

5.3 Results 37

6. Summary and conclusion **39**

Appendices

A. Investigations on noise estimation 42

B. Robustness of DCT to phase modifications 47

References 50

Acknowledgements **53**

List of abbreviations and symbols

Abbreviation/ Symbol	Explanation
AC	alternate current
ADC	analog-to-digital converter
CMBNE	cascaded-median based noise estimation
CPU	central processing unit
D	spectrum of estimated noise magnitude
DAC	digital-to-analog converter
DCT	discrete cosine transform
DFT	discrete Fourier transform
DMA	direct memory access
DQTBNE	dynamic quantile tracking based noise estimation
DSP	digital signal processor
FCT	fast cosine transform
FFT	fast Fourier transform
f_s	sampling frequency
I/O	input/ output
IDCT	inverse discrete cosine transform
IDFT	inverse discrete Fourier transform
IFFT	inverse fast Fourier transform
KLT	Karhunen-Loeve transform
L	window length
MBNE	median based noise estimation
MOS	mean opinion score
N	size of the DCT/ DFT
PC	personal computer
PESQ	Perceptual Evaluation of Speech Quality
QBNE	quantile based noise estimation
RMS	root mean square
S	window shift
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
TI	Texas Instruments
USB	Universal Serial Bus
$w(n)$	analysis-synthesis window

W_N^{kn}	complex exponential, $e^{\frac{-j2\pi kn}{N}}$
$x(n)$	real-valued discrete-time signal
X	spectrum of the unprocessed speech signal
$y(n)$	resynthesized real-valued discrete-time signal
Y	spectrum of the resynthesized speech signal
α	over subtraction factor
β	spectral floor factor
γ	exponent factor

List of Figures

Figure	Caption	Page
2.1	Example of even extensions of a sequence: (a) Original sequence $x(n)$, $0 \leq n \leq N-1$, (b) Periodic $2N$- sample even extension of $x(n)$, (c) $y(n)$, A $2N$-sample even extension of $x(n)$, (d) Division of $y(n)$ into its $v(n)$ and $w(n)$.	4
3.1	Spectral subtraction using DFT or DCT	8
3.2	A p-point q-stage cascaded-median based noise estimation	10
4.1	Mean, median, 0.25 quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (babble 0db) : (a) VHSES (b) NOIZEUS	14
4.2	Spectrograms of noises estimated as various quantiles of the noisy speech material (babble-0db) bit "aayiye aapka naam kya hai?", (a)−(f)	15
4.2	Spectrograms of noises estimated as various quantiles of the noisy speech material (babble-0db) bit "aayiye aapka naam kya hai?", (g)−(l)	16
4.3	PESQ scores for the enhanced noisy speech, Speech material : VHSES at SNR 0 dB with babble ($\alpha = 3$), car, train ($\alpha = 2.5$), pink, white ($\alpha = 3.5$), and street ($\alpha = 2$) noises. $\beta = 0.001$, $\gamma = 1$.	22
4.4	PESQ scores for the enhanced noisy speech, Speech material : NOIZEUS at SNR 0 dB with babble, street, train ($\alpha = 1.5$), pink, car and white ($\alpha = 2.5$) noises. $\beta = 0.001$, $\gamma = 1$.	23
4.5	PESQ scores vs SNR for the enhanced noisy speech, DCT (β=0.01, γ=1), DFT (β=0.01, γ=1)Speech material : VHSES various SNRs with babble, car, pink, street, train and white noises.	31
4.6	PESQ scores for the enhanced noisy speech, DCT ($\beta = 0.001$, $\gamma = 1$), DFT ($\beta = 0.01$, $\gamma = 1$)Speech material : NOIZEUS various SNRs with babble, car, pink, street, train and white noises.	32
5.1	Block diagram of TMS320C5515 eZdsp USB Stick	35
5.2	Implementation level diagram of spectral subtraction on DSP board	35
5.3	Working Of DMA for Buffering and Data Transfer Operations ($S=L/2$)	36
5.4	Setup for giving input and recording output from DSP board	37
5.5	Signals and spectrograms: Processing of "Where were you a year ago?," from a male speaker, with white noise at 3 dB SNR.	37

A.1 Mean, median, 0.25 quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (car 0db) 41
A.2 Mean, median, 0.25 quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (pink 0db) 42
A.3 Mean, median, 0.25 quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (street 0db) 43
A.4 Mean, median, 0.25 quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (train 0db) 44
A.5 Mean, median, 0.25 quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (white 0db) 45

List of Tables

Figure	Caption	Page
4.1	Processing steps and investigations	11
4.2	(a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Babble (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Babble	17
4.3	(a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Car (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Car	17
4.4	(a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Pink (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Pink	18
4.5	(a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Street (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Street	18
4.6	(a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Train (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : Train	19

4.7	(a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : White (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction (β = 0, γ = 1), Speech material : VHSES, Noise : White	19
4.8	PESQ scores obtained using spectral subtraction and DQTBNE. $\alpha = 2$, $\beta = 0$ and $\gamma = 1$, Speech material : VHSES with various noises	20
4.9	PESQ scores for enhanced speech using CMBNE and spectral subtraction with different types of windowing functions and for various noises at 0 dB. γ = 1; β = 0.01, α given in brackets corresponding to the scores. Speech material : VHSES	24
4.10	PESQ scores for enhanced speech using CMBNE and spectral subtraction with different types of windowing functions and for various noises at 0 dB. γ = 1; β = 0.001, α given in brackets corresponding to the scores. Speech material : NOIZEUS	25
4.11	PESQ score for enhanced speech using CMBNE and spectral subtraction for different types of noise at 0 dB SNR and γ = 1. Speech material : VHSES	28
4.12	PESQ score for enhanced speech using CMBNE and spectral subtraction, for different types of noise at 0 dB SNR and $\gamma = 1$. Speech material : NOIZEUS	29
4.13	PESQ scores of combinations of magnitude and phase for the speech bit "Where were you a year ago?" from speech material VHSES	30
4.14	(a)SNR advantage obtained using 3-point 4-stage CMBNE and spectral subtraction using DCT and DFT against optimal value of α, β = 0.01, γ = 1, Speech material : VHSES (b)SNR advantage obtained using 3-point 4-stage CMBNE and spectral subtraction using DCT and DFT against optimal value of α, DCT(β = 0.001), DFT(β = 0.01), γ = 1, Speech material : NOIZEUS	33
B.1	PESQ scores showing sensitivity of DCT to phase changes as percentage of inversions per frame. Speech material : VHSES	47
B.2	PESQ scores showing sensitivity of DCT to phase changes with respect to magnitude as percentage of inversions per frame. Speech material : VHSES	49
B.3	Number of phase inversions occurring as compared to that of the clean speech segment "Where were you a year ago?" from material VHSES, for a particular frequency component in the range 27−39.	51

Chapter 1

INTRODUCTION

1.1 Problem overview

Presence of noise in the input speech deteriorates the performance of speech recognition algorithms and the effectiveness of hearing aids. Several noise suppression techniques have been developed for enhancement of speech corrupted with noise for use in hearing aids, cochlear prosthesis or sensory aids. In most of these applications, a single channel input is available and use of additional microphone for providing a reference input for adaptive noise cancellation is not practical. Also, for real-time operation in hearing aids, the algorithm used must have low complexity and memory requirement. Spectral subtraction algorithm [1], [2] is a single input enhancement method and is computationally efficient. It works on an additive model of noise, which has to be estimated from the noisy signal. A dynamic estimation of noise is required as it is generally non-stationary. Some of the methods developed for noise estimation are minimal tracking algorithms, time-recursive averaging algorithms, histogram based techniques and quantile-based ones [3]−[5]. They differ in computational complexity and memory requirement, which affect their suitability for real-time implementation. Voice activity detection based methods may not work satisfactorily in low SNR conditions and may not correctly track the noise spectrum during long speech segments. So methods not requiring it are preferable. Methods based on order statistics are reported to work well [3]−[5], but they involve sorting operations which increase their complexity and memory requirements.

Choice of the transform used for processing is an important consideration. It is desirable that the coefficients generated are uncorrelated, and the transform is invertible and computationally less intensive [6]. Commonly used transforms are Karhunen-Loeve transform (KLT), discrete Fourier transform (DFT), discrete cosine transform (DCT), and discrete wavelet transform (DWT). KLT is optimal in terms of energy compaction, but depends on input statistics. The main advantage of DWT is its fast implementation because of complexity of order N, but it offers less number of frequency bands, which may lead to insufficient separability of speech and noise. DFT is the most established transform in speech processing and many fast Fourier transform (FFT) algorithms are available for its efficient computation. However, DFT produces complex coefficients and spectral subtraction involves

modifications of the magnitude spectra only. Using the original phase for reconstruction of the enhanced signal limits the enhancement that can be obtained. DCT is well established transform in image processing and compression algorithms, owing to its excellent energy compaction and simpler phase representation. Hence, its application in the field of speech processing is worth investigating.

1.2 Project objective

The objective of this project is to compare performance of DCT with DFT for speech enhancement using spectral subtraction and to verify whether it is suitable for real-time applications. The performance evaluation is carried out through offline implementation using Matlab. This is followed by implementation of DCT-based spectral subtraction on a 16-bit fixed point DSP processor, with an on-chip FFT hardware and thus devising methods for efficient DCT implementation to computations.

1.3 Dissertation outline

Chapter 2 contains the basics of DCT and a comparison between DCT and DFT. Chapter 3 presents a discussion on spectral subtraction and noise estimation. Chapter 4 describes the investigations on Matlab based offline implementation. The real-time implementation aspects using 16-bit fixed point processor TMS320C5515 are described in Chapter 5. The last chapter provides summary and conclusion of the work.

Chapter 2

DCT BASICS

2.1 Introduction

The DCT of a sequence expresses it in terms of sum of cosine functions at different frequencies. Mathematically, DCT is equivalent to the DFT of even extensions of the signal. Fig. 2.1 shows an example of even extensions of a *N*-sample sequence, wherein Fig. 2.1(b) shows an even periodic extension of *x*(n) given in Fig. 2.1(a) and Fig. 2.1(c) shows *y*(n) as 2*N*-point even extension of *x*(n). Depending upon the type of extension, there are eight standard DCT variants [7], out of which four are commonly used. For a sequence *abcd*, even extensions *abcdcba*, *abcddcba*, *dcbabcd*, and *dcbaabcd* can form a basis for DCT. DCT-II and DCT-III correspond to the extension *abcddcba*. These transforms satisfy the property of being inverses of each other and are used as the forward and inverse transforms in this thesis, respectively. These are given as the following :

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right),\ 0 \le k \le N-1 \quad (1)$$

$$x(n) = \frac{1}{2}X(0) + \sum_{k=1}^{N-1} X(k) \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right),\ 0 \le n \le N-1 \quad (2)$$

2.2 Fast cosine transform

The direct application of DCT and IDCT formulae for *N*-sample sequence would require O(N^2) operations, but by factorizing the computation similarly to the fast Fourier transform (FFT), it is possible to reduce the complexity. Such reduced O(N log N) methods to compute DCTs are known as fast cosine transform (FCT) algorithms. Though there are algorithms that directly specialize in optimizing DCT, calculation of the transform in terms of DFT is useful for implementation on a DSP platform with FFT hardware.

DCT of an *N*-sample sequence $x(n)$ can be computed by taking the *2N*-point DFT of $y(n)$ shown in Fig. 2.1(c) and multiplying it by $W_{2N}^{k/2}$, where the complex exponential $W_N^{kn} = e^{\frac{-j2\pi kn}{N}}$. Alternatively, it can be computed by taking *2N*-point DFT of the original sequence with *N* zeros padded to it, multiplying it by $W_{2N}^{k/2}$ and then taking twice the real part [8]. Similarly IDCT can be calculated by using a *2N*-point IDFT.

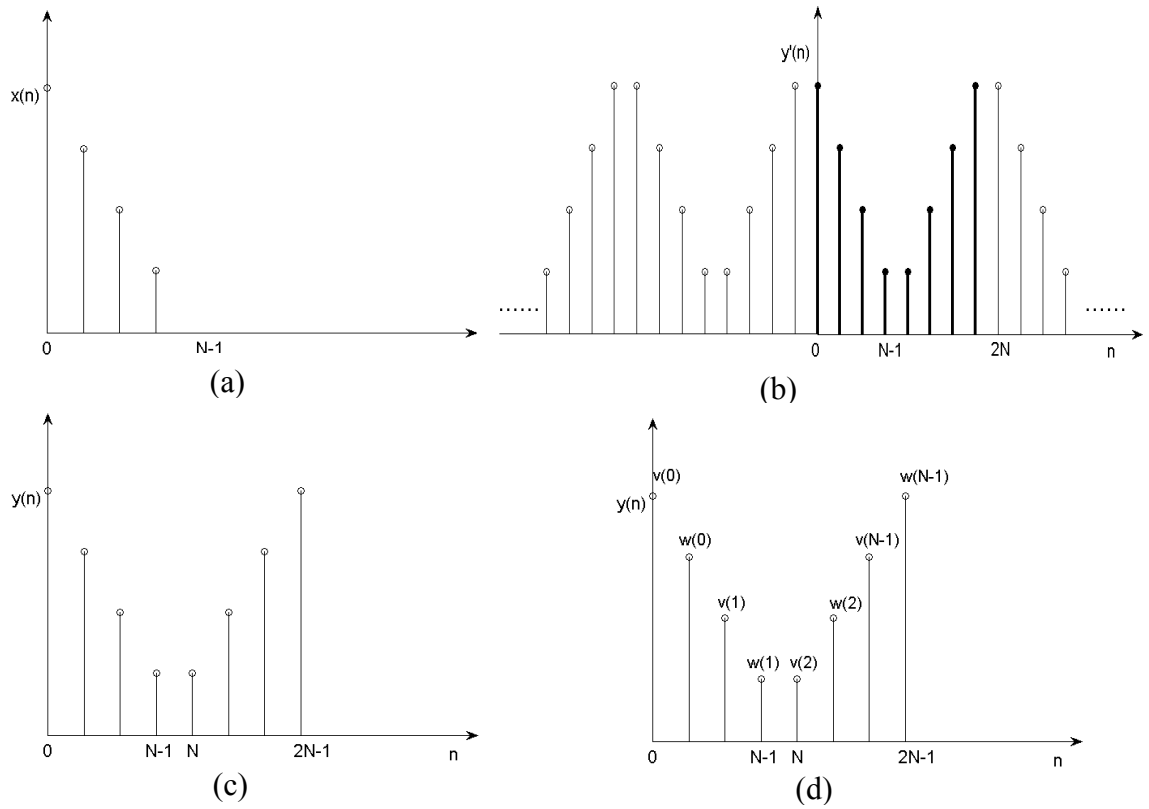


Figure 2.1. Example of even extensions of a sequence: (a) Original sequence $x(n)$, $0 \le n \le N-1$, (b) Periodic $2N$- sample even extension of $x(n)$, (c) $y(n)$, A $2N$- sample even extension of $x(n)$, (d) Division of $y(n)$ into its $v(n)$ and $w(n)$.

Makhoul [8] proposed an algorithm that can be used to obtain an N-point DCT via an N-point DFT. This algorithm is of interest owing to the reduced complexity it offers for a real-time implementation of DCT. It involves dividing the extension of the signal, i.e. y(n) into N-point even and odd sequences, v(n) and w(n) as shown in Fig. 2.1(d) and is given as

$$v(n) = y(2n);\ w(n) = y(2n+1),\ 0 \le n \le N-1 \tag{3}$$

Applying the properties of signal periodicity and symmetry, sequence v(n) can be expressed in terms of x(n) as

$$v(n) = \begin{cases} x(2n), & 0 \le n \le \left\lfloor \frac{N-1}{2} \right\rfloor \\ x(2N-2n-1), & \left\lfloor \frac{N-1}{2} \right\rfloor \le n \le N-1 \end{cases} \tag{4}$$

where $\lfloor p \rfloor$ stands for the integral part of p. The N- point DFT of this sequence v(n) is calculated as V(k), $0 \le k \le N$-1. It is multiplied by $2e^{\frac{-j\pi k}{2N}}$ and the real part, gives the DCT $C_x(k)$ of $x(n)$:

$$C_x(k) = 2\mathrm{Re}\left(W_{4N}^k \sum_{n=0}^{N-1} v(n) W_N^{nk}\right), \quad 0 \le k \le N-1 \tag{5}$$

For computing x(n) as IDCT of a given forward transform $C_x(k), 0 \le k \le N-1$, V(k) is computed as the following

$$V(k) = \frac{1}{2} W_{4N}^{-k} \left(C_x(k) - j\, C_x(N-k) \right),\ 0 \le k \le N-1 \tag{6}$$

with $C_x(N) = 0$. IDFT of $V(k)$ results in a real sequence $v(n)$ and is used to get $x(n)$, using the relation in (4). With use of FFT, the complexity of this algorithm for computing DCT is brought down from $O(N^2)$ to $O(N\log N)$ with $O(N)$ pre- and post-processing steps involving multiplications.

2.3 Comparison of DCT and DFT

The suitability of DCT and DFT for spectral subtraction is compared on the basis of phase, energy compaction, and some miscellaneous factors.

Phase: Noise addition disturbs both magnitude and phase spectra of the signal. Most of the speech enhancement operations enhance the magnitude spectrum. The phase spectrum of the noisy speech is taken as the best estimate of phase spectrum of the clean signal, and it is used to reconstruct the enhanced speech spectrum after modification of the magnitude spectrum [9]. If the phase spectrum is set to zero and speech is resynthesized, the speech sounds voiced with a constant pitch. If it is replaced by a random phase between $-\pi$ to $+\pi$, the resynthesized speech sounds unvoiced. Vary [10] has reported that random phase deviation upto a threshold of $\pi/8$ can be acceptable and that the speech sounds rough after that. The DCT coefficients have binary phase of 0 and π. In regions with significant amount of signal energy, i.e. the high SNR blocks, noise will be weak to change the sign on the coefficient, so its effect would be confined to magnitude. If the noise dominates so much that it changes the sign resulting in an erroneous phase, the subtraction algorithm will counteract it with highly modifying the magnitude, thus significantly reducing the effect of erroneous phase. Thus DCT is expected to give a better phase performance [6].

Energy Compaction: The efficiency of a transform is highly gauged by the energy compaction it offers. If the information is present in fewer coefficients, the other components can be discarded or attenuated. As shown in [11], energy compaction provided by DCT is nearly as much as that of the optimal KLT and is definitely higher than DFT. A demonstration has been given in [6]. A clean speech is divided into frames using a rectangular window with 50% overlap and the transforms are performed. The magnitude of resulting coefficients are arranged in decreasing order and l lowest valued coefficients are set to zero. The magnitude and phase of the resulting coefficients are used to reconstruct the speech using weighted overlap-add technique. A plot of square error vs l, given in Fig. 2.2., shows that DCT provides better energy compaction. Thus, use of DCT is justified for spectral subtraction as noise suppression is achieved by attenuating the transform coefficients.

Resolution and edge effect: For N-sample sequence, DCT has N independent spectral coefficients whereas DFT has $N/2+1$ independent coefficients (for an even N) due to the property of conjugate symmetry. Further, the boundary effects (effects due to block

discontinuities) in DCT are less predominant than in DFT because of the smooth transitions at the edges generated as a result of even extension [11].

Based on these considerations, we can expect DCT to perform better than DFT for spectral subtraction.

Chapter 3

SPECTRAL SUBTRACTION FOR SPEECH ENHANCEMENT

3.1 Spectral subtraction

Spectral subtraction is based on the additive model of speech and noise, along with the assumption that input signal and noise are uncorrelated, which is valid only in a quasi-statistical sense. The estimate of clean speech is obtained by the subtraction of estimated noise from the input noisy signal. Thus, the processing requires noise estimation, which can be done by using a voice activity detector or using statistical methods.

A block diagram of the spectral subtraction in transform domain is shown in Fig. 3.1, where DFT or DCT is used for spectrum calculation and corresponding inverse transform is used for resynthesis. The input signal is segmented by overlapping windows and the spectrum is calculated for each windowed segment. This is followed by magnitude and phase calculation. The phase will be in the range of [0, 2π] in case of DFT, whereas the binary value of 0 or π for DCT. Noise magnitude spectrum $D_n(k)$ is estimated using the statistics of the magnitude spectra of previous frames and it is subtracted from the magnitude value of the transform of the noise corrupted signal $|X_n(k)|$ to obtain the enhanced signal $|Y_n(k)|$. In case of DCT, the magnitude is the absolute value of the transformation of the noisy signal. The enhanced signal magnitude is then combined with the phase of noisy speech to get the spectrum for re-synthesizing speech. Time domain waveform segment is calculated from the speech by applying inverse transform (IDFT or IDCT). An overlap-add is carried out at the output to mask the discontinuities created due to the dissociation of magnitude and phase made in the short-time spectrum of the original signal.

A large number of variations of the basic technique have been developed for use in audio codecs and speech recognition. Berouti *et al.* developed a generalized spectral subtraction algorithm [1] in which the effects of musical noise are mitigated by employing an over-subtraction factor α and a spectral flooring factor β. It is given as

$$|Y_n(k)| = \begin{cases} \beta^{1/\gamma} D_n(k), & \text{if } |X_n(k)| < (\alpha+\beta)^{1/\gamma} D_n(k) \\ [|X_n(k)|^\gamma - \alpha(D_n(k))^\gamma]^{1/\gamma} & \text{otherwise} \end{cases} \quad (9)$$

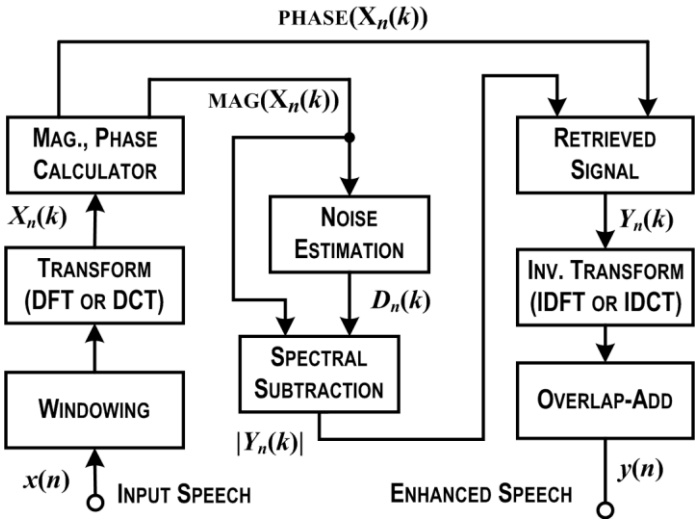


Figure 3.1. Spectral subtraction using DFT or DCT, adapted from [13]

where $\alpha > 1$, $0 < \beta << 1$, and γ is the exponent. It results in magnitude subtraction if $\gamma = 1$ and power subtraction if $\gamma = 2$. The purpose of α is to compensate for the underestimation of noise and to remove broadband spectral peaks, thus reducing the musical noise. The remnants of this operation are narrow spectral valleys, which are masked by using a flooring factor. Higher values of α may create distortions in speech. The retrieved magnitude spectrum is used to synthesize the cleaned speech signal.

In generalized spectral subtraction [1], it is assumed that noise affects the entire spectrum uniformly, which is generally not valid given the dynamic nature of noise. To overcome this, adaptive algorithms [1], with the parameter α varying with estimated SNR have been proposed. Kamath [14] proposed a multi-band approach to power spectral subtraction by using a band specific over-subtraction factor α_i and a tweaking factor δ_i. The estimated magnitude in the ith band is given as

$$|Y_n(k)| = \begin{cases} \beta^{1/2} X_n(k), & \text{if } |X_n(k)| < (\alpha_i \delta_i)^{1/2} D_n(k) \\ [|X_n(k)|^2 - \alpha_i \delta_i (D_n(k))^2]^{1/2} & \text{otherwise} \end{cases} \quad (10)$$

where α_i for band i varies with the segmental SNR (in dB) of the corresponding band and is computed as

$$\alpha_i = \begin{cases} 4.75, & \mathrm{SNR}_i < -5 \\ 4 - \frac{3}{20}(\mathrm{SNR}_i), & -5 \le \mathrm{SNR}_i \le 20 \\ 1, & \mathrm{SNR}_i > 20 \end{cases} \quad (11)$$

The band SNR_i (in dB) is estimated as

$$\mathrm{SNR}_i = 10\log_{10}\left[\left(\sum_{b_i}^{e_i} |X_n(k)|^2\right) / \left(\sum_{b_i}^{e_i} |D_n(k)|^2\right)\right] \quad (12)$$

The tweaking factor δ_i provides an additional degree of control that can be utilized to specify a frequency specific subtraction factor beside α_i which controls noise subtraction level based on the band SNR. The values are empirically determined in [14] as

$$\delta_i = \begin{cases} 1, & f_i \leq 1 \text{ kHz} \\ 2.5, & 1 \text{ kHz} < f_i \leq \left(F_s/2\right) - 2 \text{ kHz} \\ 1.5, & f_i > \left(F_s/2\right) - 2 \text{ kHz} \end{cases} \quad (13)$$

where f_i is the upper frequency of the ith band and Fs is the sampling frequency. The experiments were conducted by dividing the speech into 1 to 8 bands. Four linearly spaced frequency bands were found to be adequate in obtaining good speech quality. Similar approach can be followed for multi-band spectral subtraction in DCT domain.

3.2 Noise estimation

Noise spectrum needs to be dynamically estimated because of its non-stationary nature. The time corresponding to the number of past frames using which noise of the current frame is estimated is called the noise estimation duration. In the spectral subtraction algorithm, an under-estimation results in residual noise while an over-estimation results in distortion leading to degraded quality and possibly a low intelligibility. By using a voice activity detector or a speech/non-speech classifier, noise estimation is carried out as a moving average over several overlapping windows during silence intervals under the assumption that noise remains stationary during speech segments[15]. This method may not work satisfactorily under low SNR conditions and it may not track the variation in noise spectrum during long 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 some of which are discussed below.

3.2.1 Histogram-based techniques

In histogram-based methods [18], noise is estimated based on the histogram of each frequency bin of power spectrum. The value corresponding to the maximum of the histogram distribution over the past frames is considered as estimate of the noise spectrum in the current frame. Appropriate bin width for histogram at each frequency needs to be used to maintain a low variability yet a fine estimation of noise spectrum. Though this method was reported to perform well, it has a high computational complexity which makes it unsuitable for implementation on low-power processor.

3.2.2 Quantile-based noise estimation (QBNE)

Not all frequency bands are occupied by speech even during speech segments of the input signal and energy in each frequency band is at noise level for a significant fraction of time. Quantile-based noise estimation [16] is based on the observation that the signal energy

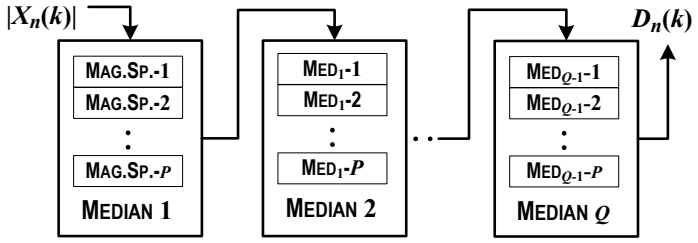


Figure 3.2. A *p*-point *q*-stage cascaded-median based noise estimation, from [13]

in a particular frequency bin is high due to the contribution by speech signal and occurs in only 10−20% of the frames. Therefore in this approach, noise spectrum is estimated by selecting a certain quantile from the previous frames of the noisy speech spectrum. The method is unsuitable for real-time operation as it involves sorting of past frames which is computation intensive and also has a large memory requirement.

3.2.3 Cascaded-median based noise estimation (CMBNE)

Out of the many statistics based noise estimation methods for single channel speech enhancement, a median based estimation has been reported to work in a robust manner [16]. As sorting operations for true median require a large amount of memory and time, it is not suitable for real time implementation. An approximation to the moving median with satisfactory saving in memory and computation is given by cascaded median [13]. In this method, the entire block of input is passed in small size frames and a local median is estimated, which is passed onto the next stage. So, in a *p*-frame *q*-stage cascaded median, an ensemble median is calculated once all the *p*-frames in a stage are filled and is output to the next stage. A block diagram of cascaded median is shown in Fig. 3.2. Thus an approximated median is obtained at the output after every $M = p^q$ input frames. Let us compare the computational complexities of true median and cascaded median, for noise estimated in each frequency bin every *M*-frames from the previous *M*-frames. For a true median based algorithm, a buffer of *M*-samples is required for sorting and the number of sorting operations are $M(M\text{-}1)/2$ i.e. $(M\text{-}1)/2$ per frame. The cascaded median requires *q p*-sample buffers i.e. a buffer of *qp* samples and $p(p\text{-}1)/2$ sorting operations per frame per frequency bin, as the median is calculated at only one stage at each frame position. Therefore, the saving ratios for storage and sorting are $M/(pq)$ and M/p.

3.2.4 Dynamic quantile tracking based noise estimation (DQTBNE)

Tiwari and Pandey [17] proposed a dynamic tracking algorithm to estimate the noise spectrum as a quantile of the input noisy spectrum. The algorithm does not involve storage and sorting of past samples. In this method, noise spectral sample in each frame is estimated by applying an increment or a decrement on the previous estimate. Real-time implementation using 0.25-quantile as the noise estimate gave satisfactory results in DFT domain.

Chapter 4

INVESTIGATIONS USING OFFLINE IMPLEMENTATION

4.1 Introduction

Real-time implementation using a low-power processor in a sensory aid requires a technique with low computational complexity. Generalized spectral subtraction with noise estimation based on cascaded-median was chosen for the same. Investigations were carried out to study the effect of processing methods and associated parameters on noise suppression. The investigations that were carried out for each of the processing steps during speech enhancement given in Fig. 3.1 are shown in Table 4.1. Processing was carried out for signal sampling frequency of 10 kHz. All implementations for investigations using offline processing were carried out using MATLAB.

Table 4.1. Processing steps and investigations

Processing step	Investigations carried out
Windowing	• Window type : rectangular, Hamming, Hanning, Griffin-Lim • Window overlap : 50%, 75% (as applicable with the window) • Window length : 10−50 ms
Transform calculation	• Transform type : DFT/ IDFT, DCT/ IDCT
Noise estimation	• Noise estimation methods : mean based noise estimation, CMBNE, DQTBNE • Noise estimation duration : past 27, 81, 161, 243 frames
Magnitude spectral subtraction	• Effect of parameters of generalized spectral subtraction. ($\gamma = 1$ throughout the report)
Spectrum calculation with phase	• Importance of phase in enhancing speech quality

Along with median, performance of other quantiles for noise tracking was also tested. The test materials and evaluation method used, and investigations are presented in the subsequent sections. Results are discussed in the last section.

4.2 Test materials and evaluation method

The speech material used for investigation is as follows -

a) Sentences taken from NOIZEUS database - This database [20] consists of 30 IEEE sentences recorded from 3 male and 3 female speakers which were originally sampled at 25 kHz and made available at 8 kHz. For testing, 7 sentences from the database (5 male and 2 female speakers) are concatenated and up-sampled to 10 kHz. The material is as the following

> "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. The clothes dried on a thin wooden rack. He wrote down a long list of items."

The total duration is 17 s and this material is rich in consonants. Hereafter this test material is referred to as NOIZEUS.

b) Lab recording - For a fast qualitative comparison, we have used speech material recorded in our lab. It consisted of a sequence of three isolated vowels, a Hindi sentence, and an English sentence as the following

"-/a/-/i/-/u/ - aayiye aapka naam kya hai? - Where were you a year ago?"

from a male speaker which was recorded with sampling frequency of 11.025 kHz. For informal listening test the signal length is increased by concatenating the recording four times one after the other to make it to ~25 s. The sampling rate is converted to 10 kHz before processing. This material is referred to as "vowel, Hindi sentence, English sentence" or VHSES.

c) Noise - Babble, car, street and train noises are taken from AURORA database [21] and have been concatenated repeatedly to match the length of the speech signals. Pink and white noises are generated using Matlab.

Test inputs to the code i.e. records of noisy speech were generated by adding the noises mentioned in (c) at SNRs (based on RMS values of signals) of 18, 15, 12, 9, 6, 3, 0, -3 and -6 dB. Implementation of the algorithms was carried out using Matlab for investigating the effects of various steps and parameters.

The outputs are qualitatively evaluated through informal listening. An objective assessment was carried out using PESQ score (scale : 0 – 4.5) [19], which is reported to have a reasonably good correlation with subjective assessment of speech quality and distortion. It is calculated from the difference between the loudness spectra of level-equalized and time aligned noise-free reference and test signals.

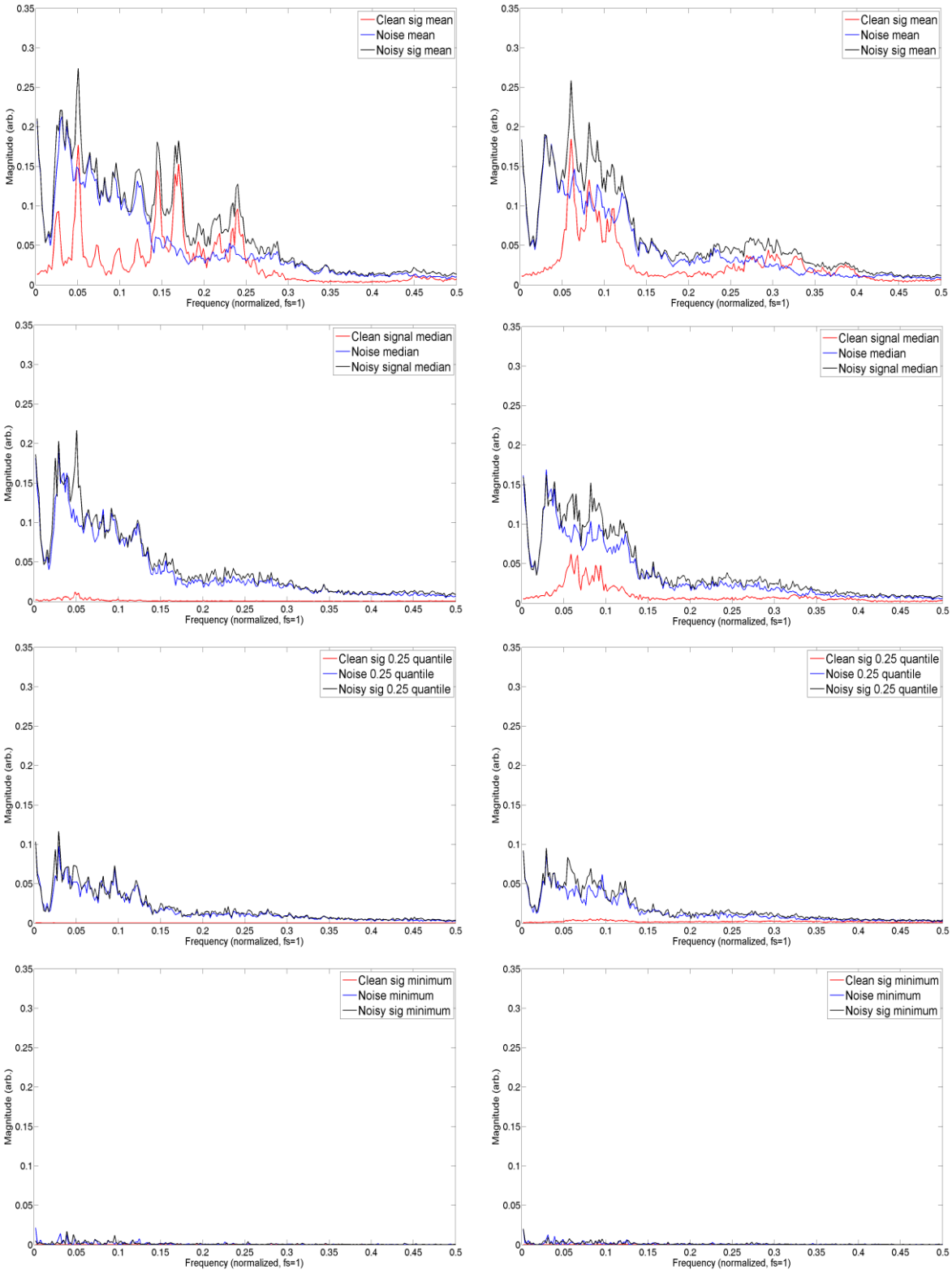
4.3 Investigations on noise estimation

Investigations were carried out to compare the noise estimation methods and determine the one that could best approximate noise spectrum in DCT domain similar to that reported in [22] for DFT. Mean, median (0.5-quantile), 0.25-quantile and minimum (0-quantile) statistics are used to observe and understand the trends in the magnitude distribution of the clean speech signal, noise signal and noisy speech signal with different noises and at

various SNRs. Processing was carried out using a rectangular window of length 30 ms, 50% overlap and 512-point DCT. From Figure 4.1, it can be seen that mean of the noisy speech poorly tracks the noise mean as against median, 0.25-quantile and minimum statistics of the noisy signal which relatively track the respective quantities of noise signal quite well at almost all frequencies. Although minimum tracks the noise it needs to be multiplied by a high factor to get the actual noise magnitude, errors in estimating which may lead to loss of speech related information from the noisy speech signal along with noise. Other noises also show similar trends and the plots are given in Appendix A.

Further investigation to examine the quantile suitable for spectral subtraction using DCT is carried out. The dynamic quantile tracking algorithm [17], as discussed in 3.2.4, is used for noise estimation with the quantiles 0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9 and 1. Since the estimation depends on the time characteristics of the noisy signal, it is assumed that the optimal values of parameters to be passed as inputs to the algorithm are similar for DFT and DCT. Spectrograms of noises estimated as various quantiles of the noisy speech material bit "aayiye aapka naam kya hai?" added with babble noise at 0 dB are shown in Figure 4.2. It is seen that noise is under-estimated in the lower quantiles and it increases for higher quantiles. The spectral subtraction factor α can be adjusted according to the quantile to attain the desired subtraction. We desire that an α value chosen with a given quantile provide us with decent enhancement over almost all kinds of broadband noises and over a wide range of SNRs. To identify the optimal alpha, the spectral subtraction algorithm is iterated with values of α ranging from 0 to 65 in steps of 0.1. Table 4.2−4.7 give the best PESQ scores that can be attained by a given quantile coupled with a suitable α value (optimal α values are also given in the tables) and β set to zero at SNRs of -6, -3, 0, 3 and 6 dB for various noises with the speech material VHSES.

It can be observed that all the quantiles track the noise spectrum considerably well and using appropriate over-subtraction factors result in similar enhancements. However, lower quantiles 0, 0.1 require extremely high values of α and its miscalculation results in sub-optimal enhanced signal quality. Higher quantiles pick-up speech and often require $\alpha < 1$ to compensate for the over-estimated noise magnitude. Error in calculation of α in this case will cause severe distortions and kills speech in low SNR frames.



(a) VHSES (b) NOIZEUS

Figure 4.1. Mean, median, 0.25-quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (babble, 0 dB)

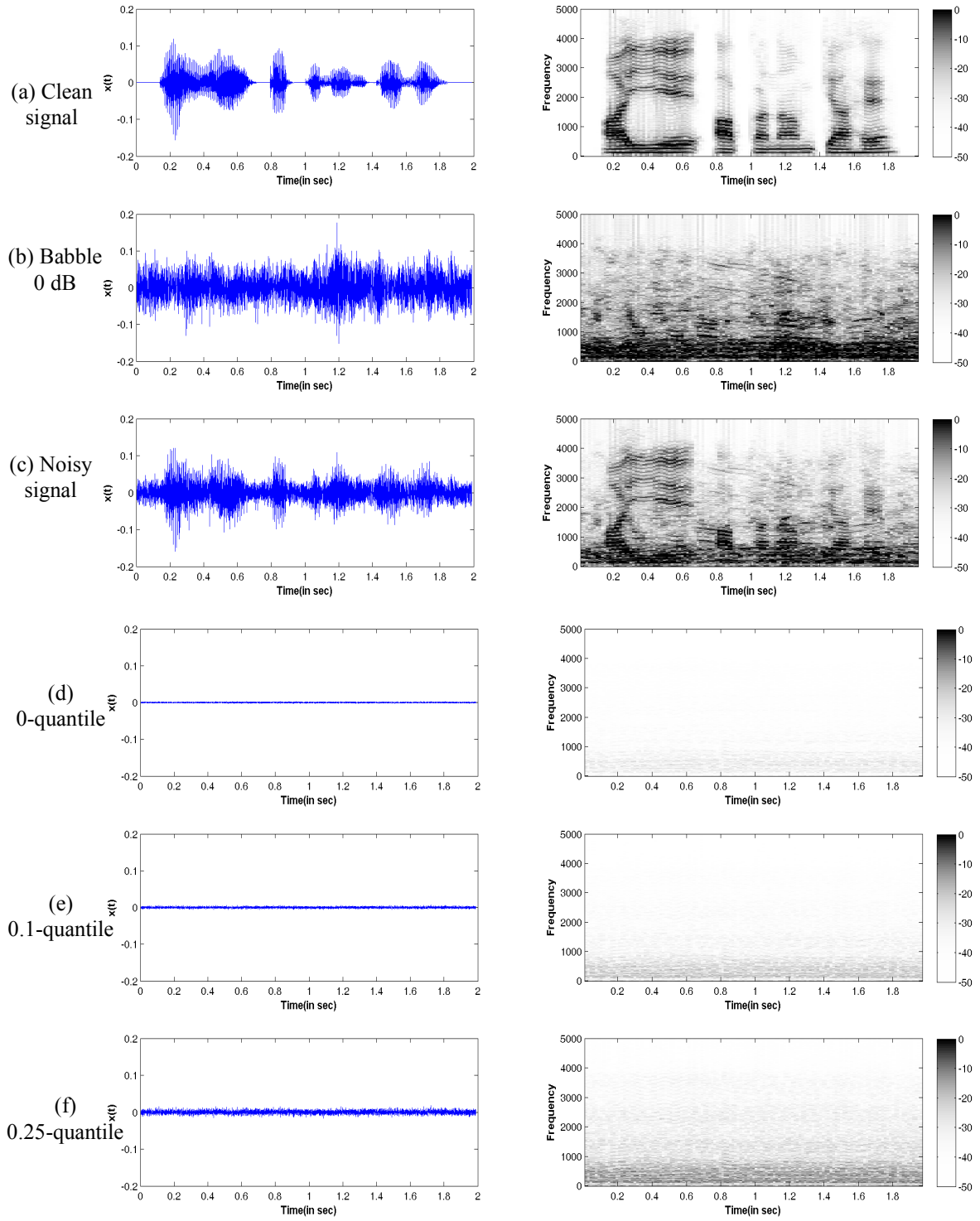


Figure 4.2. Spectrograms of noises estimated as various quantiles of the noisy speech material bit "aayiye aapka naam kya hai?" (babble, 0 dB)

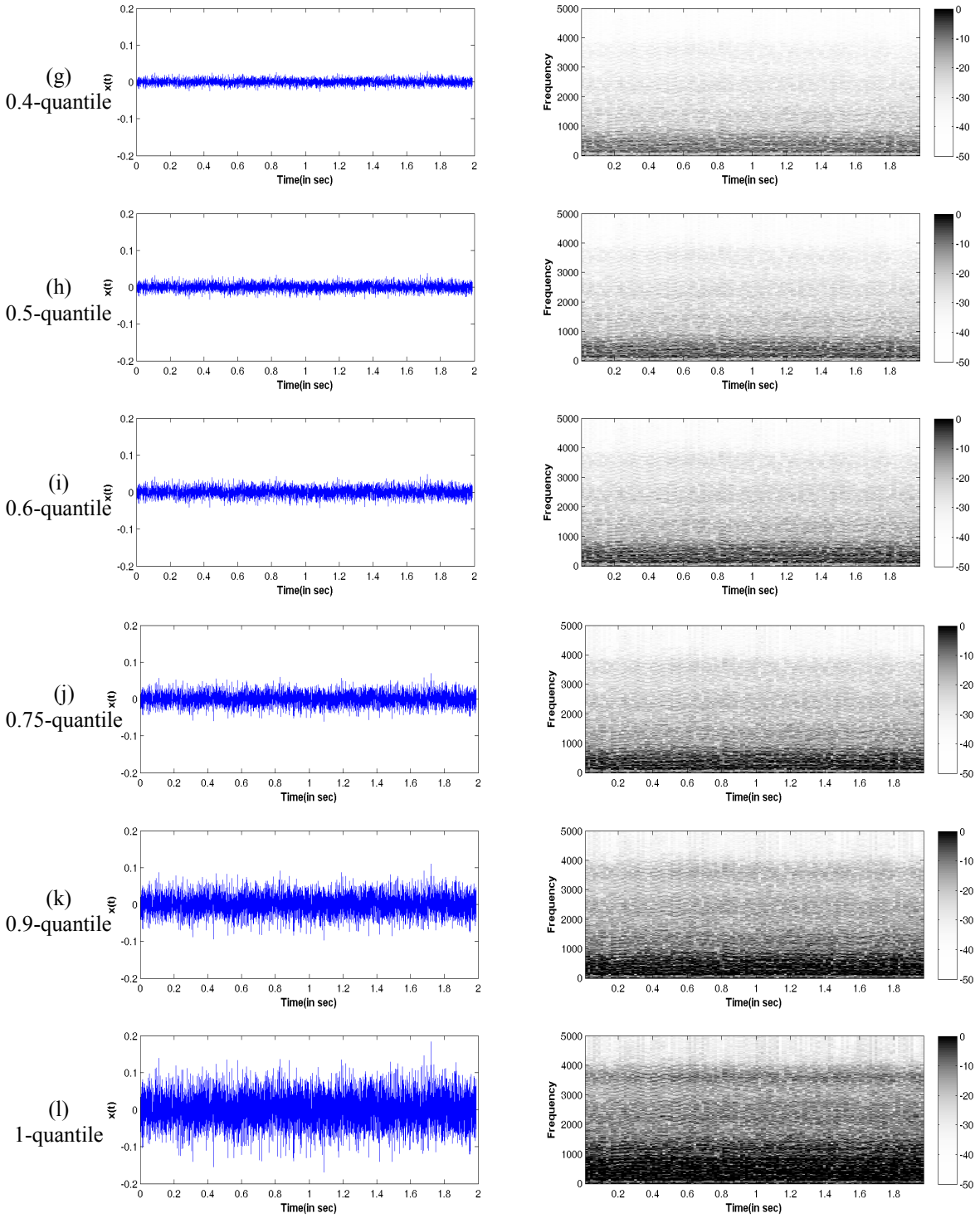


Figure 4.2. Spectrograms of noises estimated as various quantiles of the noisy speech material bit "aayiye aapka naam kya hai?" (babble, 0 dB)

Table 4.2 (a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Babble

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	25.8	11.5	3.8	2.6	2.2	1.7	1.0	0.3	0.3
-3	35.5	11.6	3.8	3.3	2.4	1.3	0.7	0.8	0.4
0	29.7	14.1	5.7	1.5	2.3	1.7	1.2	1.1	0.5
3	25.2	13.2	5.1	3.1	2.5	2.0	1.3	0.8	0.5
6	30.5	11.9	4.5	2.7	2.4	1.7	1.1	0.6	0.4

Table 4.2 (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Babble

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	1.68	1.71	1.72	1.70	1.69	1.71	1.71	1.69	1.78
-3	1.86	1.83	1.90	1.86	1.89	1.90	1.89	1.88	1.83
0	2.09	2.07	2.06	2.06	2.06	2.06	2.04	2.09	2.04
3	2.32	2.30	2.30	2.30	2.30	2.30	2.29	2.29	2.27
6	2.50	2.50	2.49	2.49	2.48	2.49	2.50	2.49	2.44

Table 4.3 (a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Car

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	31.5	15.4	5.6	3.4	2.7	2.2	1.8	0.9	0.6
-3	34.8	15.2	5.5	3.3	2.7	2.3	1.8	0.8	0.6
0	32.1	14.9	5.9	3.7	3.3	2.6	1.4	0.8	0.4
3	34.5	14.8	6.1	3.8	2.9	2.0	1.7	0.9	0.5
6	29.7	15.5	5.9	3.6	2.8	2.1	1.3	0.9	0.4

Table 4.3 (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Car

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	1.72	1.73	1.74	1.73	1.73	1.72	1.70	1.66	1.67
-3	1.96	1.96	1.94	1.95	1.94	1.95	1.91	1.90	1.91
0	2.24	2.20	2.23	2.34	2.20	2.20	2.18	2.17	2.15
3	2.39	2.37	2.39	2.37	2.38	2.39	2.36	2.33	2.30
6	2.58	2.59	2.61	2.61	2.62	2.61	2.58	2.53	2.46

Table 4.4 (a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Pink

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	53.9	17.8	6.9	4.0	3.0	2.5	1.9	1.3	0.8
-3	54.3	19.3	7.0	4.2	3.3	2.8	1.8	1.3	0.8
0	50.8	18.9	7.0	4.3	4.6	2.7	1.9	1.2	0.8
3	34.4	20.5	7.1	4.4	4.2	3.3	2.4	1.6	0.6
6	33.3	21.2	8.3	4.9	3.6	2.9	2.1	1.3	0.5

Table 4.4 (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Pink

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	1.81	1.84	1.84	1.84	1.85	1.82	1.80	1.77	1.71
-3	1.98	2.10	2.07	2.05	2.06	2.05	2.02	1.98	1.86
0	2.16	2.26	2.28	2.27	2.23	2.58	2.24	2.19	2.04
3	2.37	2.47	2.50	2.50	2.49	2.48	2.45	2.37	2.21
6	2.55	2.66	2.73	2.73	2.72	2.70	2.64	2.55	2.40

Table 4.5 (a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Street

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	28.4	9.4	2.6	2.9	1.2	0.9	0.6	0.8	0.2
-3	30.3	11.2	4.0	2.7	2.1	1.3	1.3	0.7	0.4
0	22.5	12.3	4.4	2.6	2.3	1.7	1.1	0.6	0.4
3	26.4	11.9	4.8	2.8	2.2	1.6	1.4	0.6	0.3
6	21.2	11.5	5.0	2.7	1.9	1.9	0.9	0.5	0.3

Table 4.5 (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Street

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	1.66	1.66	1.65	1.64	1.65	1.64	1.61	1.61	1.59
-3	1.85	1.83	1.83	1.84	1.83	1.82	1.81	1.81	1.80
0	2.02	2.03	2.03	2.03	2.04	2.04	2.02	2.01	1.99
3	2.27	2.26	2.27	2.27	2.26	2.26	2.25	2.25	2.23
6	2.50	2.52	2.52	2.52	2.51	2.50	2.51	2.48	2.43

Table 4.6 (a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Train

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	37.1	16.2	5.9	4.1	2.6	2.6	1.9	1.1	0.7
-3	32.0	14.7	5.3	3.5	2.7	2.0	1.7	1.1	0.5
0	27.3	15.3	5.4	2.9	2.4	2.3	1.4	0.8	0.3
3	22.0	13.5	5.9	3.3	2.8	2.2	1.2	0.7	0.3
6	17.4	13.1	5.8	4.4	3.3	1.9	1.0	0.6	0.2

Table 4.6 (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : Train

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	2.25	2.26	2.28	2.29	2.28	2.27	2.24	2.20	2.14
-3	2.45	2.43	2.46	2.45	2.45	2.45	2.43	2.4	2.34
0	2.64	2.66	2.68	2.67	2.68	2.64	2.64	2.58	2.50
3	2.80	2.84	2.87	2.87	2.88	2.86	2.83	2.78	2.68
6	2.97	3.01	3.04	3.05	3.04	3.03	3.00	2.94	2.89

Table 4.7 (a) Optimal values of α for a given quantile that result in highest PESQ scores by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : White

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	55.1	17.0	6.1	3.9	2.6	2.4	1.8	1.2	0.7
-3	58.4	16.5	6.1	4.7	2.9	2.3	2.2	1.4	0.8
0	65.9	30.9	9.4	5.8	4.4	3.6	2.5	1.7	0.9
3	59.6	17.5	9.7	5.8	4.5	3.6	2.5	1.7	0.9
6	58.7	22.6	8.1	6.1	4.3	3.3	2.3	1.6	0.8

Table 4.7 (b) PESQ scores for optimal values of α for a given quantile by DQTBNE and spectral subtraction ($\beta = 0$, $\gamma = 1$), Speech material : VHSES, Noise : White

SNR (dB)	Quantile								
	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
-6	1.54	1.65	1.64	1.64	1.62	1.64	1.62	1.57	1.56
-3	1.67	1.84	1.78	1.79	1.82	1.81	1.77	1.73	1.64
0	1.86	2.06	2.07	2.07	2.07	2.06	2.03	1.95	1.78
3	2.08	2.23	2.36	2.37	2.35	2.33	2.27	2.16	1.93
6	2.32	2.48	2.51	2.54	2.54	2.51	2.43	2.31	2.08

Table 4.8 PESQ scores obtained using spectral subtraction and DQTBNE. $\alpha = 2$, $\beta = 0$ and $\gamma = 1$, Speech material : VHSES with various noises.

(a) Babble noise

SNR (dB)	Quantile			
	0.25	0.4	0.5	0.6
-6	1.66	1.65	1.64	1.64
-3	1.78	1.82	1.82	1.82
0	1.97	2.04	2.04	2.04
3	2.18	2.25	2.29	2.30
6	2.39	2.44	2.47	2.47

(b) Car noise

SNR (dB)	Quantile			
	0.25	0.4	0.5	0.6
-6	1.51	1.64	1.64	1.68
-3	1.70	1.84	1.86	1.90
0	2.02	2.13	2.16	2.17
3	2.15	2.24	2.27	2.39
6	2.37	2.49	2.57	2.61

(c) Pink noise

SNR (dB)	Quantile			
	0.25	0.4	0.5	0.6
-6	1.44	1.54	1.67	1.75
-3	1.60	1.76	1.89	1.97
0	1.80	2.00	2.08	2.20
3	2.06	2.20	2.30	2.39
6	2.27	2.42	2.52	2.62

(d) Street noise

SNR (dB)	Quantile			
	0.25	0.4	0.5	0.6
-6	1.64	1.59	1.61	1.59
-3	1.79	1.82	1.83	1.79
0	1.96	2.00	2.03	2.00
3	2.18	2.24	2.26	2.25
6	2.39	2.50	2.51	2.49

(e) Train noise

SNR (dB)	Quantile			
	0.25	0.4	0.5	0.6
-6	1.94	2.12	2.19	2.23
-3	2.21	2.37	2.39	2.45
0	2.44	2.55	2.63	2.66
3	2.61	2.78	2.84	2.85
6	2.83	2.98	3.03	3.02

(f) White noise

SNR (dB)	Quantile			
	0.25	0.4	0.5	0.6
-6	1.41	1.47	1.53	1.60
-3	1.52	1.61	1.70	1.75
0	1.65	1.79	1.87	1.92
3	1.79	1.99	2.05	2.13
6	2.02	2.16	2.25	2.32

Hence 0, 0.1, 0.75, 0.9 and 1 quantiles were not considered for noise estimation. In the quantile range of 0.25−0.6, 0.4-quantile and 0.5-quantile exhibit almost constant α values over different SNRs for various noises. Henceforth, investigations were carried out using CMBNE as described in 3.2.3 for tracking noise as 0.5-quantile (median) of the input noisy speech spectrum. Additionally, performance of the quantiles 0.25−0.6 for $\alpha = 2$ are given in Table 4.8.

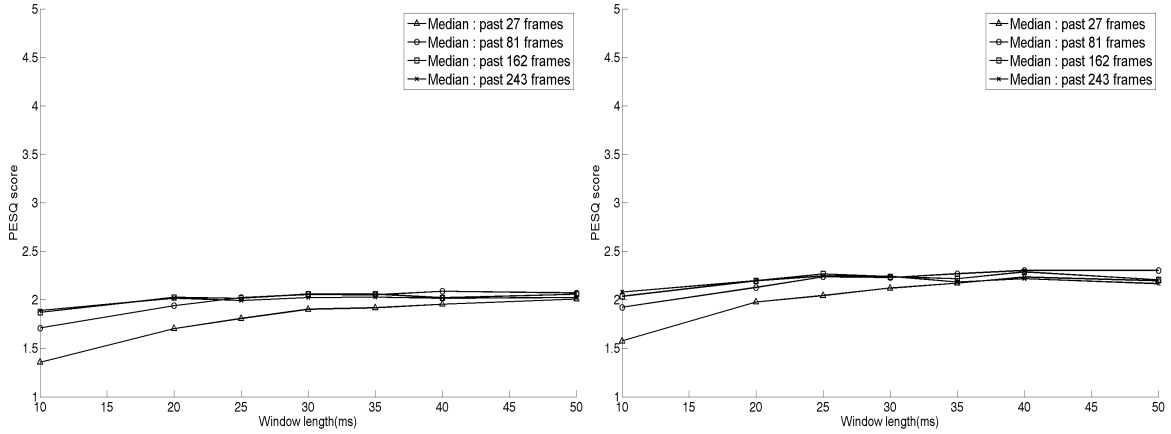
4.4 Effect of noise estimation duration and window lengths

Investigations were carried out to examine the effects of noise estimation duration and window length on speech enhancement. Processing was carried out using a rectangular window of length varying from 10 ms to 50 ms, with 50% overlap and 512-point DCT. Noise is estimated for the past 27, 81, 162 and 243 frames using the sample quantiles. Figure 4.3 and Figure 4.4 show how the PESQ scores vary with changing window lengths and how they differ across the mentioned noise estimation durations for the speech materials VHSES and NOIZEUS. Window length of 25−40 ms provide good scores with noise estimated over past 81, 162 and 243 frames. Hence, further investigations were carried out using window of length 30ms. Since use of 81, 162, and 243 frames resulted in almost similar tracking of noise spectrum, an estimation over past 81 frames is chosen which corresponds to a duration of 1.215 s. Since MBNE is computation intensive due to the sorting operations and memory requirement, the CMBNE described in 3.2.3 is used which involves a 3-frame 4-stage cascaded median as an approximation to the median over past 81 frames. In this approach, for each frequency bin, 12 memory locations are required for storage and the number of sorting operations are 1.48 per frame as compared to that of 81 and 40 in case of a true median based approach.

4.5 Effect of window type and overlap amount

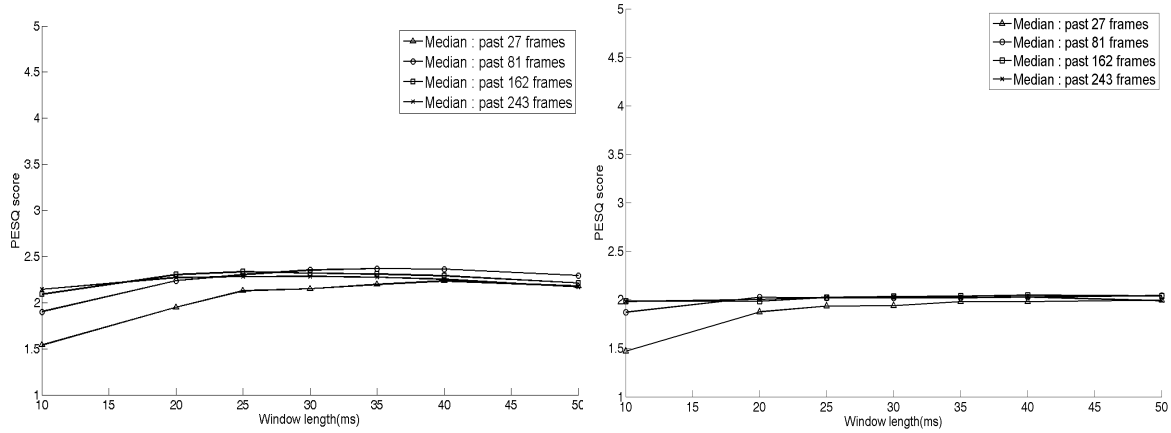
Fixed-frame processing results in discontinuities at the frame edges, to mitigate which an overlap-add is used. These artifacts effect the enhancement that can be obtained and vary depending on the length of the window, shift amount and its smoothening nature, i.e. the windowing function.

Investigations using rectangular window, Hamming window, Hanning window and the modified Hamming window proposed by Griffin-Lim [23] are carried out. In standard overlap-add method, a window should satisfy the requirement that sum of all the overlapping windows is unity. For Griffin-Lim window, sum of squares of overlapping windows should also sum to one. This requirement is met by the modified Hamming window only for 75% overlap. So comparisons are made between 50% and 75% of rectangular, Hamming and Hanning windows and 75% of Griffin-Lim window and results are given in Table 4.9 and Table 4.10. The PESQ scores obtained using rectangular window with 50% overlap were the best. Other windows and shift durations showed slight decrement ~0.01−0.02 in the scores. These results as against the expected better performance that might be obtained using Hamming or related modified windows can be attributed to the calculation of the objective measure itself. Some other quality methods and extensive listening tests are to be performed to establish the superior performance of a particular window type and overlap amount.



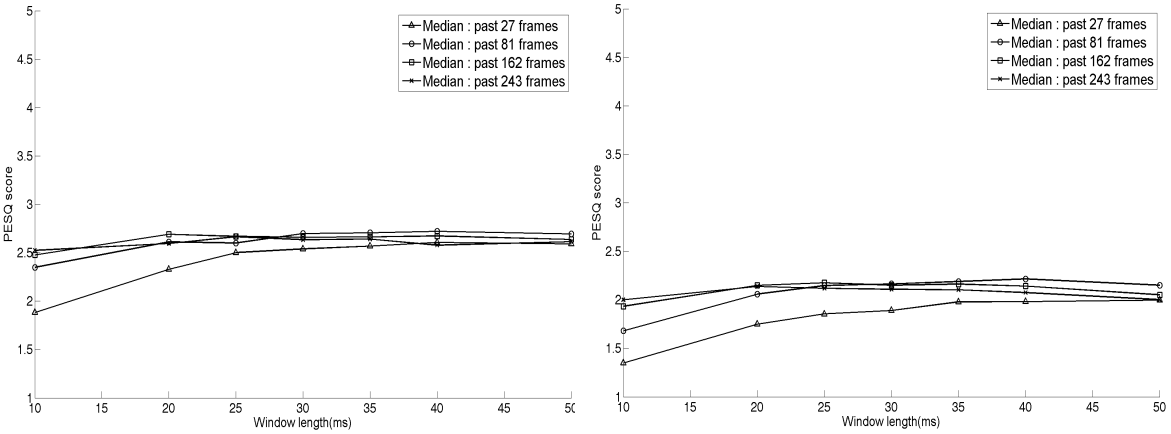
(a) Babble noise (α = 3)

(b) Car noise (α = 2.5)



(c) Pink noise (α = 3.5)

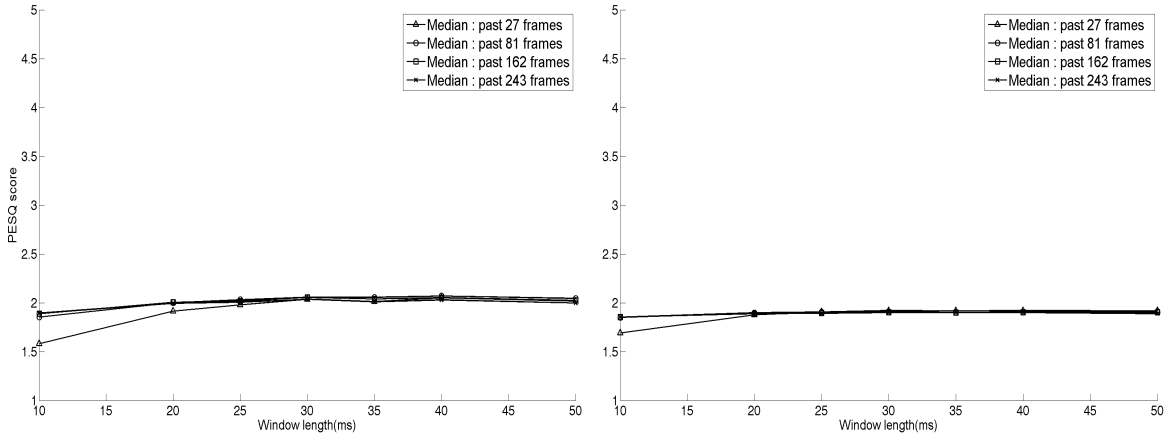
(d) Street noise (α = 2)



(e) Train noise (α = 2.5)

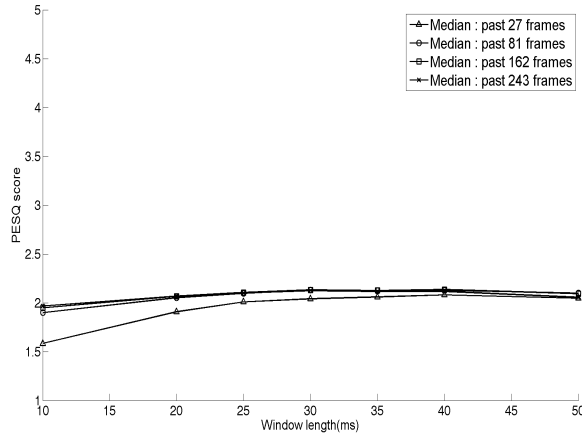
(f) White noise (α = 3.5)

Figure 4.3. PESQ scores for the enhanced noisy speech using sample quantile and spectral subtraction, β = 0.001, γ = 1 at optimum values of α. Speech material : VHSES with babble, car, train, pink, white, and street noises at 0 dB SNR.

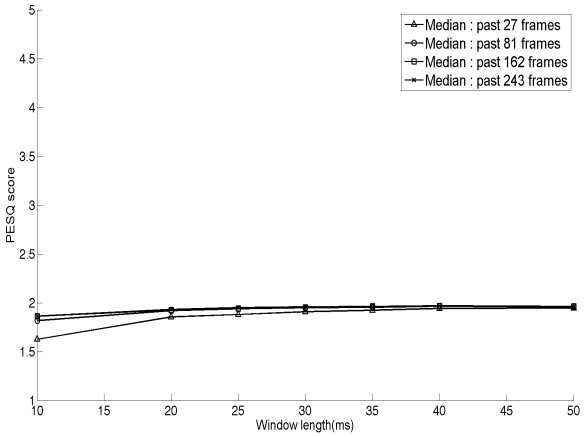


(a) Babble noise (α = 1.5)

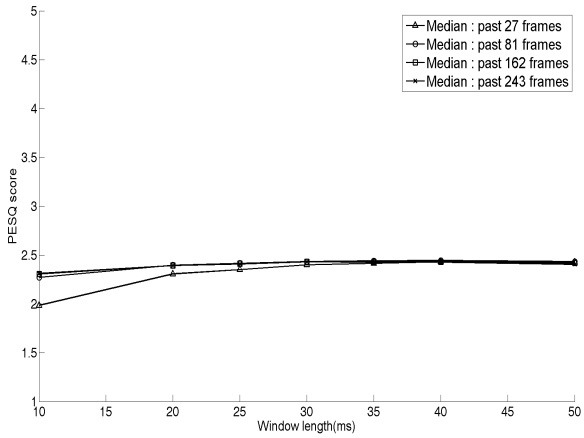
(b) Car noise (α = 2.5)



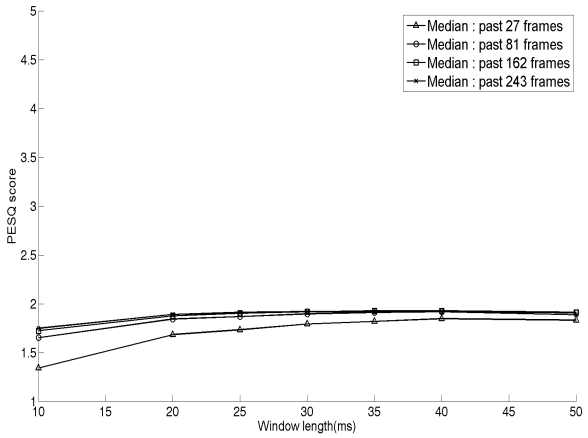
(c) Pink noise (α = 1.5)



(d) Street noise (α = 2.5)



(e) Train noise (α = 1.5)



(f) White noise (α = 2.5)

Figure 4.4. PESQ scores for the enhanced noisy speech using sample quantile and spectral subtraction, β = 0.001, γ = 1 at optimum values of α. Speech material : NOIZEUS with babble, car, train, pink, white, and street noises at 0 dB SNR.

Table 4.9 PESQ scores for enhanced speech using CMBNE and spectral subtraction with different types of windowing functions. $\gamma = 1$; $\beta = 0.01$, α given in brackets corresponding to the scores. Speech material : VHSES with various noises at 0 dB.

(a)Babble noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.07 (3.0)	2.08 (1.5)
Hamming	2.02 (2.0)	2.00 (1.5)
Hanning	2.02 (1.5)	2.00 (1.5)
Griffin-Lim	NA	2.00 (1.0)

(b)Car noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.3 (2.5)	2.22 (3.0)
Hamming	2.17 (2.5)	2.13 (2.5)
Hanning	2.17 (2.5)	2.11 (2.5)
Griffin-Lim	NA	2.09 (2.5)

(c)Pink noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.33 (4.0)	2.34 (4.0)
Hamming	2.25 (4.0)	2.24 (4.0)
Hanning	2.22 (4.0)	2.21 (4.0)
Griffin-Lim	NA	2.11 (3.0)

(d)Street noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.03 (2.0)	2.03 (2.0)
Hamming	1.96 (1.5)	1.96 (2.0)
Hanning	1.97 (1.5)	1.94 (1.5)
Griffin-Lim	NA	1.95 (1.5)

(e)Train noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.66 (2.5)	2.66 (2.5)
Hamming	2.58 (2.5)	2.57 (2.5)
Hanning	2.58 (2.5)	2.55 (2.5)
Griffin-Lim	NA	2.51 (2.0)

(f)White noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.18 (4.0)	2.14 (4.0)
Hamming	2.05 (4.0)	2.07 (4.0)
Hanning	2.03 (4.0)	2.03 (4.0)
Griffin-Lim	NA	1.94 (4.0)

Table 4.10 PESQ scores for enhanced speech using CMBNE and spectral subtraction with different types of windowing functions. $\gamma = 1$; $\beta = 0.001$, α given in brackets corresponding to the scores. Speech material : NOIZEUS with various noises at 0 dB.

(a)Babble noise

Window	Overlap (α value)	
	50%	75%
Rectangular	1.91 (2.0)	1.91 (2.0)
Hamming	1.87 (1.5)	1.88 (1.5)
Hanning	1.87 (1.5)	1.87 (1.5)
Griffin-Lim	NA	1.88 (1.5)

(b)Car noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.01 (2.5)	2.07 (2.5)
Hamming	1.97 (3.0)	2.04 (2.0)
Hanning	1.99 (2.0)	2.02 (2.0)
Griffin-Lim	NA	2.04 (2.0)

(c)Pink noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.13 (3.0)	2.15 (3.0)
Hamming	2.07 (3.0)	2.10 (2.5)
Hanning	2.04 (2.5)	2.08 (2.5)
Griffin-Lim	NA	2.10 (2.5)

(d)Street noise

Window	Overlap (α value)	
	50%	75%
Rectangular	1.96 (1.5)	1.95 (1.5)
Hamming	1.92 (2.0)	1.92 (1.0)
Hanning	1.93 (1.5)	1.92 (1.0)
Griffin-Lim	NA	1.93 (1.0)

(e)Train noise

Window	Overlap (α value)	
	50%	75%
Rectangular	2.45 (2.5)	2.50 (3.0)
Hamming	2.42 (1.5)	2.48 (3.0)
Hanning	2.44 (2.0)	2.46 (3.0)
Griffin-Lim	NA	2.48 (3.0)

(f)White noise

Window	Overlap (α value)	
	50%	75%
Rectangular	1.93 (3.0)	1.95 (3.0)
Hamming	1.88 (2.5)	1.90 (3.0)
Hanning	1.86 (2.5)	1.88 (3.0)
Griffin-Lim	NA	1.90 (3.0)

4.6 Effect of over-subtraction factor

The over-subtraction factor α is the most important of all the parameters which affects the residual noise and distortion and this affects intelligibility and quality of the processed output. Investigations were carried out by varying α in the range of $0 - 4$ with the floor factor β taking values 0, 0.001, 0.01 and 0.1. Analysis was carried out using rectangular window of duration 30 ms with 50% overlap and a 3-point 4-stage CMBNE. Table 4.11 and Table 4.12 show the PESQ scores of speech enhanced from the input speech materials VHSES and NOIZEUS respectively, which are added with various noises at 0 dB, using different sets of α and β. For VHSES, α values of $2.5 - 3.0$ are found to be optimal for babble, car, street, and train noises and $3.5 - 4.0$ for pink, white noises. For NOIZEUS, the scores were best for α as 2 for car, babble, street, and train noises and 2.5−3 for pink, white noises. Floor factor $\beta = 0.01$ for VHSES and $\beta = 0.001$ for NOIZEUS are found to be optimal. Investigations need to be carried out to see the scope of further improvement that can be obtained in the quality when SNR-dependent α or frequency-dependent α is used.

4.7 Phase estimation for spectral subtraction

The estimated magnitude spectrum of noise subtracted from the magnitude spectrum of noisy speech gives us the magnitude spectrum of enhanced speech signal. This resulting magnitude is recombined with the original phase i.e. phase of the input noisy signal which is followed by applying an inverse transformation and overlap-add for resynthesis. It may be expected that quality will improve if the phase spectrum is also corrected. To understand the enhancement that phase can provide in the signal quality, a set of simulations were carried out by using combinations of clean magnitude (CM), clean phase (CP), noisy magnitude (NM), noisy phase (NP), and modified magnitude (MM) using CMBNE and spectral subtraction. Results are given in Table 4.13. Clean magnitude and clean phase correspond to the clean signal whose PESQ score is 4.5. It can be seen from the table that an improvement of upto 0.2 in the scores can be obtained by cleaning the phase, whereas a cleaner estimate of magnitude can bring about much higher improvement.

The human ear may be relatively insensitive to phase, but a phase discontinuity may significantly degrade the perceptual quality especially when the discontinuities occur during voiced segments. Investigations were carried out regarding impact of phase changes at different magnitude levels on the quality of speech material VHSES and are given in Appendix B. It can be observed that, phase is important in the voiced segments of high energy and thus an attempt to mitigate the phase discontinuity in such frames is made which is presented in Appendix B. Mowlaee [28] proposed a method based on minimum mean square error (MMSE) for phase estimation in DFT domain. This method was adapted to DCT, resulting in an approach similar to the one described in Appendix B. Both, the methods were

found to introduce a monotonic voice to the speech signal and hence were considered to be unsuitable for speech enhancement.

4.8 Comparison between DCT and DFT

A comparison in made by applying the enhancement algorithm in DCT and DFT domains for signals with different types of noise and with SNR of 18, 15, 12, 9, 6, 3, 0, −3 and −6 dB. Plots of PESQ scores vs input SNR are given in Figure 4.5 and Figure 4.6 for VHSES and NOIZEUS. The PESQ scores increased progressively as SNR increases and both the transforms proved to improve the quality of the noisy signal as their scores remained above that of the unprocessed while processing of noise-free speech decreased the score to around 3.5 ± 0.2 depending on the α value. The scores obtained from both the transforms are comparable with DCT slightly better than DFT for street and train noises for the material VHSES. Table 4.14 shows the SNR advantage for different types of noises and optimal α used for DFT and DCT, which are again similar. For VHSES, an advantage of 4 − 13 dB in SNR is obtained using DFT while DCT resulted in an advantage of 4 − 12.5 dB. For NOIZEUS, the advantage was 2 − 8 dB with DFT while it was 2.5 − 8 dB with DCT.

4.9 Conclusion

Optimal parameters for real-time implementation are obtained through these investigations. Processing was carried out with sampling frequency of 10 kHz keeping the FFT/FCT length at 512. Window length of 30 ms i.e. 300 samples for a noise estimation duration over past 81 frames was chosen. Magnitude spectral subtraction ($\gamma = 1$) was carried out throughout. Good scores were obtained with α in the range of 2 − 3.5 and $\beta = 0.01$ was suitable for all the cases. Median of the noisy input speech was found to be a good approximate of noise spectrum and its approximation, the cascaded-median was used for noise estimation. Processing significantly enhanced the speech for all noises with different SNRs. Also, the enhancement provided by DCT based method is comparable to that of existing DFT-based spectral subtraction. PESQ scores did not show any difference in the results obtained by the two methods.

Table 4.11 PESQ score for enhanced speech using CMBNE and spectral subtraction, α and β varying, γ = 1. Speech material : VHSES with different types of noise at 0 dB SNR.

(a)Babble noise

α	β			
	0	0.001	0.01	0.1
0	1.82	1.82	1.82	1.82
1	1.97	1.97	1.97	1.97
1.5	2.04	2.04	2.04	2.05
2	2.06	2.06	2.07	2.09
2.5	2.09	2.09	2.09	2.13
3	2.07	2.07	2.08	2.14
3.5	2.05	2.05	2.07	2.15
4	2.01	2.01	2.04	2.16

(b)Car noise

α	β			
	0	0.001	0.01	0.1
0	1.78	1.78	1.78	1.77
1	2.03	2.03	2.03	2.03
1.5	2.13	2.13	2.13	2.13
2	2.20	2.20	2.20	2.20
2.5	2.22	2.22	2.23	2.24
3	2.23	2.23	2.24	2.25
3.5	2.21	2.21	2.22	2.23
4	2.17	2.18	2.19	2.21

(c)Pink noise

α	β			
	0	0.001	0.01	0.1
0	1.61	1.61	1.61	1.61
1	1.92	1.92	1.92	1.91
1.5	2.09	2.09	2.09	2.07
2	2.15	2.15	2.15	2.13
2.5	2.24	2.24	2.24	2.2
3	2.30	2.30	2.31	2.23
3.5	2.32	2.32	2.35	2.24
4	2.33	2.34	2.36	2.24

(d)Street noise

α	β			
	0	0.001	0.01	0.1
0	1.93	1.93	1.93	1.93
1	2.15	2.15	2.16	2.16
1.5	2.23	2.23	2.23	2.25
2	2.27	2.27	2.28	2.29
2.5	2.28	2.28	2.29	2.32
3	2.25	2.26	2.28	2.33
3.5	2.28	2.29	2.31	2.32
4	2.21	2.22	2.25	2.33

(e)Train noise

α	β			
	0	0.001	0.01	0.1
0	2.12	2.12	2.12	2.12
1	2.48	2.48	2.48	2.47
1.5	2.57	2.57	2.57	2.56
2	2.64	2.64	2.64	2.64
2.5	2.67	2.67	2.68	2.68
3	2.67	2.67	2.68	2.68
3.5	2.64	2.64	2.65	2.67
4	2.61	2.61	2.63	2.63

(f)White noise

α	β			
	0	0.001	0.01	0.1
0	1.47	1.47	1.47	1.47
1	1.65	1.65	1.65	1.65
1.5	1.79	1.79	1.78	1.77
2	1.93	1.93	1.93	1.89
2.5	1.97	1.97	2.00	1.99
3	2.10	2.09	2.1	2.04
3.5	2.14	2.14	2.15	2.00
4	2.13	2.13	2.16	2.03

Table 4.12 PESQ score for enhanced speech using CMBNE and spectral subtraction, α and β varying and $\gamma = 1$. Speech material : NOIZEUS with different types of noise at 0 dB SNR.

(a)Babble noise

α	β			
	0	0.001	0.01	0.1
0	1.74	1.74	1.74	1.74
1	1.87	1.87	1.87	1.87
1.5	1.91	1.91	1.91	1.92
2	1.91	1.91	1.91	1.94
2.5	1.90	1.91	1.9	1.96
3	1.88	1.90	1.88	1.97
3.5	1.85	1.88	1.86	1.97
4	1.82	1.85	1.82	1.97

(b)Car noise

α	β			
	0	0.001	0.01	0.1
0	1.75	1.75	1.75	1.75
1	1.93	1.93	1.93	1.92
1.5	1.98	1.98	1.98	1.98
2	2.01	2.02	2.01	2.02
2.5	2.01	2.02	2.01	2.04
3	2.00	2.02	2.00	2.05
3.5	1.96	2.00	1.96	2.04
4	1.88	1.95	1.89	2.02

(c)Pink noise

α	β			
	0	0.001	0.01	0.1
0	1.66	1.66	1.66	1.66
1	1.88	1.87	1.88	1.87
1.5	1.98	1.98	1.98	1.96
2	2.06	2.05	2.06	2.04
2.5	2.12	2.12	2.12	2.08
3	2.13	2.15	2.13	2.11
3.5	2.09	2.13	2.09	2.11
4	2.03	2.09	2.03	2.09

(d)Street noise

α	β			
	0	0.001	0.01	0.1
0	1.82	1.82	1.82	1.82
1	1.94	1.94	1.94	1.94
1.5	1.96	1.96	1.96	1.96
2	1.94	1.94	1.94	1.96
2.5	1.90	1.91	1.90	1.94
3	1.84	1.86	1.84	1.91
3.5	1.77	1.80	1.77	1.89
4	1.71	1.75	1.72	1.86

(e)Train noise

α	β			
	0	0.001	0.01	0.1
0	2.14	2.14	2.14	2.14
1	2.34	2.34	2.34	2.34
1.5	2.41	2.41	2.41	2.41
2	2.45	2.45	2.45	2.46
2.5	2.45	2.46	2.45	2.49
3	2.44	2.46	2.44	2.49
3.5	2.40	2.45	2.40	2.47
4	2.36	2.43	2.37	2.43

(f)White noise

α	β			
	0	0.001	0.01	0.1
0	1.53	1.53	1.53	1.53
1	1.70	1.70	1.70	1.70
1.5	1.78	1.78	1.78	1.77
2	1.85	1.85	1.85	1.84
2.5	1.90	1.90	1.90	1.89
3	1.93	1.94	1.93	1.92
3.5	1.90	1.93	1.90	1.92
4	1.81	1.88	1.82	1.90

Table 4.13 PESQ scores of combinations of magnitude and phase for the speech bit "Where were you a year ago?" from speech material VHSES.

(a) Babble noise

SNR	NM + NP	NM + CP	CM + NP	MM + NP	MM + CP
-3 dB	1.53	1.88	2.83	1.47	1.62
0 dB	1.68	2.04	2.95	1.77	1.90
3 dB	1.86	2.20	3.06	1.99	2.14

(b) Car noise

SNR	NM + NP	NM + CP	CM + NP	MM + NP	MM + CP
-3 dB	1.45	1.91	2.87	1.65	1.75
0 dB	1.62	2.06	2.99	1.73	1.92
3 dB	1.80	2.22	3.11	2.18	2.33

(c) Pink noise

SNR	NM + NP	NM + CP	CM + NP	MM + NP	MM + CP
-3 dB	1.24	1.70	2.79	1.63	1.71
0 dB	1.41	1.85	2.94	2.02	2.10
3 dB	1.59	2.02	3.10	2.28	2.33

(d) Street noise

SNR	NM + NP	NM + CP	CM + NP	MM + NP	MM + CP
-3 dB	1.24	1.70	2.79	1.63	1.71
0 dB	1.41	1.85	2.94	2.02	2.10
3 dB	1.59	2.02	3.10	2.28	2.33

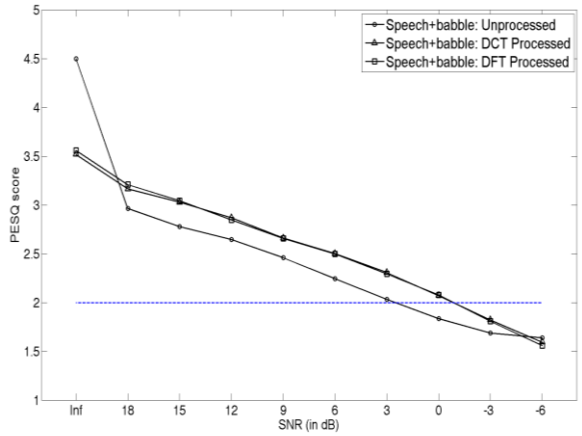
(e) Train noise

SNR	NM + NP	NM + CP	CM + NP	MM + NP	MM + CP
-3 dB	1.84	2.19	2.88	2.30	2.37
0 dB	2.04	2.37	3.04	2.48	2.49
3 dB	2.24	2.56	3.15	2.81	2.87

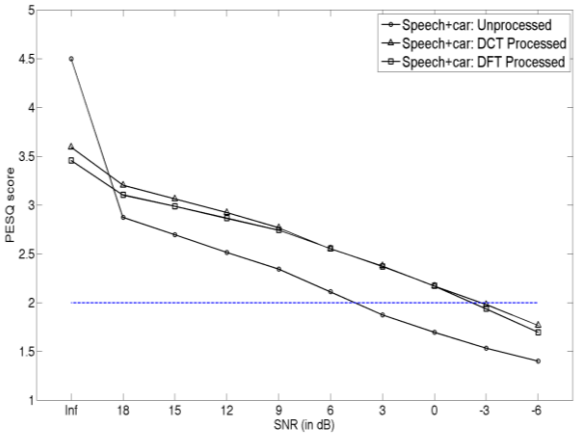
(f) White noise

SNR	NM + NP	NM + CP	CM + NP	MM + NP	MM + CP
-3 dB	1.17	1.56	2.88	1.53	1.59
0 dB	1.27	1.66	2.97	1.75	1.81
3 dB	1.40	1.79	3.10	2.04	2.11

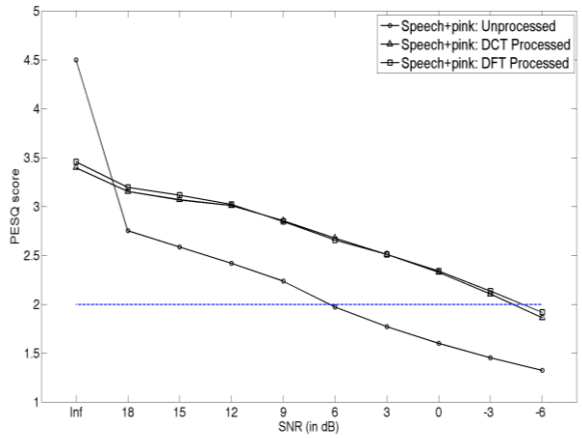
*Magnitude modified using CMBNE and spectral subtraction, $\beta = 0$, $\gamma = 1$, $\alpha = 2.5$ for babble, street, train and $\alpha = 3$ for car, pink, white noises.



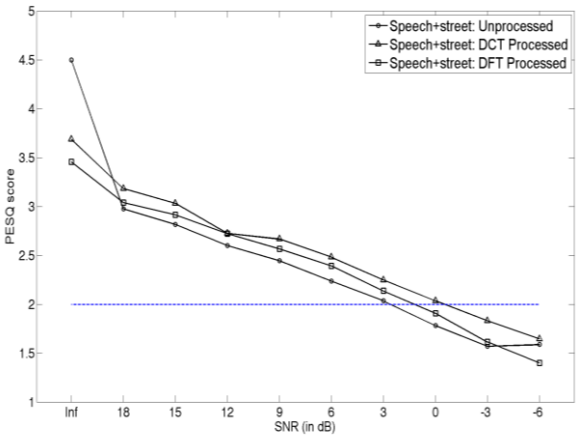
(a) Babble noise



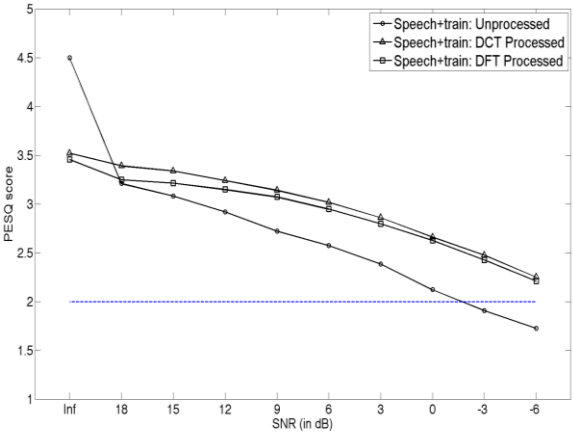
(b) Car noise



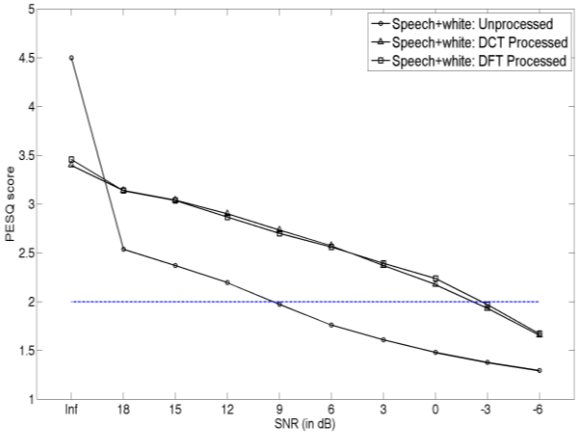
(c) Pink noise



(d) Street noise

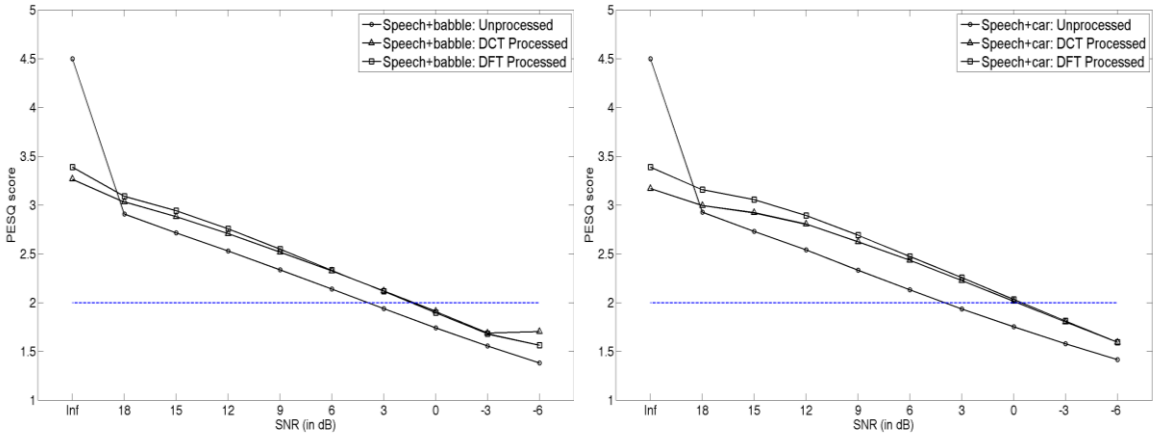


(e) Train noise



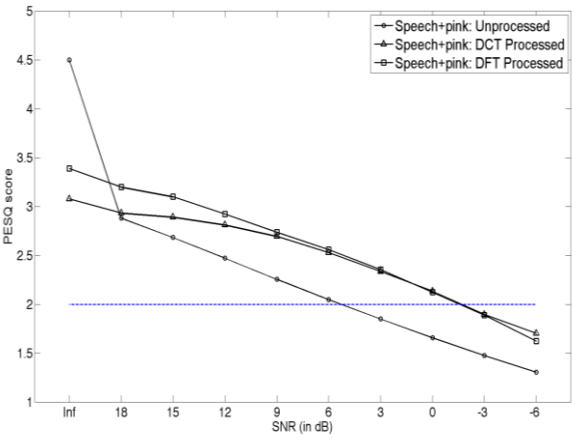
(f) White noise

Figure 4.5. PESQ scores for the enhanced noisy speech, DCT (β=0.01, γ=1), DFT (β=0.01, γ=1) at optimal values of α. Speech material : VHSES with babble, car, pink, street, train and white noises at various SNRs.

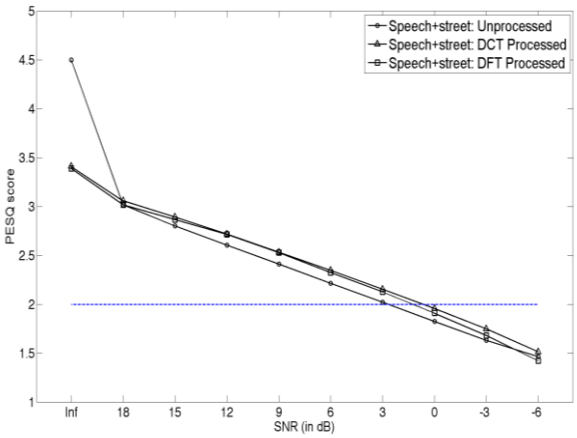


(a) Babble noise

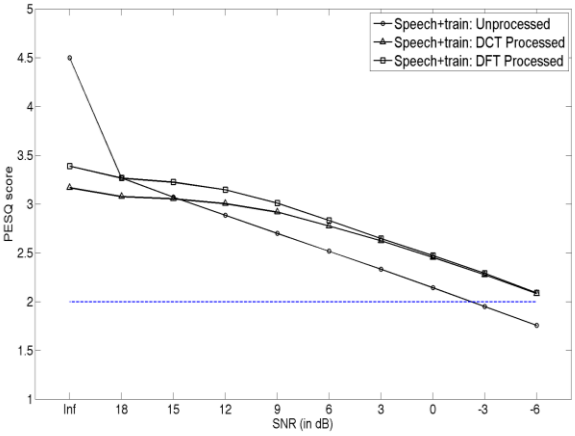
(b) Car noise



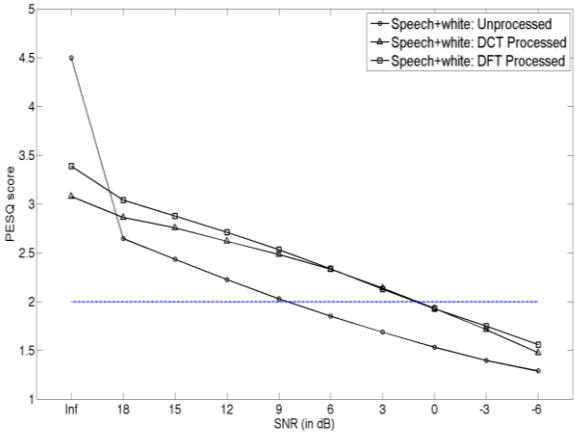
(c) Pink noise



(d) Street noise



(e) Train noise



(f) White noise

Figure 4.6. PESQ scores for the enhanced noisy speech, DCT ($\beta = 0.001$, $\gamma = 1$), DFT ($\beta = 0.01$, $\gamma = 1$) at optimal values of α. Speech material : NOIZEUS with babble, car, pink, street, train and white noises at various SNRs.

Table 4.14 (a) SNR advantage obtained using 3-point 4-stage CMBNE and spectral subtraction using DCT and DFT using optimal value of α, and β = 0.01 and γ = 1, Speech material : VHSES with various noises.

Noise	Optimal α		SNR advantage	
	DFT	DCT	DFT	DCT
Babble	2.0	3.0	4.0	4.0
Car	2.5	2.5	7.0	7.0
Pink	2.5	4.0	13.0	12.5
Street	2.5	2.0	2.0	4.0
Train	2.5	3.0	6.0	6.5
White	2.5	4.0	12.0	11.5

Table 4.14 (b) SNR advantage obtained using 3-point 4-stage CMBNE and spectral subtraction using DCT and DFT against optimal value of α, and for DCT β = 0.001, for DFT β = 0.01, γ = 1, Speech material : NOIZEUS with various noises.

Noise	Optimal α		SNR advantage	
	DFT	DCT	DFT	DCT
Babble	1.5	2.0	3.0	3.0
Car	1.5	2.5	4.0	4.0
Pink	1.5	3.0	7.0	7.0
Street	1.5	1.5	2.0	2.5
Train	1.5	2.5	5.0	5.0
White	1.5	3.0	8.0	8.0

This page is intentionally left blank

Chapter 5

REAL-TIME IMPLEMENTATION

5.1 Introduction

The DCT-based spectral subtraction technique is implemented for real-time processing on DSP Board, “Spectrum Digital eZdsp USB stick” [24] with 16-bit fixed point processor TI-TMS320C5515 [25] and a maximum clock frequency of 120 MHz. 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. The DSP-board has 4MB flash memory for user program. The hardware includes an audio codec TLV320AIC3204 [26] with stereo ADC and DAC supporting 16/20/24/32-bit quantization and sampling frequency of 8−192 kHz. It also has four 4-channel DMA controllers, three 32-bit timers, tightly coupled FFT hardware accelerator for efficient computation of 8−1024 point FFT. The block diagram of the DSP board is shown in Figure 5.1.

5.2 Implementation Details

The program was written in C, using TI's 'Code Composer Studio, ver. 4.0 as the integrated development environment. For reducing conversion overheads, the input samples, spectral values, and the processed samples are all stored as 4-byte words, with 16-bit real and 16-bit imaginary parts. The imaginary part of input sample is set to zero.

A block diagram of implementation of DCT based spectral subtraction on the DSP board with L-sample window and N-point FFT ($L = 300$, $N = 512$) is shown in the Figure. 5.2. Codec and DMA are used to continuously acquire and output the speech signal. The data transfer and buffering operations are shown in Figure 5.3. At a sampling frequency of 10 kHz, DMA channel 2 is programmed to acquire the ADC values into an input cyclic buffer which is divided into 3 memory blocks of size $S = L/2$ to facilitate input windowing with 50% overlap. Samples from the just-filled and previous input blocks which together form the desired window of length L, are copied to the input data buffer. The input buffer is of the size N, with the remaining $N−L$ words set to zero. Pointers initialized as (..1,2,3,1..) are used and cyclically updated to monitor the just-filled and the current-input blocks. A DMA interrupt is generated when a block gets filled. A 2 memory block cyclic buffer is employed to write the processed output to the DAC. The current output and write-to output blocks are tracked by

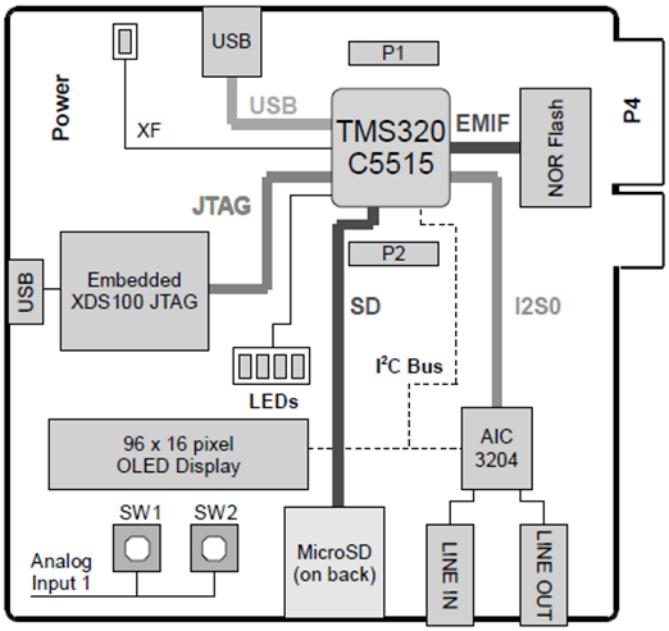


Figure 5.1. Block diagram of TMS320C5515 eZdsp USB Stick [24].

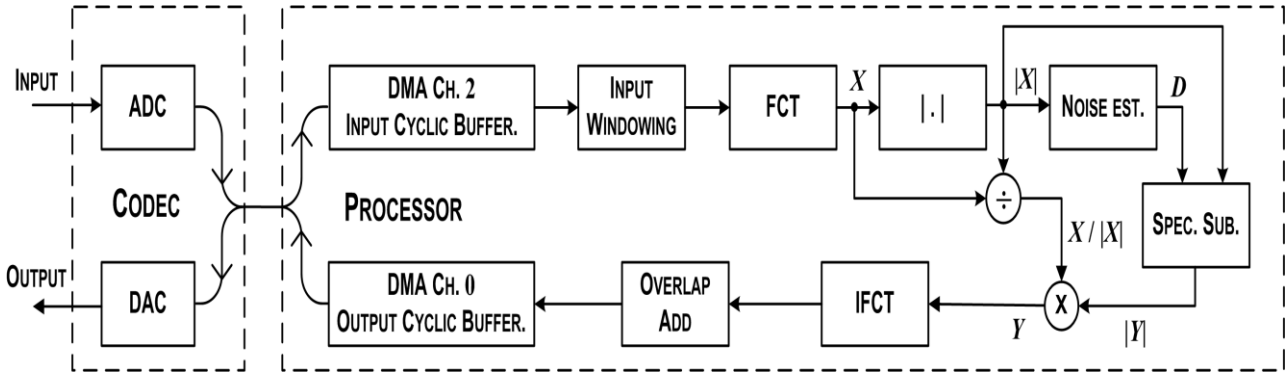


Figure 5.2. Implementation level diagram of spectral subtraction on DSP board, adapted from [22].

pointers with toggling values of 1 and 2. These are devised for an efficient realization of analysis-synthesis with 50% overlap. The resynthesis is done using overlap-add.

FFT-based DCT described in section 2.2 is implemented by making use of the hardware FFT accelerator on chip. The input buffer serves as the input to *N*-point FCT. All the twiddle factors to be multiplied for computing a DCT and IDCT are generated using Matlab and are stored in arrays in the program memory, keeping the number of multiplications needed to be done in real time as low as possible. The code is optimized by exploiting the conjugate symmetry property of FFT and avoiding extra multiplications in DCT computation. The operations are performed with numbers in Q15 format owing to the fixed-point processor and intrinsic assembly functions of the processor provided by the manual [25] like saturate multiply, saturate add, saturate subtract are employed as essential to handle overflow. Processed values are scaled up/down to prevent overflows and underflows at potential points in the code.

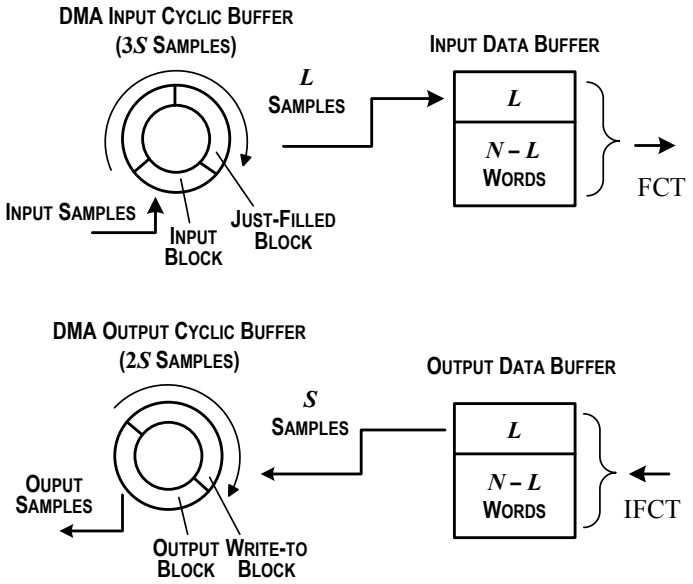


Figure 5.3. Working Of DMA for Buffering and Data Transfer Operations (S=L/2), adapted from [22].

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 rectangular window with 50% overlap at a sampling frequency of 10 kHz, L = 300, and N = 512. Noisy phase is used for the reconstruction of the spectrum. For the experiment, α is kept at 2.5 and β at 0.01 and in real-time processing code, these two factors are defined as macros, assigned with values as ALPHA = round (8α), BETA = $1/\beta$.

5.3 Test results

For testing the implementation on the DSP board, the input signal for processing was generated from a PC sound card and given to the DSP board through one channel of its stereo-in audio connector. The processed output signal from one channel of stereo-out audio connector was acquired through a notebook PC sound card (not connected to AC mains) as shown in Figure 5.4. The DSP board is powered through its USB port connected to the desktop PC. Use of two PCs whose grounds are not connected avoids formation of ground loop and associated noise.

An example showing the clean, noisy and processed signals and corresponding spectrograms processed using CMBNE and DCT based spectral subtraction through offline and real-time implementations are shown in Figure 5.5 for comparison. Real-time processed output was found to be similar to the corresponding offline processed output and there was no perceptible difference in processed outputs obtained using DFT and DCT. Measurements using a DSO showed that real-time processing introduced a signal delay of 49 ms. Out of this delay, the processing delay (algorithmic delay of one frame and computational delay of 0.5 frame) contributed 45 ms and the remaining delay is attributed to audio input-output latency

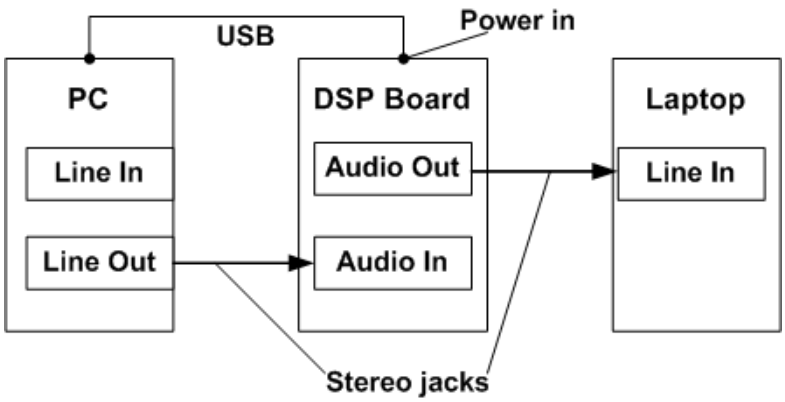


Figure 5.4. Setup for giving input and recording output from DSP board.

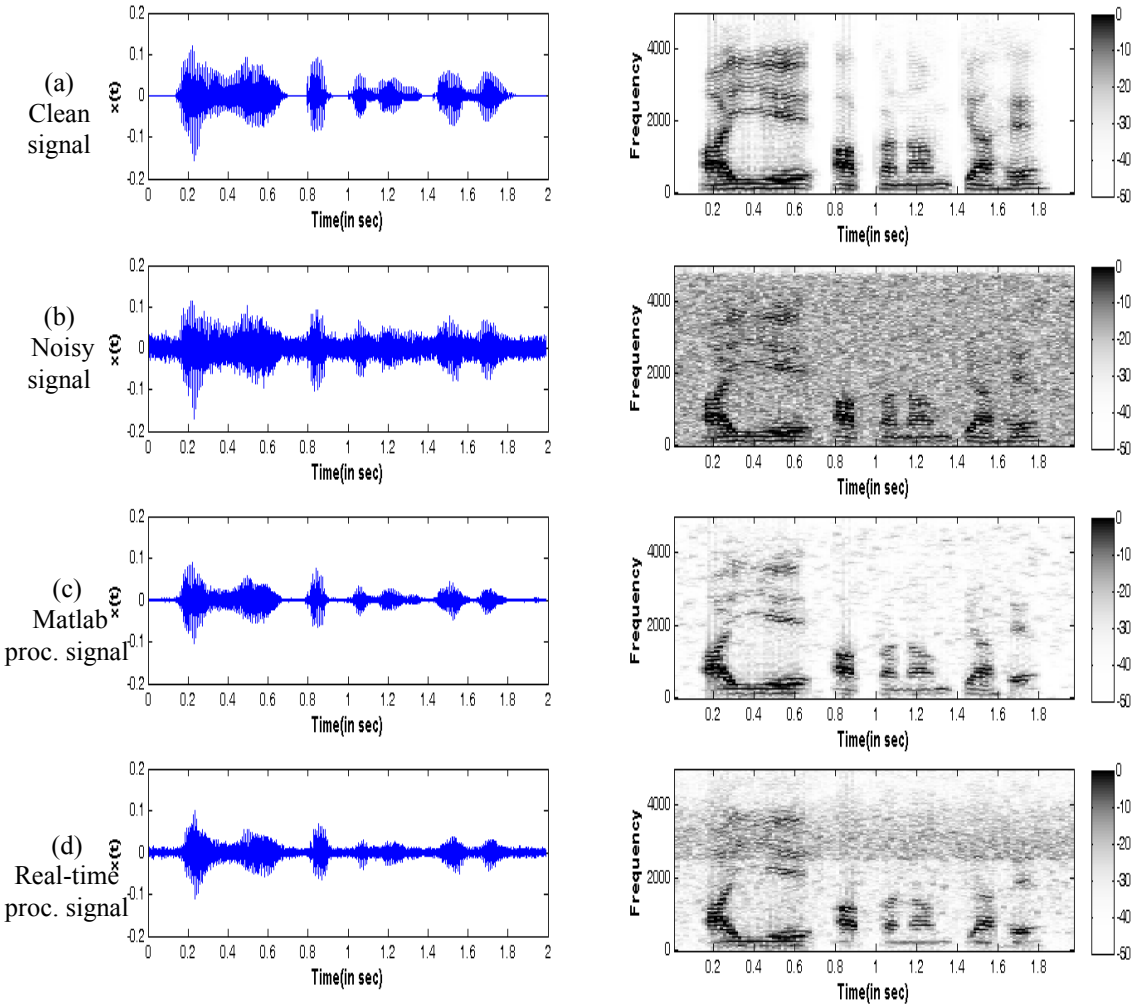


Figure 5.5. Signals and spectrograms: Processing of "Where were you a year ago?," from a male speaker, with white noise at 3 dB SNR.

of the DSP-board hardware. The algorithm is satisfactorily implemented for a clock frequency down to 20 MHz. Hence it is informed that the implementation needs 1/6th of the available processing capacity providing scope for integrating other signal processing steps like dynamic range compression etc in hearing aids and other audio communication devices.

Chapter 6

SUMMARY AND CONCLUSION

The feasibility of using DCT for real-time implementation was investigated through the algorithm of speech enhancement using cascaded median based spectral subtraction and a comparison has been made with DFT-based processing. Offline investigations on DQTNE are carried out to identify the best suitable quantile for noise estimation. A major drawback with DCT as reported in [27] is that the processing is highly dependent on window position. A pitch synchronous implementation can address this issue, but such an implementation is unsuited for real-time processing in hearing aids. The performances with a fixed shift but different noise estimation techniques and quantiles are explored. It was found that median-based noise estimation tracks the noise spectrum well at different SNR values. Hence an approximation of the median which is the cascaded-median is used for noise estimation. Investigations were carried out to find out the optimum window length and noise estimation duration. Window of length 30 ms and noise estimation using a 3-point 4-stage cascaded median were chosen for the processing. Investigations showed that use of clean phase provided up to 0.2 improvement in PESQ scores. However, phase enhancement methods which are compatible with real-time processing were not successful in improving speech quality. Future work may be carried out on minimizing phase discontinuities.

Further improvement may be achieved by using an adaptive over-subtraction factor in the spectral subtraction algorithm to use SNR-dependent and frequency-dependent factors. Processing the noisy speech in the SNR range of -6 dB to 18 dB resulted in an SNR advantage of 3−13 dB. An improvement of 0.24−0.7 and 0.13−0.47 in PESQ scores was achieved using different noises at 0 dB for speech materials VHSES and NOIZEUS respectively.

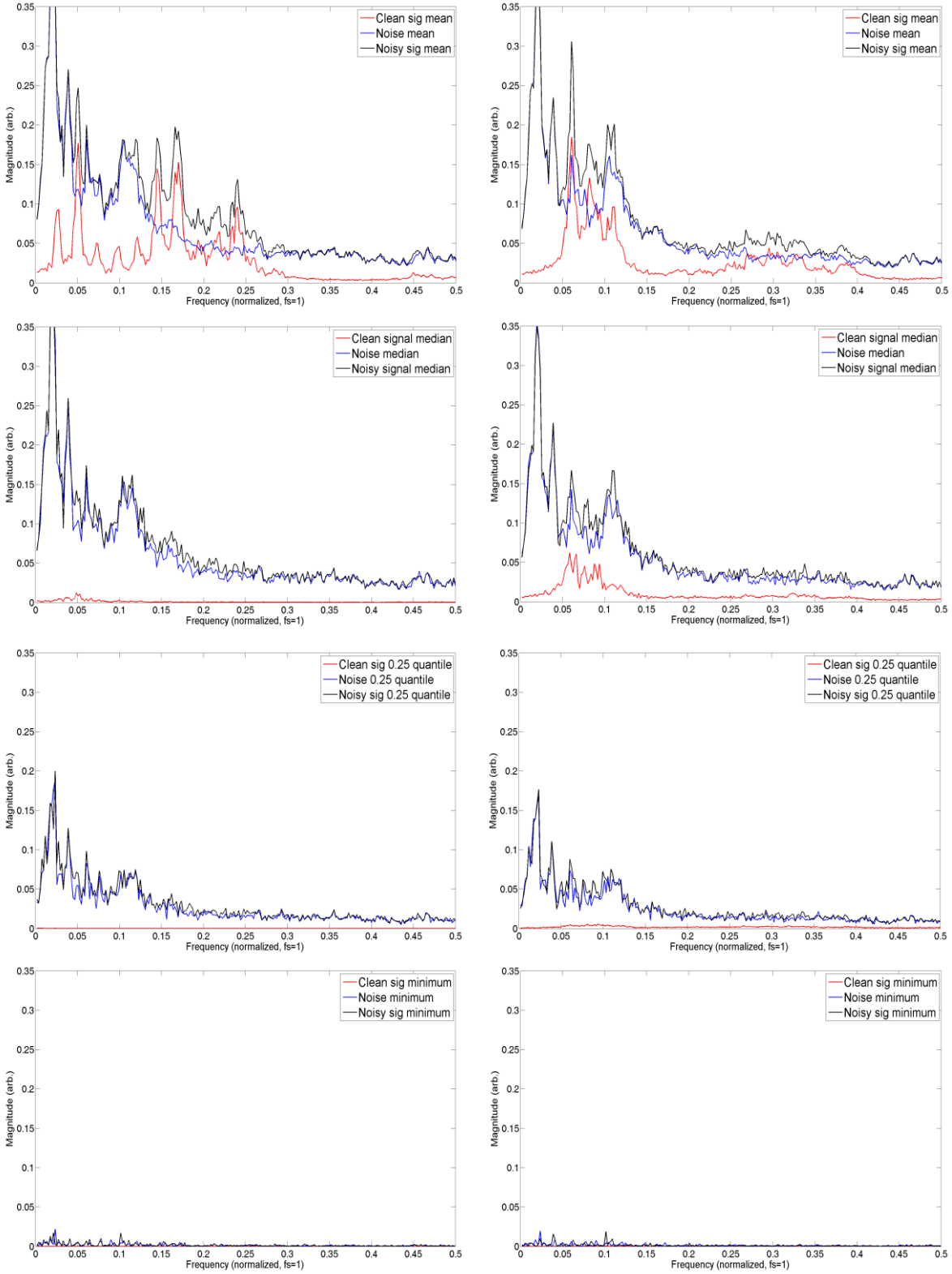
For real-time operation, the method was implemented on a 16-bit fixed point processor TMS320C5515 using the DSP board. Codec and DMA were used at a sampling rate of 10 kHz for continuous acquisition of the input signal and outputting of the processed signal. Use of FCT algorithm facilitated the application of hardware accelerator for the real-time processing. The data transfer and buffering operations were devised for an efficient realization of analysis-synthesis with 50 % overlap. The real-time processing with analysis window length of 30 ms and 512-point FFT was implemented, using about one-sixth of the

computing capacity of the processor, with a processing delay 49 ms, making it suitable for hearing aid applications. Informal listening tests showed that the processed output from the DSP board was perceptually similar to the corresponding output from the offline implementation for speech as well as other audio signals.

This page is intentionally left blank

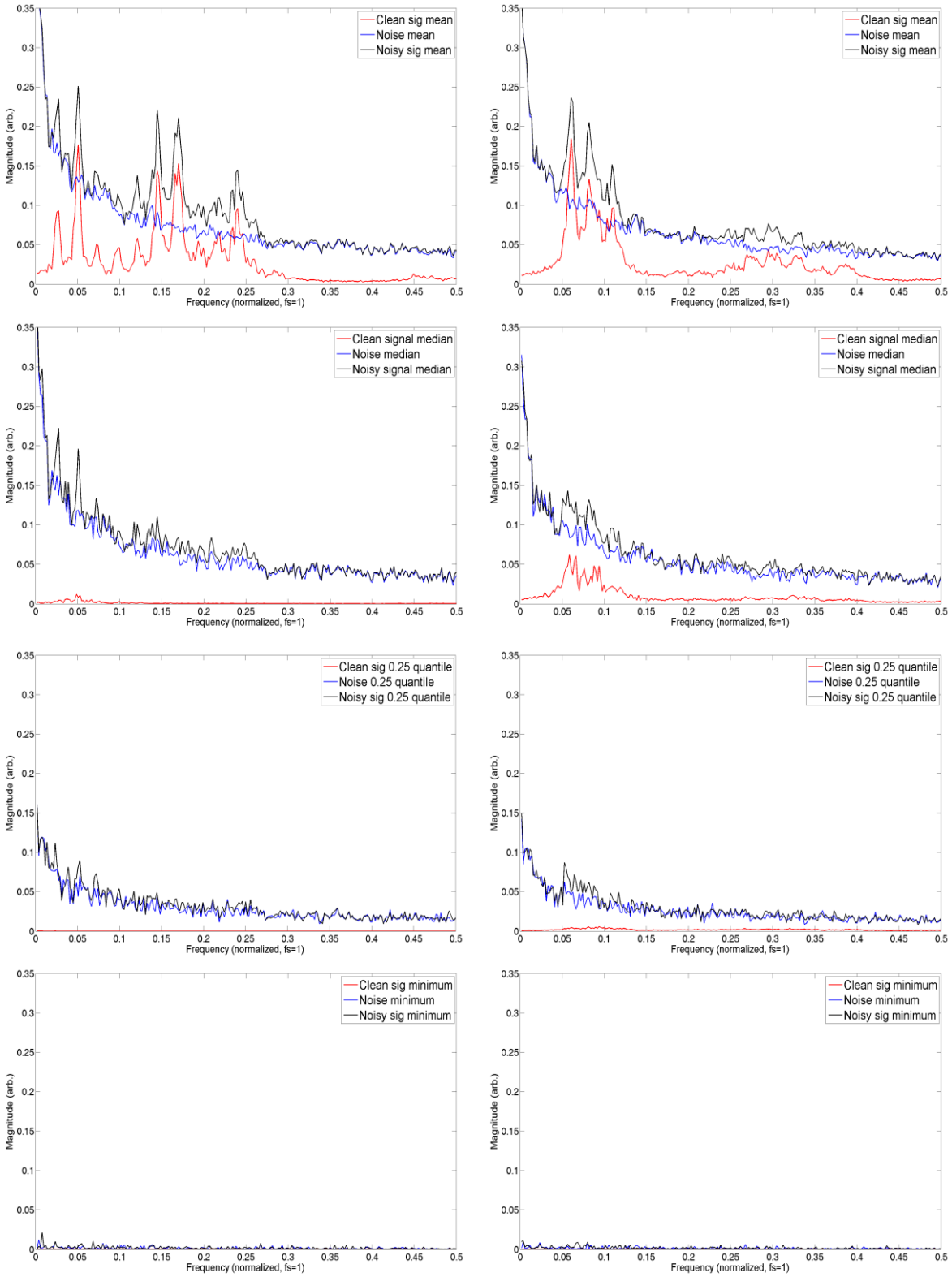
Appendix A

INVESTIGATIONS ON NOISE ESTIMATION



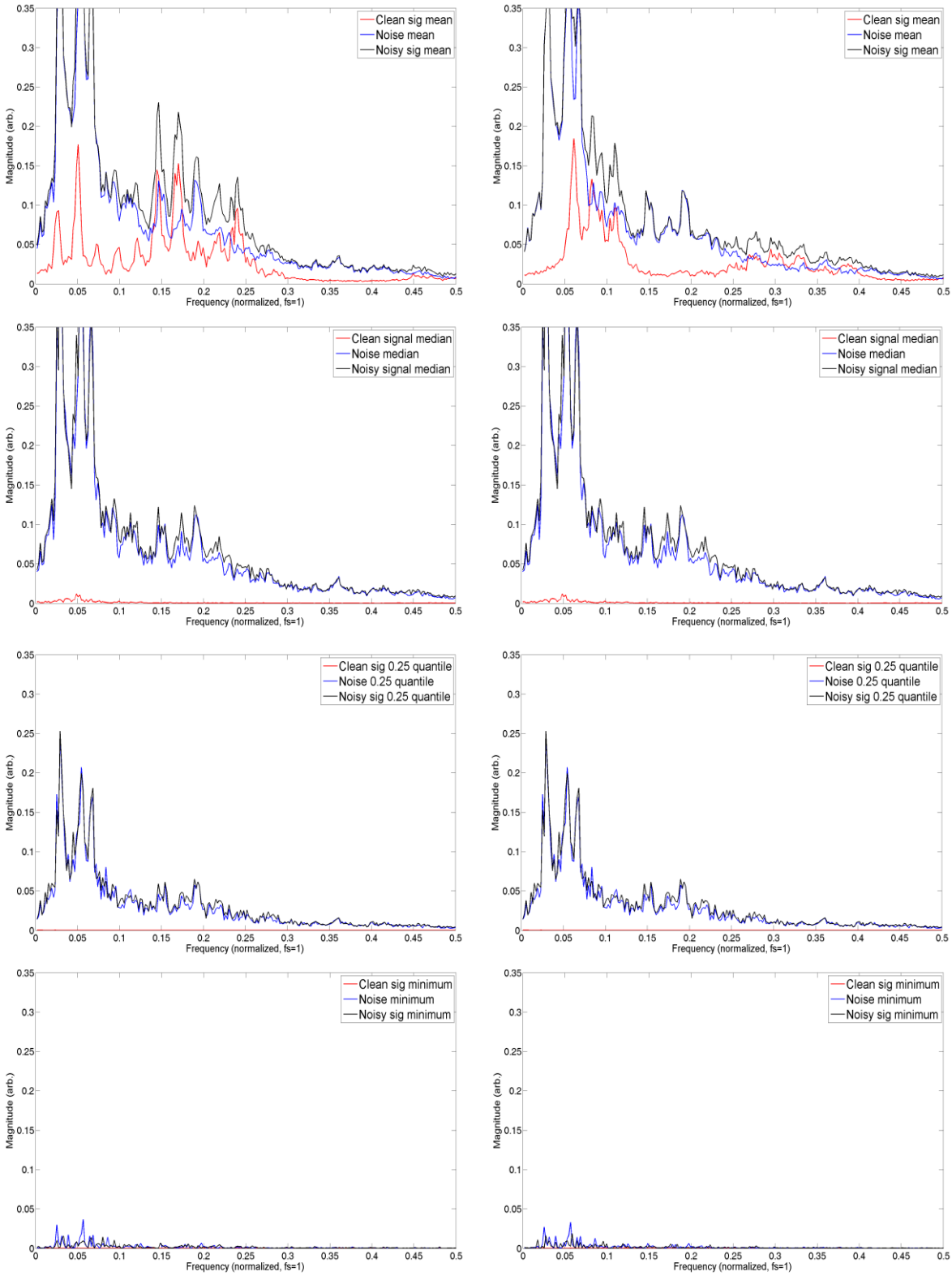
(a) VHSES (b) NOIZEUS

Figure A.1. Mean, median, 0.25-quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (car, 0 dB)



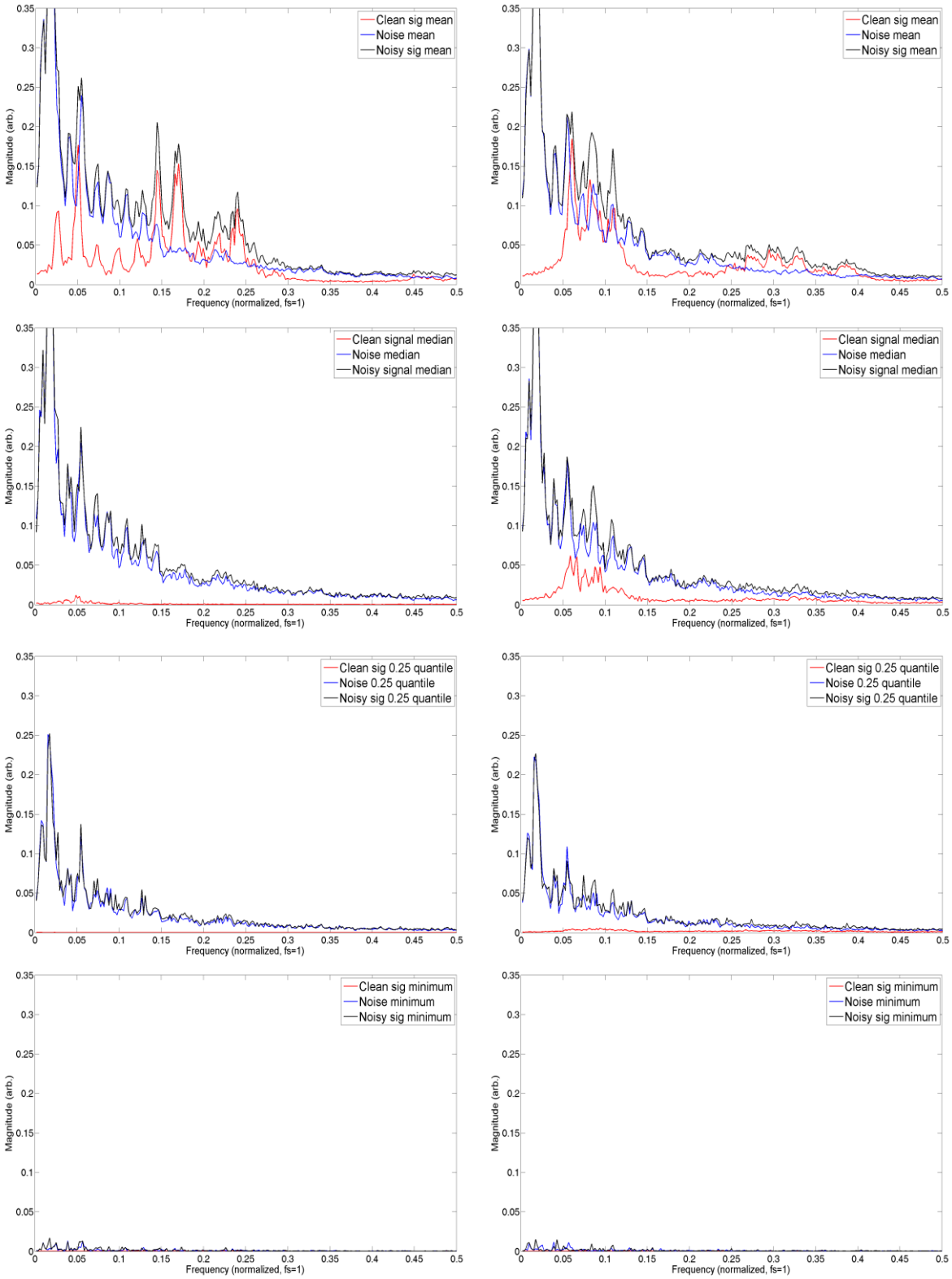
(a) VHSES (b) NOIZEUS

Figure A.2. Mean, median, 0.25-quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (pink, 0 dB)



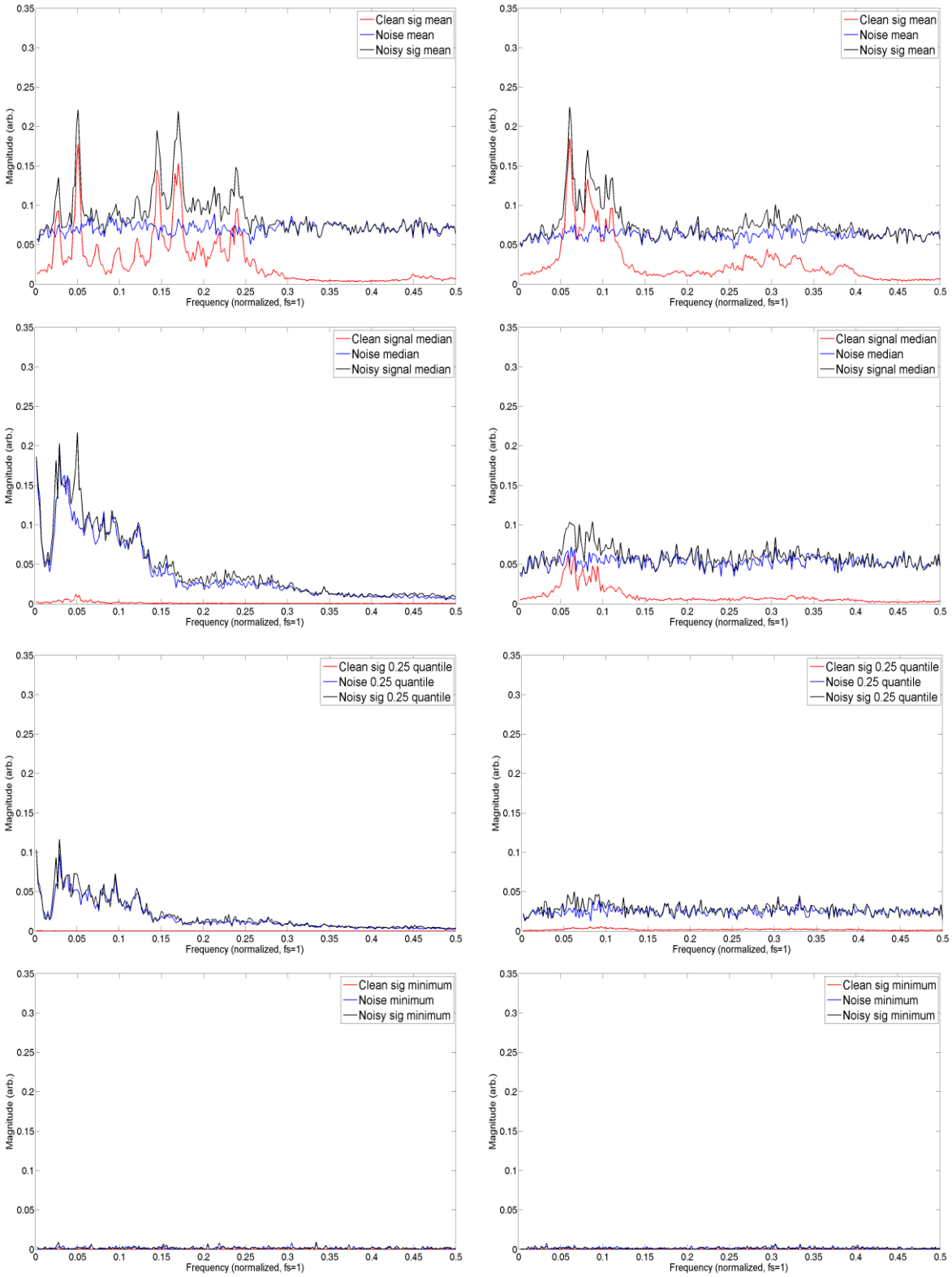
(a) VHSES (b) NOIZEUS

Figure A.3. Mean, median, 0.25-quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (street, 0 dB)



(a) VHSES (b) NOIZEUS

Figure A.4. Mean, median, 0.25-quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (train, 0 dB)



(a) VHSES (b) NOIZEUS

Figure A.5. Mean, median, 0.25-quantile and minimum of magnitude spectra of clean speech signal, noise and noisy speech (white, 0 dB)

Appendix B

ROBUSTNESS OF DCT TO PHASE INVERSIONS

To understand the sensitivity of DCT to phase changes and its impact on the PESQ scores, an experiment was conducted. Random phase inversions were introduced in the transform domain of the clean speech material and the resultant phase was used for resynthesis of the signal. Processing was carried out with rectangular window of 30 ms and 512-point DCT with 50% and 75% overlaps. Effect of these modifications depending on the percentages of samples modified per frame are given in Table B.1.

Yet another experiment was conducted to see the effect of phase changes with respect to the magnitude of the signal. Here, the phase inversions were concentrated in coefficients with higher magnitudes and the PESQ scores were noted. For example, Out of the overall samples in a window (number = 300), effect of inverting top 10% of all the samples (number = 30) and 20% of the top 150 samples (number = 30) were compared. As can be expected, modifications of phase made to higher magnitudes degraded the scores more than the changes distributed over all magnitudes. Results related to the same are shown in Table B.2. The experiment was also performed on a piece of music material and it was observed that the effects were adverse. The resultant PESQ scores were ~1.1 lower than the corresponding scores obtained with the VHSES material.

A method was devised to reduce the noise-related phase discontinuities in DCT-based analysis-synthesis. The processing steps are as the following :

1) Sort the enhanced magnitude coefficients of the current window in descending order.
2) Mark the co-efficient values and frequencies of first 10% of them.
3) If any of these frequencies occurred in the previous frame, assign the sign of that particular coefficient in the previous frame to the corresponding coefficient in the current frame. For all other frequency components, retain the phase. For the first frame noisy phase is retained.

The similarity of phase estimated by the proposed method and the noisy phase to the original phase are compared and the number of phase inversions are noted. Phase inversions are the number of times the phases differ between the spectra of the signals under comparison. Table B.3 gives the phase inversions for frequencies in auditory critical bands 7 and 8 which correspond to frequency range 0.63 − 0.92 kHz [29] where there was considerable speech energy for most of the time frames i.e. frequency components 27 − 39 of the 512 length transform. It is evident that the noisy phase is closer to the original phase than the phase modified using proposed method.

Table. B.1 PESQ scores showing sensitivity of DCT-based analysis-synthesis to phase changes as percentage of inversions per frame. Speech material : VHSES

Percentage of inversions/ frame	PESQ scores (0 − 4.5)			
	Rect., 50% overlap	Rect., 75% overlap	Hamming, 50% overlap	Hamming, 75% overlap
0	4.49	4.49	4.48	3.04
10	3.2	3.07	3.11	2.79
20	2.78	2.62	2.79	2.55
25	2.66	2.47	2.68	2.48
40	2.37	2.18	2.46	2.24
50	2.3	2.14	2.42	2.21
60	2.35	2.19	2.44	2.26
75	2.64	2.49	2.69	2.45
80	2.79	2.6	2.79	2.53
90	3.17	3.02	3.19	2.8
100	4.49	4.49	4.48	3.04

Table B.2 PESQ scores showing sensitivity of DCT-based analysis-synthesis to phase changes with respect to magnitude as percentage of inversions per frame. Speech material : VHSES, window : Rect., 50% overlap

No. of samples inverted	PESQ scores			
	Top 10% samples	Top 25% samples	Top 50% samples	Top 100% samples
0	4.49	4.49	4.49	4.49
15	2.31	2.99	3.43	4.15
30	1.83	2.47	3.01	3.20
60	-	1.91	2.50	2.78
75	-	1.80	2.27	2.66
120	-	-	1.90	2.37
150	-	-	1.82	2.30
300	-	-	-	4.49

Table B.3 Number of phase inversions occurring as compared to that of the clean speech segment "Where were you a year ago?" from material VHSES, for a particular frequency component in the range 27−39. (Out of total number of frames - 1668)

Frequency component	No of phase inversions								
	3 dB			0 dB			-3 dB		
	NM + NP	CM + MP	MM + MP	NM + NP	CM + MP	MM + MP	NM + NP	CM + MP	MM + MP
27	421	408	586	482	476	604	502	518	605
28	457	467	553	504	502	558	565	569	635
29	546	547	571	588	592	600	627	647	649
30	603	630	635	655	670	672	693	700	719
31	582	578	618	638	643	665	702	704	740
32	596	589	622	653	651	669	658	659	689
33	622	640	627	679	676	684	692	691	702
34	603	604	623	654	657	685	691	691	714
35	528	541	571	619	624	644	642	644	676
36	507	504	567	548	549	603	608	616	640
37	471	493	581	487	513	588	580	584	650
38	481	487	593	527	522	623	577	583	663
39	509	513	637	552	556	657	604	591	673

To explicate the results, the continuity of phase in clean signal has been observed. Out of a total of 1668 frames that corresponded to the input clean speech segment, there are a total of 1002 inversions in the frequency component 29 and a similar trend is carried in the other frequency components. Thus our approach of smoothening the phase did not have positive results.

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process. 1979 (IEEE ICASSP '79)*, vol. 4, Washington, D. C., pp. 208−211.

[2] 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.

[3] V. Stahl, A. Fisher, and R. Bipus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process. 2000 (IEEE ICASSP '00)*, vol. 3, Istanbul, Turkey, pp. 1875−1878.

[4] R. Martin, "Spectral subtraction based on minimum statistic," in *Proc. 7th Eur. Signal Process. Conf. (EUSIPCO-94)*, Edinburgh, Scotland, 1994, pp. 1182−1185.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process*, vol. 9, no. 5, pp. 504−512, 2001.

[6] I. Y. Soon, "Transform based speech enhancement techniques," PhD thesis, School of Elect. and Electron. Eng., Nanyang Technological Univ., Singapore, 2002.

[7] V. Britanak, "Discrete cosine and sine transforms" in *The Transform and Data Compression Handbook*, K. R. Rao et al., Eds. Boca Raton, Florida: CRC, 2001.

[8] J. Makhoul, "A fast cosine transform in one and two dimensions," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 27−34, 1980.

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

[10] P. Vary, "Noise suppression by spectral magnitude estimation—mechanism and theoretical limits," *Signal Process.*, vol. 8, no. 4, pp. 387−400, 1985.

[11] R. J. Pinnell, "Adaptive transform coding of speech signals," B. Eng. thesis, McGill Univ., Montreal, Canada, 1982.

[12] S. Zhao, "Performance analysis and enhancements of adaptive algorithms and their applications," PhD thesis, School of Elect. and Electron. Eng., Nanyang Technological Univ., Singapore, 2009.

[13] S. K. Basha and P. C. Pandey, "Real-time enhancement of electrolaryngeal speech by spectral subtraction," in *Proc. 18th Nat. Conf. Commun. (NCC-2012)*, Kharagpur, India, 2012, pp. 516−520.

[14] S. D. Kamath, "A multi-band spectral subtraction method for speech enhancement," M. Sc. thesis, Dept. Elect. Eng., Univ. Texas, Dallas, 2001.

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

[16] V. Stahl, A. Fisher, and R. Bipus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process. 2000 (IEEE ICASSP '00)*, vol. 3, Istanbul, Turkey, pp. 1875−1878.

[17] N. Tiwari and P. C. Pandey, "Speech enhancement using noise estimation based on dynamic quantile tracking for hearing impaired listeners," in *Proc. 21st Nat. Conf. Commun. (NCC-2015)*, Mumbai, India, 2015, paper no. 1570056299.

[18] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. 1995 (IEEE ICASSP-95)*, Detroit, MI, pp. 153−156.

[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229−238, 2008.

[20] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", *Speech Commun.,* vol. 49, no. 7, pp. 588−601, 2007.

[21] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. 6th Int. Conf. Spoken Language Process. (ICSLP 2000)*, Beijing, China, 2000, pp. 29−32.

[22] S. K. Waddi, "Real-time enhancement of noisy speech using spectral subtraction," M. Tech. dissertation, Dept. Elect. Eng., IIT Bombay, Mumbai, India, 2013.

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

[24] Spectrum Digital, Inc., "TMS320C5515 eZdsp USB Stick Technical Reference," 2010, [Online]. Available: support.spectrumdigital.com/boards/usbstk5515/reva/files/usbstk5515_TechRef_RevA.pdf.

[25] Texas Instruments, Inc., "TMS320C5515 Fixed-Point Digital Signal Processor," 2011, [Online]. Available: focus.ti.com/lit/ds/symlink/tms320c5515.pdf.

[26] Texas Instruments, Inc., "TLV320AIC3204 Ultra Low Power Stereo Audio Codec," 2008, [Online]. Available: focus.ti.com/lit/ds/symlink/tlv320aic3204.pdf.

[27] H. Ding, I. Y. Soon, and C. K. Yeo, "A DCT-based speech enhancement system with pitch synchronous analysis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 5, pp. 2614−2623, 2011.

[28] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. 13th Int. Conf. Spoken Language Process. (INTERSPEECH 2012)*, Portland, U. S., pp. 1−4, 2012.

[29] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, pp. 248−248, 1961.

This page is intentionally left blank

Acknowledgements

I would like to express my sincere gratitude towards my guide Prof. P. C. Pandey for giving me the opportunity to work under his esteemed guidance and to explore new concepts. I am thankful to him for sparing his invaluable time to help me understand and implement the project and for the support he gave me during the project.

I am much obliged to Nitya Tiwari for the insightful views she has provided all through the project. I am thankful to all my friends and seniors in SPI lab for sharing interesting discussions and for their whole-hearted support during the tenure of this project. I would like to thank Vidyadhar Kamble for helping me in lab related issues.

Finally, I'm indebted to my family and friends for their unconditional love and encouragement in every phase of my life.

Madhu Lekha

June 2016