12-2013

Speech Enhancement By Exploiting The Baseband Phase Structure Of Voiced Speech For Effective Non-Stationary Noise Estimation

Sanjay Patil
Clemson University, sanjayp@g.clemson.edu

Speech Enhancement By Exploiting The Baseband Phase Structure Of Voiced Speech For Effective Non-Stationary Noise Estimation

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Electrical Engineering

---

by
Sanjay Patil
December 2013

Accepted by:
Dr. John Gowdy, Committee Chair
Dr. Adam Hoover
Dr. Richard Groff

ABSTRACT

Speech enhancement is one of the most important and challenging issues in the speech communication and signal processing field. It aims to minimize the effect of additive noise on the quality and intelligibility of the speech signal. Speech quality is the measure of noise remaining after the processing on the speech signal and of how pleasant the resulting speech sounds, while intelligibility refers to the accuracy of understanding speech. Speech enhancement algorithms are designed to remove the additive noise with minimum speech distortion.The task of speech enhancement is challenging due to lack of knowledge about the corrupting noise. Hence, the most challenging task is to estimate the noise which degrades the speech. Several approaches has been adopted for noise estimation which mainly fall under two categories: single channel algorithms and multiple channel algorithms. Due to this, the speech enhancement algorithms are also broadly classified as single and multiple channel enhancement algorithms.In this thesis, speech enhancement is studied in *acoustic* and *modulation* domains along with both *amplitude* and *phase* enhancement. We propose a noise estimation technique based on the spectral sparsity, detected by using the harmonic property of voiced segment of the speech. We estimate the frame to frame phase difference for the clean speech from available corrupted speech. This estimated frame-to-frame phase difference is used as a means of detecting the noise-only frequency bins even in voiced frames. This gives better noise estimation for the highly non-stationary noises like babble, restaurant and subway noise. This noise estimation along with the phase difference as an additional prior is used to extend the standard spectral subtraction algorithm. We also verify the effectiveness of this noise estimation technique when used with the Minimum Mean Squared Error Short Time Spectral Amplitude Estimator (MMSE STSA) speech enhancement algorithm. The combination of MMSE STSA and spectral subtraction results in further improvement of speech quality.

ACKNOWLEDGMENTS

I thank Dr. John N. Gowdy for introducing me to the field of speech enhancement, and for his guidance, suggestions and providing all of the databases and material required during the course of this work.

I thank Dr. Adam W. Hoover and Dr. Richard E. Groff for serving on my advisory committee.

I would like to take this opportunity to thank my family (dad,mom and brother) for their love and encouragement to pursue my dreams. I would also like to thank my friends and group mates (Sujit and Shamama) for all the technical discussions that we had during the course of this thesis work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Speech signals from uncontrolled environment may contain degradation components along with the natural speech components. The degradation components include background noises (train-noise, machine-gun noise etc.), speech from other speakers, etc. Speech degraded by additive noise makes listening difficult and gives poor performance in automatic speech processing tasks like speech recognition, speaker identification, hearing aids, speech coders, etc. Consequently, it is desirable to develop speech enhancement technique to minimize the influence of noise with minimum speech distortion. This scenario is pictorially shown in figure 1.1

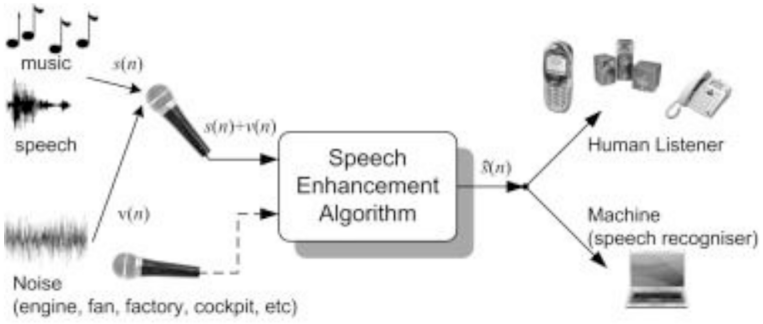


Figure 1.1: Scenario for speech enhancement. Taken from [1]

Speech enhancement algorithms aim to improve the quality and/or intelligibility of noisy speech. Speech quality relates to the ease of listening and listening comfort while the intelligibility is related to the word error rate of the perceived speech. It has been shown in [2] that the noise

reduction algorithms which try to increase the speech quality mostly fail to improve the speech intelligibility due to inaccurate noise estimation. Hence, noise estimation is the most important and challenging stage in a speech enhancement algorithm. In general, a speech enhancement algorithm consists of three major steps as given below:

1. Transform time domain noisy speech to frequency domain.
2. Estimate the amount of noise added to the clean speech.
3. Use the noise estimate to process the noisy speech.

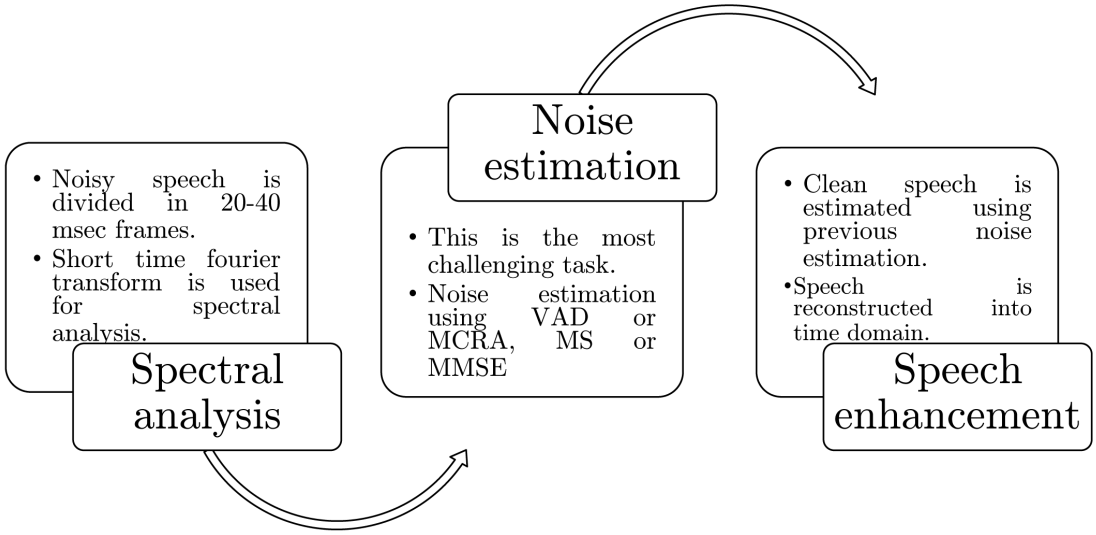


Figure 1.2: Typical speech enhancement algorithm.

Various approaches [3, 4, 5, 6, 7] can be used to estimate the noise trajectory in the spectral domain. Accurate noise estimation is critical for better performance of speech enhancement algorithm. For the reference algorithms in this thesis, noise estimation is carried out using 'Voice Activity Detector'(VAD) due to its simplicity.

The problem of speech enhancement in presence of additive noise has received considerable attention in the literature since the mid-1970 [3]. Various approaches exist to improve the quality and intelligibility of speech signal. Those approaches can be classified based upon various criteria as discussed below:

Various ways to classify the existing algorithms-

- Single channel or multi-channel depending on number of available microphones [8, 9].
- Time domain or spectral domain algorithms [10, 11].
- Inventory based algorithms.(HMMs or Code-books are used to model speech and noise characteristics) [12, 13, 14].

Furthermore, single channel speech enhancement algorithms are classified as:

- Spectral subtraction [3].
- Statistical based algorithms. (Minimum mean squared error algorithms like the Wiener filter and Short Time Spectral Amplitude (STSA) estimator) [15, 16, 17].
- Subspace based algorithms. (For example -Decomposition of noisy speech into speech and noise subspaces using SVD) [18, 19].

The choice of the algorithm depends on the application and the problem issued. We may process the speech for a human listener in order to improve its quality (e.g., in noisy environments such as offices, streets, and motor vehicles), or to improve its intelligibility in harsh conditions (such as airports). Transcription of recorded tapes degraded by additive noise is also of interest. We may use speech enhancement as a preprocessing mechanism for speech compression algorithms or as a front-end to Automatic Speech Recognition (ASR) systems.

In this thesis, we propose the single-channel noise estimation algorithm. When this algorithm is combined with the existing speech enhancement algorithm, perceptual speech quality is improved as confirmed by Perceptual Evaluation of Speech Quality (PESQ) score. The noise is assumed to be additive. The improvement is verified against babble, restaurant and subway noises.

Chapter 2

OVERVIEW OF SPEECH ENHANCEMENT TECHNIQUES

Typical single-channel speech enhancement methods make two assumptions about the observed noisy speech signal: (1) the underlying clean speech and the additive noise are uncorrelated and (2) noise statistics vary slower than the speech statistics. The signal model for single-channel speech enhancement scheme is shown in figure below:

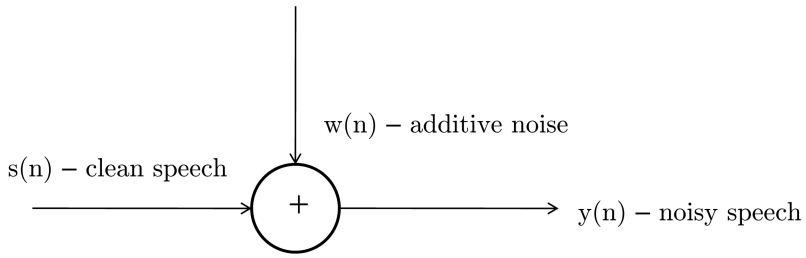


Figure 2.1: The signal model for single channel speech enhancement shows speech and the additive noise.

Some basic speech enhancement algorithms are: *spectral subtraction* [3], *Wiener filter* [20], *Minimum Mean Square Error* [15] and some recent advancements in this field like *spectral subtraction in modulation domain* [21] and *Phase estimation based speech enhancement* [22] are explained in the following sections.

2.1 Spectral Subtraction

Spectral subtraction [3] is historically the first algorithm proposed to reduce the noise from the speech signal. It is based on the simple noise reduction technique: the estimated noise spectrum is subtracted from the noisy speech to obtain the estimate of the clean speech signal. The noise is estimated from the initial 10-15 noisy speech segments in which speech is assumed to be absent and this estimate is updated accordingly whenever a speech-absent segment is observed in future. The noise is assumed to be varying slowly and not changing significantly between updating periods. This processing of the noise reduction is carried out in the frequency domain. Once noise is subtracted from the noisy speech, the enhanced speech is reconstructed using inverse Fourier transform and overlap-add technique [23].

2.1.1 Mathematical Formation of Spectral Subtraction Algorithm

Assume that y(n), the noisy(noise-corrupted) input signal, is composed of the clean speech signal s(n) and the additive noise signal, $w(n)$ i.e.,

$$y(n) = s(n) + w(n). \tag{2.1}$$

Taking the discrete-time Fourier transform of both sides gives,

$$Y(\omega) = S(\omega) + W(\omega). \tag{2.2}$$

We can express $Y(\omega)$ in polar form as:

$$Y(\omega) = |Y(\omega)|e^{j\Phi_y(\omega)}. \tag{2.3}$$

where, $|Y(\omega)|$ is the magnitude spectrum, and $\Phi_y(\omega)$ is the phase spectrum of the noisy speech.

Similarly, noise spectrum $W(\omega)$ can be expressed in polar form as:

$$W(\omega) = |W(\omega)|e^{j\Phi_w(\omega)}. \tag{2.4}$$

where, $|W(\omega)|$ is the magnitude spectrum, and $\Phi_w(\omega)$ is the phase spectrum of the additive noise. We don't know the $|W(\omega)|$ and $\Phi_w(\omega)$, and need to estimate each of these to get the estimate of the clean speech.

In speech enhancement algorithms $|W(\omega)|$ is replaced by its average value computed during non-speech activity(e.g., during speech pauses detected by *voice activity detector*). Noise phase spectrum $\Phi_w(\omega)$ is replaced by noisy speech phase spectrum $\Phi_y(\omega)$. This phase replacement is motivated by the fact that phase does not affect the speech intelligibility though it can affect speech quality to some extent [24]. After making those substitutions in Equation (2.2) we get,

$$\hat{S}(\omega) = [|Y(\omega)| - |\hat{W}(\omega)|]e^{j\Phi_y(\omega)} \tag{2.5}$$

where $|\hat{W}(\omega)|$ is the estimate of the noise magnitude spectrum. So, the task becomes simple to estimate the noise and subtract it from the noisy speech.

To avoid the negative magnitude the spectral subtraction rule was modified to

$$|\hat{S}(\omega)| = \begin{cases} |Y(\omega)| - |\hat{W}(\omega)|, & \text{if} |Y(\omega)| > |\hat{W}(\omega)| \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

This is similar to half-wave rectification. This equation for magnitude domain spectral subtraction can be easily extended to higher order spectra like power spectrum for example. Multiplying both sides of Equation(2.2) by $|Y^*(\omega)|$ leads to,

$$\begin{aligned} |Y^2(\omega)| &= |S^2(\omega)| + |W^2(\omega)| + |S^*(\omega)||W(\omega)| + |W^*(\omega)||S(\omega)| \\ &= |S^2(\omega)| + |W^2(\omega)| + 2\text{Re}(S(\omega) * W(\omega)). \end{aligned} \tag{2.7}$$

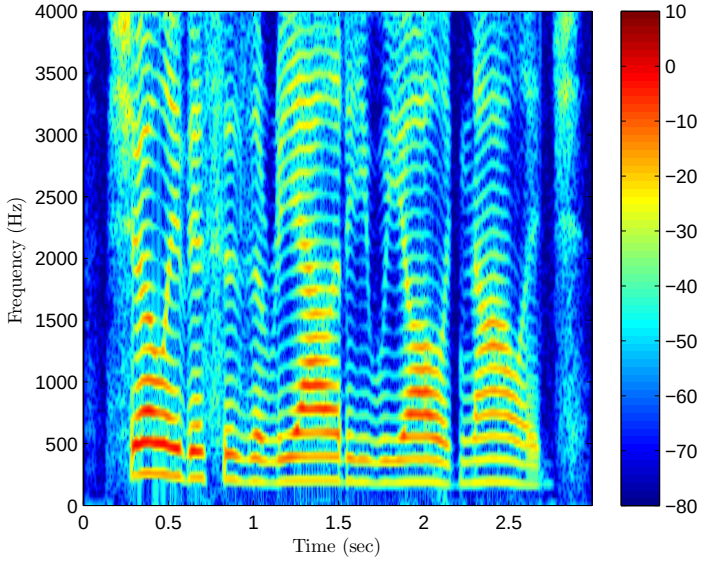
The terms $|W^2(\omega)|$,$|S^*(\omega)||W(\omega)|$ and $|W^*(\omega)||S(\omega)|$ are approximated by their expectations, i.e., E($|W^2(\omega)|$),E($|S^*(\omega)||W(\omega)|$) and E($|W^*(\omega)||S(\omega)|$). If $w(n)$ is assumed to be zero mean and independent of $s(n)$ then E($|S^*(\omega)||W(\omega)|$) and E($|W^*(\omega)||S(\omega)|$) reduce to zero and we have

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{W}(\omega)|^2, & \text{if} |Y(\omega)| > |\hat{W}(\omega)| \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

2.1.2 Shortcomings of Spectral Subtraction Algorithm

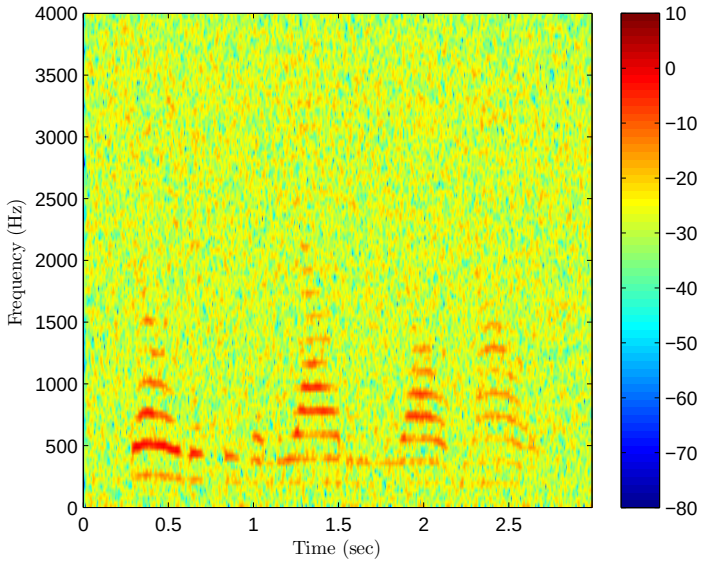
- **Musical noise**

Due to half-wave rectification in the spectral subtraction rule, the enhanced speech power spectrum may have small, isolated peaks occurring at random frequencies within the frame. When speech is reconstructed into time domain, it includes tones with frequencies that change randomly from frame to frame; that is, tones that are turned on and off at analysis frame rate (20-40 msec). This type of artifact is called as *musical noise* in the literature [25]. Musical noise can be observed in figure 2.2c due to presence of isolated peaks from time to time frames.

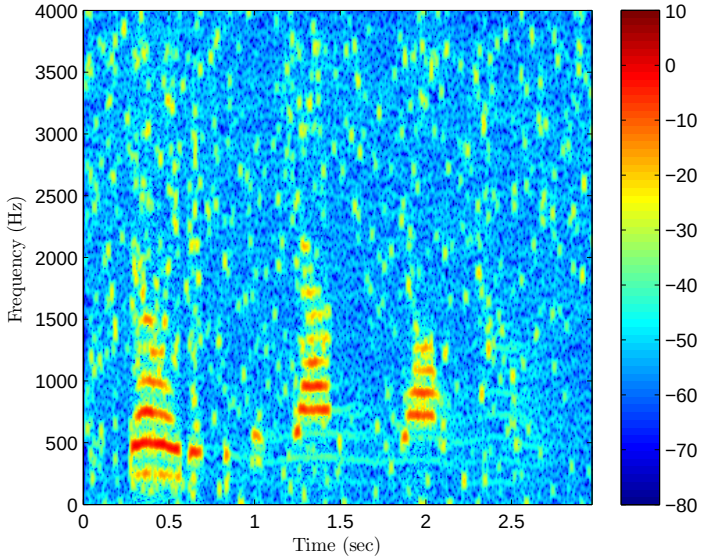


(a) Clean speech

Figure 2.2: Spectral subtraction processing:(a) Clean speech spectrogram,(b) Noisy speech spectrogram and (c) Spectrogram for speech after spectral subtraction processing.



(b) Noisy speech



(c) Enhanced speech

Figure 2.2: (Continued).

Some of the factors that contribute to musical noise are listed below:

1. Nonlinear processing of the negative subtracted spectral components.

2. Inaccurate estimate of the noise spectrum due to the fact that we are forced to use the average estimates of the noise. Hence, there are some significant variations between true noise and the

estimated noise spectrum. Using this averaged noise estimate may lead to isolated spectral peaks in the enhanced speech which contributes to annoying musical noise.

3. Large variance in the estimate of noisy and noise signal spectra even when long analysis window is used.

4. Large variability in gain.

To minimize the annoying effect of *musical noise*, the spectral subtraction rule is modified to [25],

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\hat{W}(\omega)|^2, & \text{if} |Y(\omega)| > (\alpha+\beta)|\hat{W}(\omega)| \\ \beta|\hat{W}(\omega)|^2, & \text{otherwise.} \end{cases} \tag{2.9}$$

There are several algorithms designed to minimize the amount of musical noise in processed speech [26, 27, 28, 29]. It is very difficult to minimize musical noise without affecting the speech signal itself. Hence, there exists a trade-off between noise reduction and speech distortion.

- **Usage of noisy phase instead of true noise phase**

For reconstructing speech, the original noisy phase is used without enhancement of phase. Though phase is usually considered to be insignificant for human perception as compared to amplitude, this is true only for high SNR(>5 dB). For lower SNRs phase leads to audible speech distortion. But enhancing the phase is much more difficult and complex than enhancing the amplitude [24]. This is applicable for all amplitude-only estimators. Hence, more stress is given on minimizing the effect of *musical noise* than enhancing phase.

Before leaving this section, it is very important to notice that there are several versions of standard spectral subtraction (which is mentioned above). Those are listed below [23]:

1. Nonlinear spectral subtraction.

2. Multiband spectral subtraction.

3. MMSE spectral subtraction.

4. Spectral subtraction based on perceptual properties.

5. Selective spectral subtraction.

2.2 Wiener Filter

Spectral subtraction algorithms are based largely on the intuitive and heuristically based principle. Noise being additive, it is intuitively appealing to obtain the clean speech estimate by subtracting the noise estimate from the noisy speech. This algorithm is not optimal in any sense. Wiener filter and MMSE STSA are the optimal estimators of the clean speech in the 'minimum mean square error' sense.

The Wiener filter is an optimal filter that minimizes the estimation error e(n), as shown in the figure below:

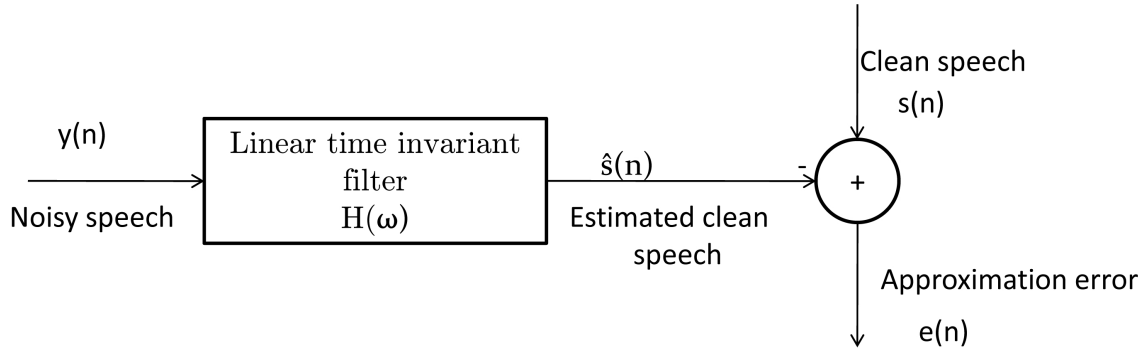


Figure 2.3: Block diagram for statistical filtering

The transfer function for Wiener filter can be derived in both time and frequency domain. For simplicity, it is presented here in frequency domain.

$$\hat{S}(\omega) = H(\omega)Y(\omega). \tag{2.10}$$

Then, estimation error at frequency ω_k can be written as:

$$\begin{aligned} E(\omega_k) &= S(\omega_k) - \hat{S}(\omega_k). \\ &= S(\omega_k) - H(\omega)Y(\omega). \end{aligned} \tag{2.11}$$

We need to compute $H(\omega)$ that minimizes the mean-square error, i.e.,$\mathrm{E}[|E(\omega_k)|^2]$,

$$\begin{aligned}\mathrm{E}[|E(\omega_k)|^2] &= \mathrm{E}[(S(\omega_k) - H(\omega)Y(\omega))^*(S(\omega_k) - H(\omega)Y(\omega))]. \\ &= \mathrm{E}[|S(\omega_k)|^2] - H(\omega_k)\mathrm{E}[S^*(\omega_k)Y(\omega_k)] - H^*(\omega_k)\mathrm{E}[Y^*(\omega_k)S(\omega_k)] + |H(\omega_k)|^2\mathrm{E}[|Y(\omega_k)|^2].\end{aligned} \tag{2.12}$$

Since, $P_{yy}(\omega_k) = \mathrm{E}[|Y(\omega_k)|^2]$ is the power spectrum of $y(n)$, and $P_{ys}(\omega_k) = \mathrm{E}[Y(\omega_k)S^*(\omega_k)]$ the cross-power spectrum of $y(n)$ and $s(n)$, the above equation can be written as:

$$J_2 = \mathrm{E}[|E(\omega_k)|^2] \quad = \mathrm{E}[|S(\omega_k)|^2] - H(\omega_k)P_{ys}(\omega_k) - H^*(\omega_k)P_{sy}(\omega_k) + |H(\omega_k)|^2P_{yy}(\omega_k). \tag{2.13}$$

To find the optimal filter $H(\omega_k)$ we take the complex derivative of J_2 with respect to $H(\omega_k)$ and set it to zero:

$$\begin{aligned}\frac{\partial J_2}{\partial H(\omega_k)} &= H^*(\omega_k)P_{yy}(\omega_k) - P_{ys}(\omega_k). \\ &= [H(\omega_k)P_{yy}(\omega_k) - P_{sy}(\omega_k)]^*. \qquad (2.14) \\ &= 0. \qquad (2.15)\end{aligned}$$

Solving for $H(\omega_k)$ we get

$$H(\omega_k) = \frac{P_{sy}(\omega_k)}{P_{yy}(\omega_k)}. \tag{2.16}$$

Note that $H(\omega_k)$ is complex valued, since the cross-power spectrum is generally complex quantity. For our signal model,$P_{yy}(\omega_k) = P_{ss}(\omega_k) + P_{ww}(\omega_k)$ and $P_{sy}(\omega_k) = P_{ss}(\omega_k)$, so we have

$$H(\omega_k) = \frac{P_{ss}(\omega_k)}{P_{ss}(\omega_k) + P_{ww}(\omega_k)}. \tag{2.17}$$

where $P_{yy}(\omega_k)$ is complex power spectrum of noisy speech, $P_{ss}(\omega_k)$ is complex power spectrum of clean speech and $P_{ww}(\omega_k)$ is the complex power spectrum of noise. This suggests that for our problem, $H(\omega_k)$ is real and even valued. This means h_k is non-causal and therefore, the Wiener filter is not realizable as it also requires the power spectrum of clean speech. This limitation of the Wiener filter is resolved by using Wiener filtering iteratively where first iteration noisy speech is taken as the clean speech [30].

The subtractive-type speech enhancement methods such as spectral subtraction Wiener fltering as discussed above are heavily dependent on the accuracy of voice detection, because noise estimation cannot be correct unless the non-speech frames are known. Due to this, such algorithms suffer from annoying *musical noise* artifacts.

2.3 MMSE Estimator

The Wiener filter, covered in the last section is an optimal complex spectral estimator for clean speech. It attempts to estimate the spectrum of clean speech from the given noisy speech complex spectrum. But the short time spectral amplitude (STSA) is acknowledged to be more important from speech intelligibility and quality perspectives. So, many approaches have been invented to estimate the amplitude of the clean speech from the given noisy speech. MMSE STSA estimator is an optimal estimator (in MSE sense) for clean speech amplitude. That is, it minimizes the following error function:

$$e = \mathrm{E}(\hat{S_k} - S_k)^2. \tag{2.18}$$

where $\hat{S_k}$ is the estimate of the clean speech amplitude and S_k is a true clean speech amplitude. In the Bayesian MSE approach the expectation is obtained with respect to the joint pdf $p(\mathbf{Y}, X_k)$, i.e., both $\mathbf{Y}$ and X_k are assumed to be random with Gaussian pdfs. The Bayesian MSE is given by:

$$Bmse(\hat{X_k}) = \int\int (X_k - \hat{X_k})^2 p(\,Y, X_k)\, d\mathbf{Y}\, dX_k. \tag{2.19}$$

Minimization of Bayesian MSE with respect to $\hat{X_k}$ leads to the optimal MMSE estimator given by [23]:

$$\hat{X_k} = \mathrm{E}(X_k | Y(\omega_0), Y(\omega_1), ..., Y(\omega_N - 1)) \tag{2.20}$$

where $\mathbf{Y} = [Y(\omega_1), ..., Y(\omega_N - 1)]$ is the noisy speech spectrum and 'N' is order of FFT. Assuming statistical independence between Fourier coefficients, we get $\mathrm{E}(X_k | Y(\omega_0), Y(\omega_1), ..., Y(\omega_N - 1)) = \mathrm{E}(X_k | Y(\omega_k))$. So we have

$$\begin{aligned} \hat{X_k} &= \mathrm{E}[X_k | Y(\omega_k)]. \\ &= \int_0^\infty x_k p(x_k | Y(\omega_k))\, dx_k. \end{aligned} \tag{2.21}$$

$$= \frac{\int_0^\infty x_k p(Y(\omega_k)|x_k) p(x_k)\, dx_k}{\int_0^\infty p(Y(\omega_k)|x_k) p(x_k)\, dx_k}. \tag{2.22}$$

But $p(Y(\omega_k)|X_k)p(X_k) = \int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_x)p(x_k,\theta_x)\,d\theta_x$, where θ_x is the realization of the phase random variable of $X(\omega_k)$.With this simplification we get,

$$\hat{S_k} = \frac{\int_0^{\infty}\int_0^{2\pi} x_k p(Y(\omega_k)|x_k,\theta_x)p(x_k,\theta_x)\,d\theta_x\,dx_k}{\int_0^{\infty}\int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_x)p(x_k,\theta_x)\,d\theta_x\,dx_k}. \tag{2.23}$$

From the assumed statistical model, $Y(\omega_k)$ is the sum of two zero-mean complex Gaussian random variables. Therefore, $p(Y(\omega_k)|x_k,\theta_x)$ will also be Gaussian:

$$p(Y(\omega_k)|s_k,\theta_s) = p_w(Y(\omega_k) - S(\omega_k)) \tag{2.24}$$

where $p_W(.)$ is pdf of the noise Fourier transform coefficients, $W(\omega_k)$. Then,

$$p(Y(\omega_k)|s_k,\theta_s) = \frac{1}{\pi\lambda_w(k)}\exp[\frac{-1}{\lambda_w(k)}|Y(\omega_k) - X(\omega_k)|^2]. \tag{2.25}$$

where $\lambda_w(k) = E(|W(\omega_k)|^2)$, is the variance of the kth spectral component of noise. Similarly,

$$p(s_k,\theta_s) = \frac{s_k}{\pi\lambda_s(k)}\exp[\frac{-s_k^2}{\lambda_s(k)}]. \tag{2.26}$$

Using above two pdfs form, we get [23]:

$$\hat{S_k} = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k}\exp[\frac{-v_k}{2}][(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})]Y_k. \tag{2.27}$$

where I_0 and I_1 denote the modified Bessel functions of zero and first order.
In eqn.(2.27),

$$v_k = \frac{\zeta_k}{1+\zeta_k}\gamma_k. \tag{2.28}$$

where

$$\gamma_k = \frac{Y_k^2}{\lambda_w(k)}. \tag{2.29}$$

is *a posteriori SNR* and,

$$\zeta_k = \frac{\lambda_s(k)}{\lambda_w(k)}. \tag{2.30}$$

is *a priori SNR*. The *a posteriori SNR* can be calculated easily from noisy speech using a *voice activity detector*. The *a priori* SNR is determined using *a decision-directed approach* given below:

$$\hat{\zeta_k}(m) = a\frac{S_k^2(\hat{m-1})}{\lambda_w(k, m-1)} + (1-a)max(\gamma_k(m) - 1, 0) \tag{2.31}$$

where m is the frame index. For the first frame,

$$\hat{\zeta_k}(0) = a + (1-a)max(\gamma_k(0) - 1, 0). \tag{2.32}$$

where the value of a is typically set to 0.98.

2.3.1 Significance of a Decision-directed Approach

When *a decision-directed approach* is used to determine *a priori* SNR, the enhanced speech had almost no musical noise. In the MMSE suppression rule, Equation (2.26), *a priori* SNR is a dominant factor affecting the noise reduction [31]. This *a priori* SNR is calculated using *a decision-directed* approach. The *decision-directed* approach exhibits two behaviors depending on the value of γ_k. When γ_k stays below 0dB (e.g., in the low energy speech segments), the ζ_k estimate corresponds to smooth version of γ_k. When γ_k is considerably larger than 0dB, the ζ_k estimate follows γ_k but with the delay of one frame as shown in figure 2.4. This smoothed estimate of *a priori* SNR results in smooth MMSE attenuation (unlike spectral subtraction). So, musical noise will be reduced or eliminated altogether as shown in figure 2.5c.

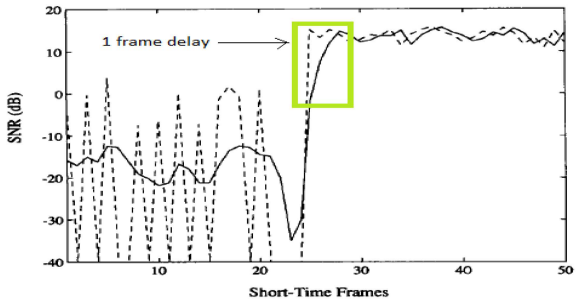
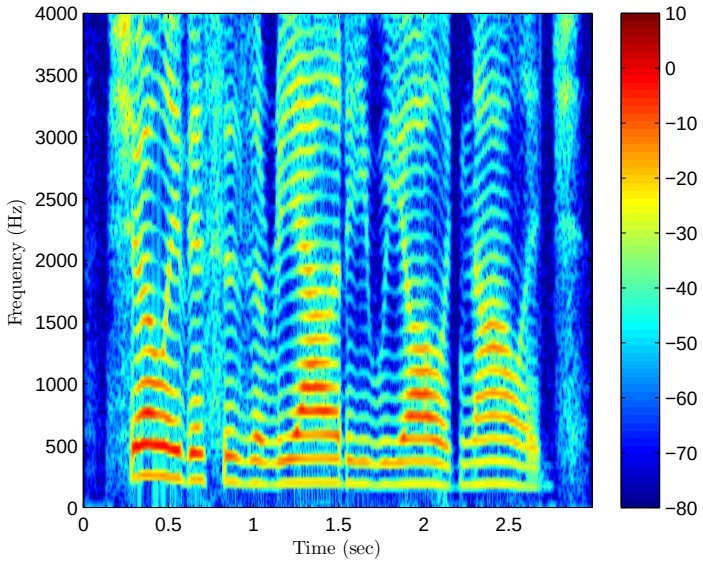
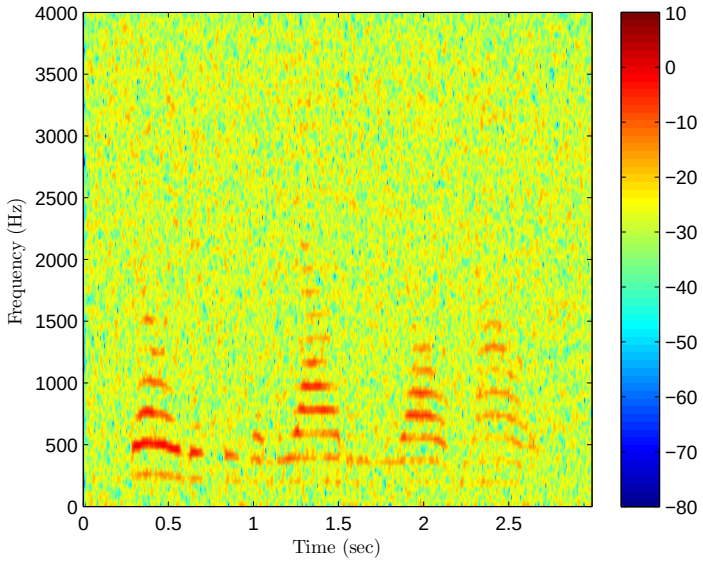


Figure 2.4: Behavior of *a priori* SNR due to *a decision-directed* approach. Solid line indicates *a priori* SNR and dotted line indicates *a posteriori* SNR. Taken from [15]

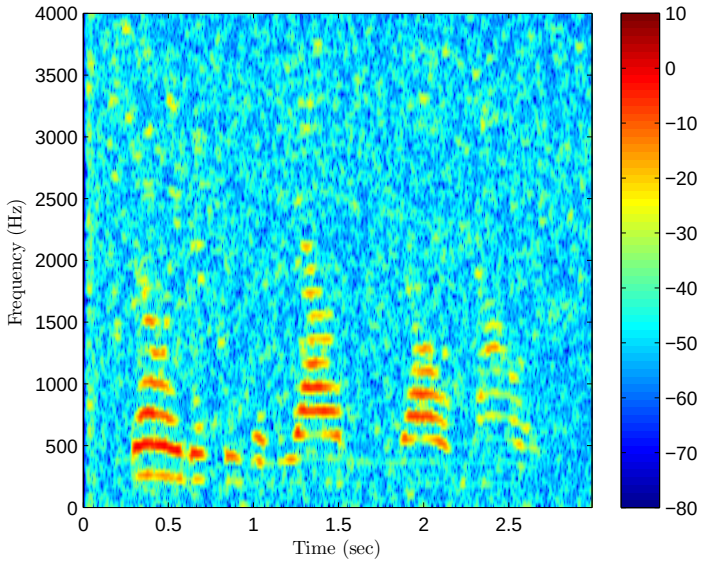


(a) Clean speech



(b) Noisy speech

Figure 2.5: MMSE STSA processing:(a) Clean speech spectrogram,(b) Noisy speech spectrogram and (c) Spectrogram for speech after MMSE processing.



(c) Enhanced speech

Figure 2.5: (Continued).

2.4 Speech Enhancement in Modulation Domain

Speech enhancement algorithms discussed in previous sections have been implemented in Fourier transform domain. Speech signal is divided into frames and those frames are transformed into the frequency domain. This domain is referred as *acoustic domain* in the literature to differentiate it from the *modulation domain*. The concept of *modulation domain* was proposed by Zadeh in 1950 [32]. Acoustic frequency is defined as the axis of the first STFT of the speech signal and modulation frequency is defined as the frequency axis of second STFT as shown in figure below [33]. The acoustic spectrum is the STFT of speech signal, while the modulation spectrum at a given acoustic frequency is the STFT of time series of the acoustic spectral magnitudes at that frequency. The short-time modulation spectrum is thus a function of time, acoustic frequency and modulation frequency [21].

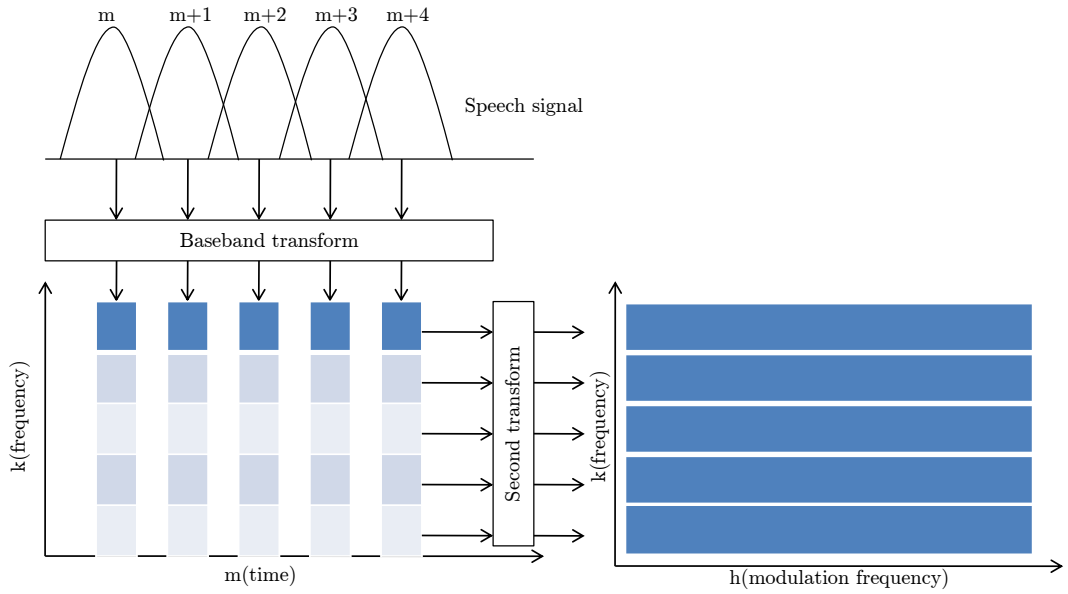


Figure 2.6: Acoustic domain to modulation domain transformation.

The modulation domain has been deeply studied for the processing of speech signals [34, 35, 36]. It has been shown that our perception of temporal dynamics corresponds to our perceptual filtering of the speech signal into modulation frequency channels. Also, most of the speech information is located in low frequency region (2-16 Hz) of the modulation spectrum, and this property can be exploited for better noise and speech separation. These findings have motivated the noise

reduction in the modulation domain instead of the acoustic domain. For this, standard *Analysis-Modification-Synthesis* framework is extended to the modulation domain [21] as discussed below.

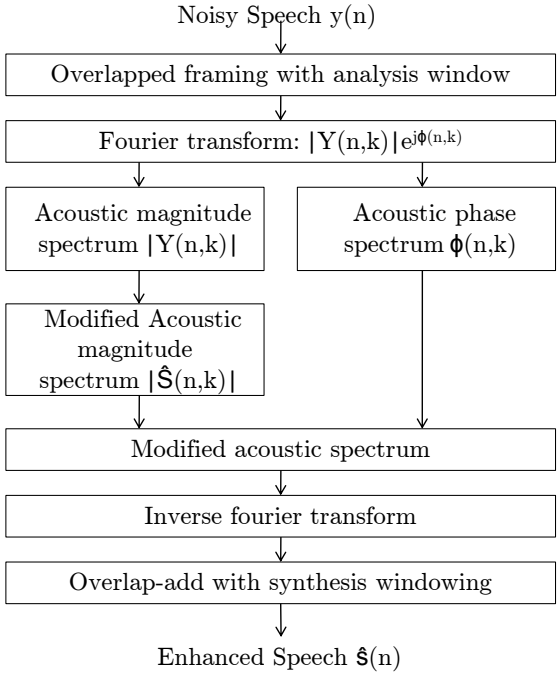


Figure 2.7: Analysis-Modification-Synthesis framework for acoustic domain.

For our signal model, $y(n) = s(n) + w(n)$. The STFT of the corrupted speech is given by,

$$Y(n,k) = \sum_{l=-\infty}^{\infty} y(l)\omega(n-l)e^{-j2\pi kl/N}. \tag{2.33}$$

where k is the index of discrete acoustic frequency, N is the acoustic frame duration, $\omega(n)$ is analysis window function. In polar form,

$$Y(n,k) = |Y(n,k)|e^{j\phi(n,k)} \tag{2.34}$$

where, $|Y(n,k)$ and $\phi(n,k)$ are magnitude and phase spectrum of the noisy speech, respectively. The modulation spectrum is calculated using second STFT as

$$\mathcal{Y}(\eta,k,m) = \sum_{l=-\infty}^{\infty} |Y(n,k)|\nu(\eta-l)e^{-j2\pi ml/M}. \tag{2.35}$$

where η is the acoustic frame number, k is index of discrete acoustic frequency, m is index of discrete modulation frequency, M is modulation frame duration and $\nu(\eta)$ is modulation domain window function. In the polar form,

$$\mathcal{Y}(\eta, k, m) = |\mathcal{Y}(\eta, k, m)|e^{j\varphi(n,k)} \tag{2.36}$$

where, $|\mathcal{Y}(\eta, k, m)|$ and $\varphi(n, k)$ are magnitude and phase spectrum of the noisy speech modulation transform, respectively. So, in the modulation domain we can write,

$$\mathcal{Y}(\eta, k, m) = \mathcal{S}(\eta, k, m) + \mathcal{W}(\eta, k, m). \tag{2.37}$$

For this signal model, spectral subtraction rule becomes

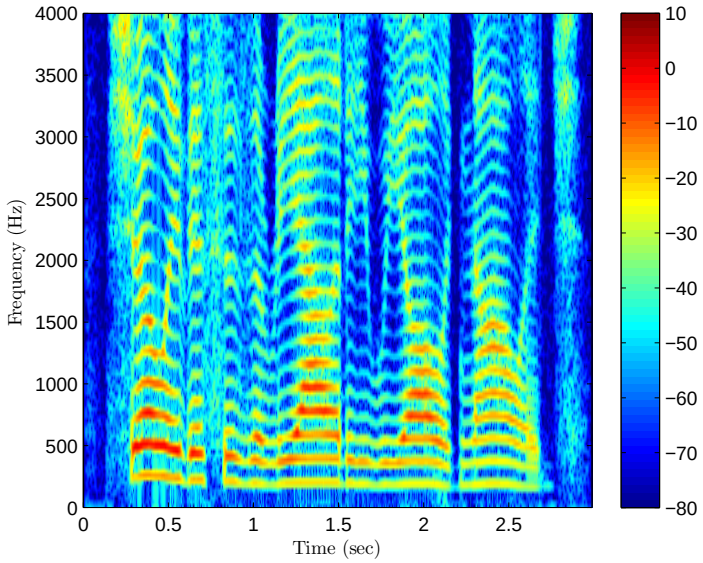
$$|\mathcal{S}(\eta, k, m)|^2 = \begin{cases} |\mathcal{Y}(\eta, k, m)|^2 - \rho|\hat{\mathcal{W}}(\eta, k, m)|^2, & \text{if} |\mathcal{Y}(\eta, k, m)|^2 > (\rho + \beta)|\hat{\mathcal{W}}(\eta, k, m)|^2 \\ \beta|\hat{\mathcal{W}}(\eta, k, m)|^2, & \text{otherwise.} \end{cases} \tag{2.38}$$

Acoustic domain window length is set to 30-40 msec and modulation domain window length is 256 msec. The noise is estimated in same manner as in acoustic domain algorithms, but in the modulation domain. After modulation spectral subtraction, modified modulation spectrum is transformed back into acoustic domain spectrum by inverse STFT and overlap-add synthesis. Finally, acoustic spectrum is transformed into time domain by inverse STFT and overlap-add synthesis.

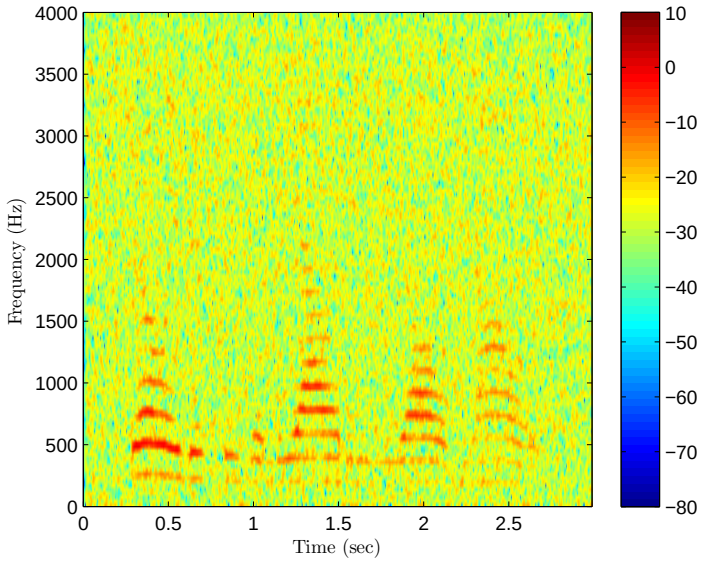
2.4.1 Advantages of Spectral Subtraction in Modulation Domain over Spectral Subtraction in Acoustic Domain

1. As modulation domain is more closely related to human's perceptual system, speech enhancement in the modulation domain results in better perceptual speech quality. Also, the speech distortion is much lower than in acoustic domain.

2. The enhanced speech has a very low amount of musical noise if the modulation window length is large (180-280 msec). This results in smoothing in temporal dimension and hence less musical noise as can be seen in figure below.

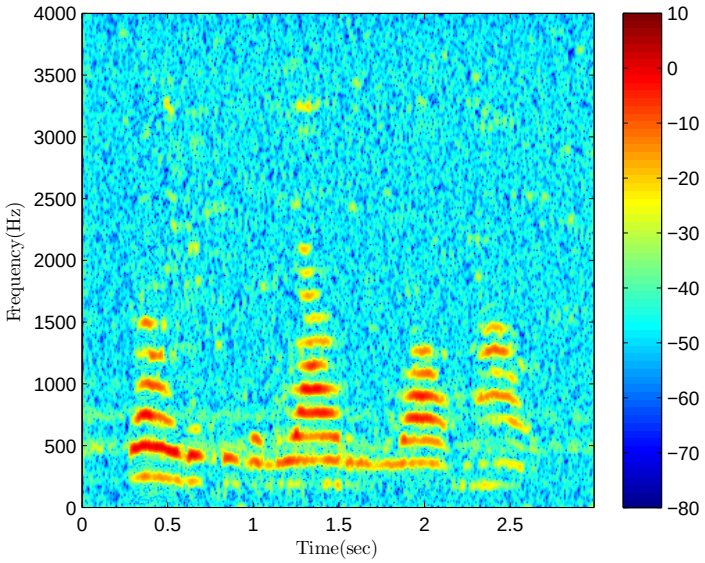


(a) Clean speech



(b) Noisy speech

Figure 2.8: Spectral subtraction processing:(a) Clean speech spectrogram,(b) Noisy speech spectrogram and (c) Spectrogram for speech after spectral subtraction in modulation domain.



(c) Enhanced speech

Figure 2.8: (Continued).

Note: This is the result of our implementation of the mentioned algorithm.

2.5 Harmonicity Based Speech Enhancement

Most earlier speech enhancement methods do not consider the structure of the speech. Each frame of the speech signal is treated similarly and suppression gain differs depending upon the SNR of that frame. But, the voiced segment (vowels and semivowels) of the speech signal exhibits quasi-periodicity, also known as harmonicity. So, the speech signal can be decomposed into voiced (vowels and semi-vowels) and unvoiced (consonants) segments. This voiced and unvoiced nature of the speech signal is due to the behavior of the *vocal folds*, which provide the excitation to the *vocal tract*. During the voiced segment of the speech, vocal folds vibrate periodically while during unvoiced segment no such periodicity exists. This mechanism of the vocal folds and vocal tract is used to design the engineering model of speech production as shown below:

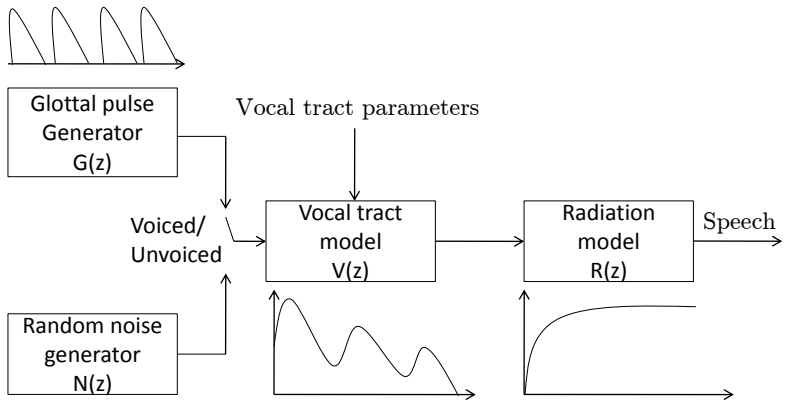


Figure 2.9: Engineering model of speech production.

The opening and closing of the vocal folds during the voiced segment produces the periodic input signal. The time duration of one cycle of opening or closing of vocal folds is called fundamental period and reciprocal is called fundamental frequency (F0). The fundamental frequency varies from a low around 80 Hz for male speakers to a high of 280 Hz for children. The periodicity is broadly distributed across frequency and time and is robust in presence of noise. This motivates the use of this clue to gain more knowledge about underlying speech. Many speech enhancement algorithms have been developed to exploit the harmonicity of the voiced speech [37, 38, 39, 40, 41, 42, 43, 44, 45, 46]. Below, we discuss one of such algorithms [22] which exploits harmonicity of voiced segment to enhance the *phase* of the voiced speech using sinusoidal speech model.

2.5.1 Phase Enhancement for Voiced Speech

For our signal model, $y(n) = s(n) + w(n)$. The Fourier transform of $y(n)$ is

$$Y(\omega) = |Y(\omega)|e^{j\phi_y(\omega)} \tag{2.39}$$

where $|Y(\omega)|$ is the magnitude spectrum of the noisy speech and $\phi_y(\omega)$ is phase spectrum of noisy speech. Due to additive noise both $|Y(\omega)|$ and $\phi_y(\omega)$ are corrupted. Though the effect of this corrupted phase spectrum is inaudible at higher SNRs (>5 dB), at lower SNRs the speech sounds distorted. Hence, phase enhancement at such low SNR can further enhance the quality of speech [47]. The voiced speech can be modeled as a weighted superposition of H sinusoids, leading to harmonic signal model,

$$\tilde{s}(n) = \sum_{h=0}^{H} A_h cos(\Omega_h n + \psi_h) \tag{2.40}$$

with real valued amplitude A_h, time domain phase ψ_h and normalized angular frequency,

$$\Omega_h = 2\pi f_h / f_s = 2\pi(h+1)f_0/f_s \tag{2.41}$$

where f_s, f_0, f_h denote sampling frequency, fundamental frequency and harmonic frequency, respectively. Phase enhancement is carried out in *baseband* STFT domain instead modulated STFT due to high correlation between phase and magnitude spectrum in the *baseband* domain. We provide the brief introduction to those two versions of STFT below:

2.5.2 Two Versions of STFT

Baseband STFT

In this version STFT is implemented by following equation,

$$X_B(n,\omega) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m} = \sum_{m=n}^{n+N-1} x(m)w(n-m)e^{-j\omega m} \tag{2.42}$$

where n is STFT frame index, ω is STFT frequency, N is order of FFT, $x(m)$ is the time domain speech signal and $w(n)$ is the window function. As STFT is a function of two parameters, it can be interpreted in two ways: 1) If n is fixed and ω is varied then we get standard frequency analysis interpretation. 2) If n is varied and ω is fixed then we have the filtering interpretation. We will

focus more on filtering interpretation of STFT. If we fix value of ω at ω_0 then

$$X_B(n, \omega_0) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega_0 m}. \tag{2.43}$$

This is a convolution of $x(m)e^{-j\omega m}$ with $w(n)$. In this view, the signal $x(m)$ is modulated by $e^{-j\omega m}$ and passed through a filter whose impulse response is a window function $w(n)$. We can view this as modulation a band of frequencies centered around ω_0 down to base-band (hence this version is named so), and then filtering by $w(n)$. This is illustrated in following figure:

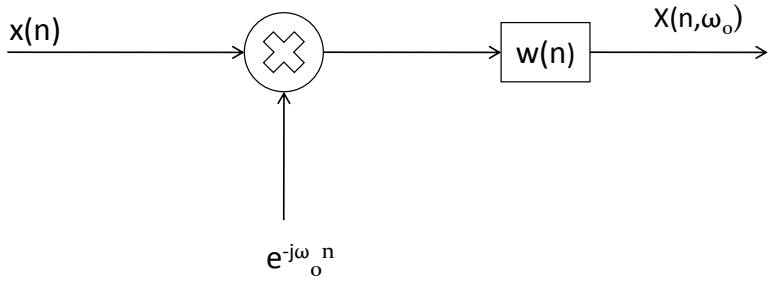


Figure 2.10: Time domain view of Baseband STFT.

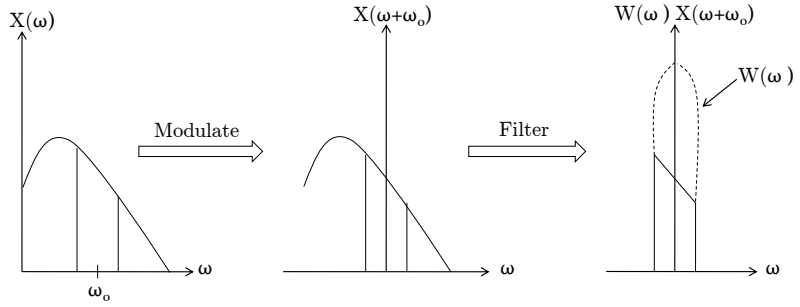


Figure 2.11: Frequency domain view of Baseband STFT.

Modulated STFT

In the *baseband* STFT, the frames are extracted by keeping the signal as it is and shifting and flipping the window function, but instead, if we keep the window at the constant position and shift signal instead, then we get the *modulated* STFT. This is given by following equation,

$$X_M(n, \omega) = \sum_{m=0}^{N-1} x(n+m)w(m)e^{-j\omega m}. \tag{2.44}$$

The name comes due to its relationship with the baseband domain STFT as derived below:

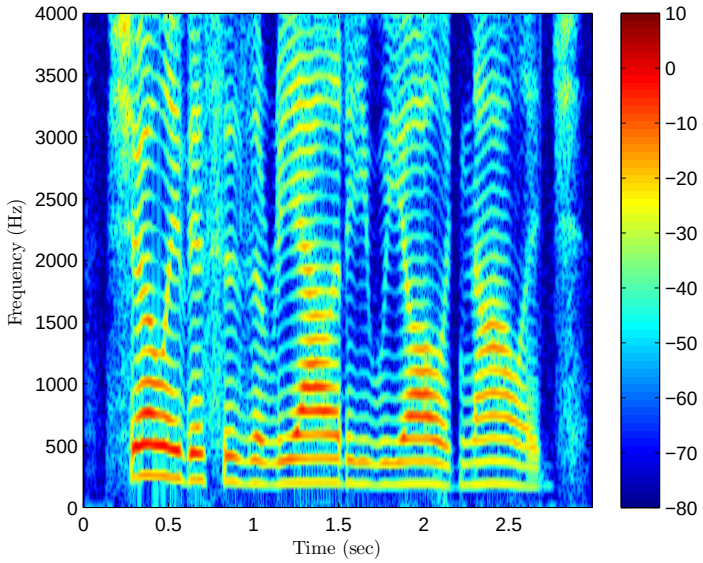
$$\begin{aligned} X_B(n,\omega) &= \sum_{m=n}^{n+N-1} x(m)w(n-m)e^{-j\omega m}. \\ &= \sum_{m=0}^{N-1} x(n+m)w(-m)e^{-j\omega(n+m)} \text{.....Putting, m = n + m.} \\ &= e^{-j\omega n}\sum_{m=0}^{N-1} x(n+m)w(-m)e^{-j\omega m}. \\ &= e^{-j\omega n}\sum_{m=0}^{N-1} x(n+m)w(m)e^{-j\omega m} \text{.....assuming symmetric window.} \\ &= e^{-j\omega n}X_M(n,\omega). \end{aligned} \tag{2.45}$$

From Equation (2.44),

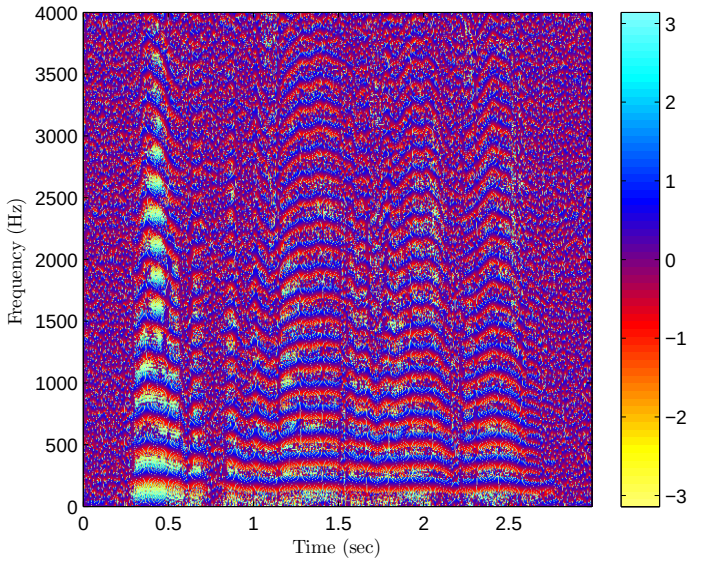
$$X_M(n,\omega) = e^{j\omega n}X_B(n,\omega). \tag{2.46}$$

$$\angle X_M(n,\omega) = \omega n + \angle X_B(n,\omega). \tag{2.47}$$

From Equation (2.45), it is clear that $X_M(n,\omega)$ is a modulated version of $X_B(n,\omega)$. Hence, it is named as *modulated* STFT. Also, from Equation (2.46), the phase of $X_M(n,\omega)$ has larger dynamic range, as it depends on the frame number n. So, it suffers from phase wrapping. On the other hand, the phase of $X_M(n,\omega)$ lies between $-\pi$ to π. Hence, it avoids phase wrapping. Due to this behavior of the *baseband* STFT, phase difference spectrum appears to be highly correlated to amplitude spectrum in the voiced region of the speech. This can be seen in figure below. In the clean speech, phase difference spectrum is correlated with the clean amplitude spectrogram but it is corrupted in noisy speech phase difference.

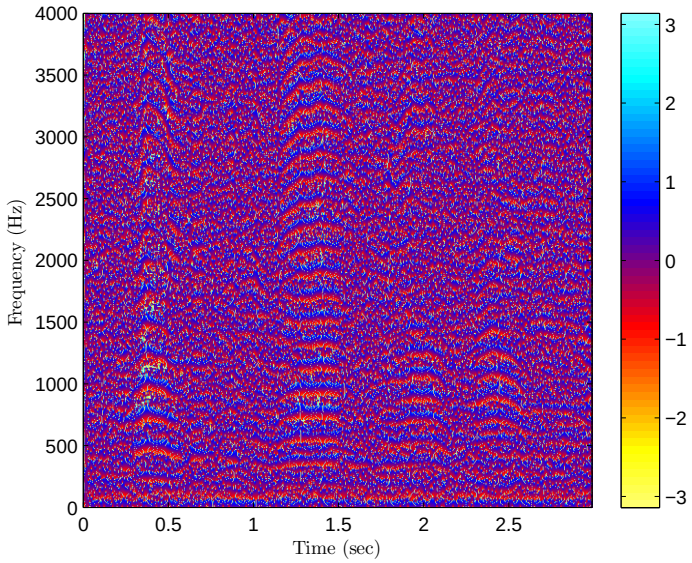


(a) Clean speech spectrogram



(b) Clean speech phase difference

Figure 2.12: Phase difference from frame to frame for clean and noisy speech.



(c) Noisy speech phase difference

Figure 2.12: (Continued).

Note: The results are generated by our implementation of this algorithm.

Assuming the harmonic signal model for voiced speech in (2.39), the phase can be reconstructed in baseband domain for voiced speech using following formulas [22]:

$$\phi_{\tilde{S}_B}(k,n) = \phi_{\tilde{S}_B}(k,n-1) + (\Omega_h^k - \Omega_k)L \tag{2.48}$$

where $\phi_{\tilde{S}_B}(k,n)$ stands for phase for voiced speech Fourier coefficient at index k, and frame n, L is the window shift in number of samples. This equation is used recursively to find the phase values at the frequency coefficient directly containing the harmonic component [22]. Also,

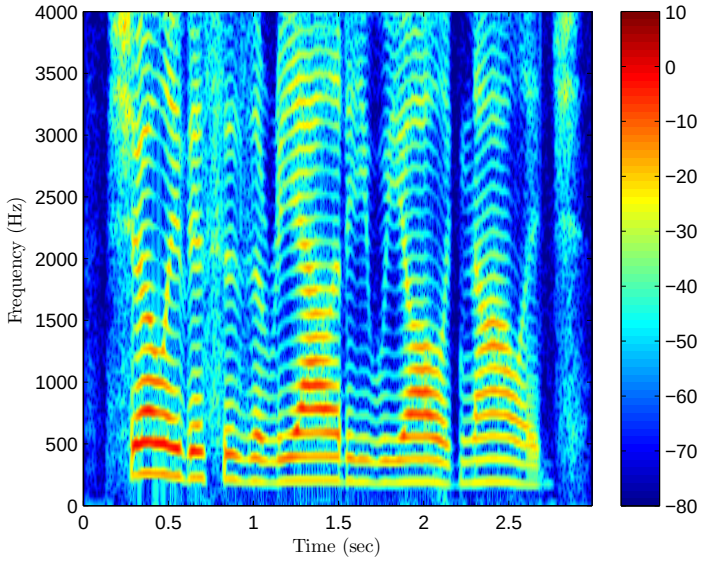
$$\Omega_h^k = \underset{\Omega_h}{\operatorname{argmin}}(|\Omega_k - \Omega_h|)$$

where Ω_k is angular frequency corresponding to current DFT bin, k. Ω_h^k is angular frequency of the harmonic closest to current DFT bin, k.

To estimate the phase between the harmonics in the frame, the following equation is used:

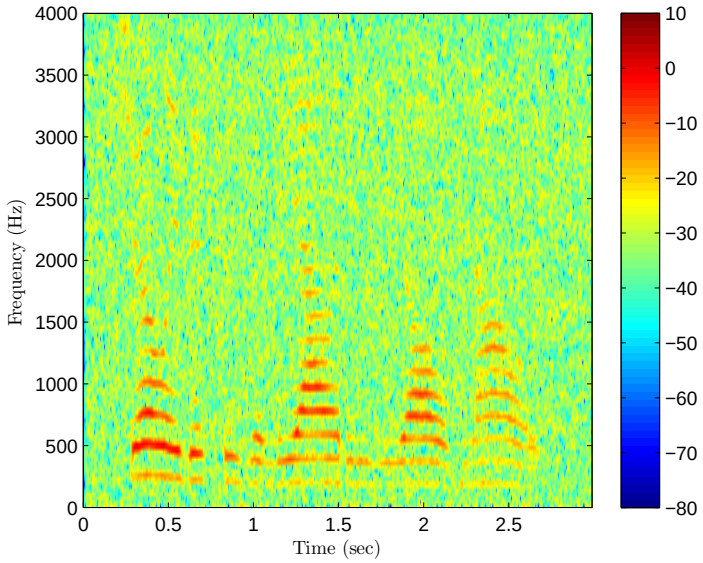
$$\phi_{\tilde{S}_B}(k+i,n) = \phi_{\tilde{S}_B}(k,n) + i\pi - i\frac{2\pi nL}{N} \tag{2.49}$$

where $i \in [\lceil \frac{-f_0/2}{f_s} N \rceil, ..., \lceil \frac{f_0/2}{f_s} N \rceil]$. Once the phase is reconstructed in the baseband domain, the STFT is transformed to the modulation domain and speech is reconstructed using overlap-add synthesis. Amplitude of the transform is left unchanged. If the reconstructed speech is processed again to plot the magnitude spectrogram, then even the amplitude spectrum looks enhanced as shown in figure below. Noise is effectively suppressed between the harmonics due to this harmonic model processing.

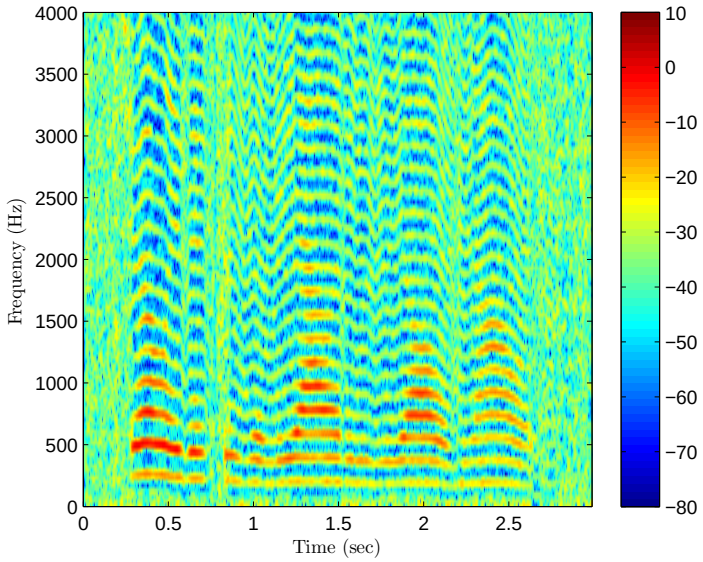


(a) Clean speech spectrogram

Figure 2.13: Figure show the output of the phase enhancement algorithm.



(b) Noisy speech spectrogram



(c) Enhanced speech spectrogram

Figure 2.13: (Continued).

Note: The results are generated by our implementation of this algorithm.

Chapter 3

OVERVIEW OF SPEECH QUALITY ASSESSMENT TECHNIQUES

As discussed earlier, speech enhancement algorithms attempt to improve the speech quality and/or intelligibility. Speech quality is related to how pleasant the speech sounds to the listener, while the speech intelligibility is related to the recognition accuracy for the processed speech. To evaluate the performance of speech enhancement algorithms, we need to quantify these properties. This has motivated researchers to devise the measures for speech quality and intelligibility. These measures can be classified into two groups: 1) Subjective measures. 2) Objective measures. Subjective measures are based on the response of the human listeners to speech and are calculated by experiments with various listeners and speech samples. Objective measures are based on the mathematical evaluation of the speech quality and intelligibility. Subjective quality assessments are often accurate and reliable, provided they are performed under stringent conditions [48, 49]. However, subjective evaluation is time consuming. Objective assessment, on the other hand, requires knowledge of the clean speech to evaluate the performance of the speech enhancement algorithm. We will describe some of the widely used measures for the speech quality in the following sections.

3.1 Subjective Speech Quality Assessment

Subjective listening tests provide the most reliable method to assess the quality of the enhanced speech. In this approach, listeners are subjected to the training and the testing phase. In the training phase, listeners are provided the reference speech samples to bring all of them to the same level of judgment, and in the testing phase actual enhanced speech is assessed. These approaches are broadly classified into two categories: 1) Approaches based on a relative preference task 2) Approaches based on assigning a numerical value to the speech quality. We will briefly summarize both of these approaches below.

3.1.1 Relative Preference Methods

The *isopreference* test was perhaps the earliest paired-comparison test to measure the speech quality [50, 51]. In [51], the test involved all possible forward and reverse combinations of test and reference signals (as given in table 3.1). Listeners are asked to mark the preferred speech utterance in each combination. The count of preferred test and reference signals are averaged for multiple listeners. With this score the reference signal that is equally preferred to the test signal is obtained, and it indicates the speech quality. Several extensions of this method are proposed in literature which uses the different reference signals for the test [52, 53].

Table 3.1: Reference Conditions

System	Signal Description
A	High-fidelity speech(clean)
B	Speech band-pass-filtered (800-3000Hz)
C	Speech low-pass-filtered (3000 Hz) and combined with low-pass-filtered white noise (500 Hz). Peak SNR 10 dB
D	Speech combined with reverberant echo. Delay of first echo 150 msec.
E	Speech peak-clipped, then band-pass-filtered (300-2000 Hz)

3.1.2 Absolute Category Rating Methods

Preference tests typically answer the question "How well the listener liked the test signal over the reference signal?". So, these tests just compare the test signal against the reference signal.

Due to such approach, all kinds of the distortions in the test signal can not be represented as only a limited number of reference signals are available. Also, the reason a particular signal is preferred over others is not evident in such tests. To address such issues, the *rating methods* are used. In such tests, reference signals are not required and listeners are asked to rate the test signal over some range of options.

3.1.2.1 Mean Opinion Score

This is the most widely used subjective speech quality test, in which the listeners are asked to rate the quality of the speech over the five-point numerical scale (as in table 3.2). The measured quality of speech is obtained by averaging the ratings from all listeners. This average is commonly called as 'Mean Opinion Score (MOS)'. This test is carried out in two stages: training and evaluation. Training is required to equalize the subjective range of the speech quality across all the listeners. In the evaluation phase, the test utterance is given to the listeners and the scores are recorded [23].

Table 3.2: MOS rating Scale

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

3.1.2.2 Diagnostic Acceptability Measure

The MOS requires the listener to state the overall quality value of the speech but it does not ask for the basis of this judgment. So, two listeners may report the same quality of the speech but for different attributes of the signal. Thus, MOS is known as a single dimension measure of the speech quality, and it can not easily be used to improve the performance of the speech enhancement algorithm. To eliminate this limitation of the subjective test *Diagnostic Acceptability Measure* (DAM) test was proposed. DAM is a multidimensional speech quality test, and it evaluates the speech quality over three dimensions namely, parametric, metametric and isometric as shown in table 3.3. Listeners are asked to rate the speech and noise distortions along with metametric and

isometric attributes over the range of 0 - 100 [54].

Table 3.3: Scales Used in the DAM Test

		Parametric Scales	
Name	**Abbreviation**	**Description**	**Example**
Signal	SF	Fluttering,bubbling	AM Speech
	SH	Distant,thin	High-pass Speech
	SD	Rasping,crackling	Peak-clipped Speech
	SL	Muffled,smothered	Low-pass Speech
	SI	Irregular,interrupted	Interrupted Speech
	SN	Nasa,whining	Band-pass Speech
	TSQ	Total Signal Quality	
Background	BN	Hissing,rushing	Gaussian noise
	BB	Buzzing,humming	60-Hz hum
	BF	Chirping,bubbling	Narrow-band noise
	BR	Rumbling,thumping	Low-frequency Speech
	TBQ	Total Background Quality	
		Metametric Scales	
	I	Intelligibility	
	P	Pleasantness	
		Isometric Scales	
	A	Acceptability	
	CA	Composite Acceptability	

3.2 Objective Speech Quality Assessment

Subjective speech quality provides the most reliable approach to assess the speech quality. However, the tests are time consuming and require multiple listeners. Due to these limitations, several researchers have worked to find an objective way to assess speech quality. Ideally, an objective measure should be able to assess the speech quality of the enhanced speech without need of the original clean speech samples. Objective measures must take into account the low-level processing

(e.g, psychoacoustics) and higher level processing such as prosodics, semantics and pragmatics. But, most of the objective assessment algorithms require access to the original clean speech and some of them can exploit the low-level processing. Despite of these limitations, some of the objective measures are significantly correlated with the subjective measures like MOS.

Objective measures are implemented by segmenting the speech signal into the frames of 10-30 msec, and then computing the distortion measure between original and enhanced speech signal. Frame level measures are then averaged to obtail the final objective speech quality score. The measures can be calculated in both time and frequency domain as can be seen in the following methods. In the frequency domain the speech spectrum magnitude is assumed to be correlated to the speech quality [23, 55, 56].

3.2.1 Segmental SNR

Segmental SNR can be evaluated in both time and frequency domain. Time domain segmental SNR is one of the easiest one to compute. This requires that both original clean speech and the enhanced speech are time-aligned. The segmental SNR is defined as:

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}(\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2}.) \tag{3.1}$$

where $x(n)$ is original clean speech, $\hat{x}(n)$ is enhanced speech, N is frame length and M is number of frames in signal. One potential problem with this measure is that during silent frames the value can be a large negative number which will bias overall SNR value. One way to avoid this is to exclude the silent frames from the speech. Another version of this method which attempts to deal with the problem of large negative SNR values is proposed in [57].

The segmental SNR can be extended to the frequency domain as follows [57]:

$$fwSNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}B_j\log_{10}[F^2(m,j)/(F(m,j)-\hat{F}(m,j))^2]}{\sum_{j=1}^{K}B_j)}. \tag{3.2}$$

where B_j is the weight for j^{th} frequency band, K is the number of bands, M is the total number of frames, $F(m,j)$ is the filter-bank amplitude of the clean signal in j th frequency band and at m th frame and $F(m,j)$ is the filter-bank amplitude of the enhanced signal in j^{th} frequency band and at m^{th} frame. The advantage of using SNR in the frequency domain is to have different weights for

different frequency bins.

3.2.2 Spectral Distance Measures Based on LPC

Several objective measures have been proposed based on the dissimilarity between the all-pole model of clean speech and the enhanced speech signals. These measures assume that over the short time intervals, speech can be represented by the p^{th} order all pole model of the form [23]:

$$x(n) = \sum_{i=1}^{p} a_x(i)x(n-i) + G_x u(n). \tag{3.3}$$

where $a_x(i)$ are the coefficients of the all-pole model, G_x is the filter gain and $u(n)$ is unit variance white noise excitation. Two common all-pole model based measures used to evaluates speech quality are the log-likelihood ratio and Itakura-Saito(IS) measure.

The log-likelihood ratio(LLR) measure is defined as:

$$d_{LLR}(a_x, \bar{a}_{\hat{x}}) = \log \frac{\bar{a}_{\hat{x}}^T R_x \bar{a}_{\hat{x}}}{a_x^T R_x a_{\hat{x}}}. \tag{3.4}$$

where a_x^T are the LPC coefficients of the clean signal, $\hat{a}_x^T$ are the LPC coefficients of the enhanced signal and R_x is the auto-correlation matrix of the clean signal.

The IS measure is defined as:

$$d_{IS}(a_x, \bar{a}_{\hat{x}}) = \frac{G_x}{\bar{G}_{\hat{x}}} \frac{\bar{a}_x^T R_x \bar{a}_{\hat{x}}}{a_x^T R_x a_{\hat{x}}} + \log(\frac{\bar{G}_{\hat{x}}}{G_x}) - 1. \tag{3.5}$$

where G_x and $\bar{Gx}$ are the all-pole gains of the clean and enhanced signal, respectively.

3.2.3 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ) is an objective measure which is well correlated to the subjective MOS, and it predicts the speech quality accurately for distortions which include channel losses in telecommunication network, packet loss, signal delays, and codec distortion [58]. The speech is processed as shown in the following figure to compute this objective measure.

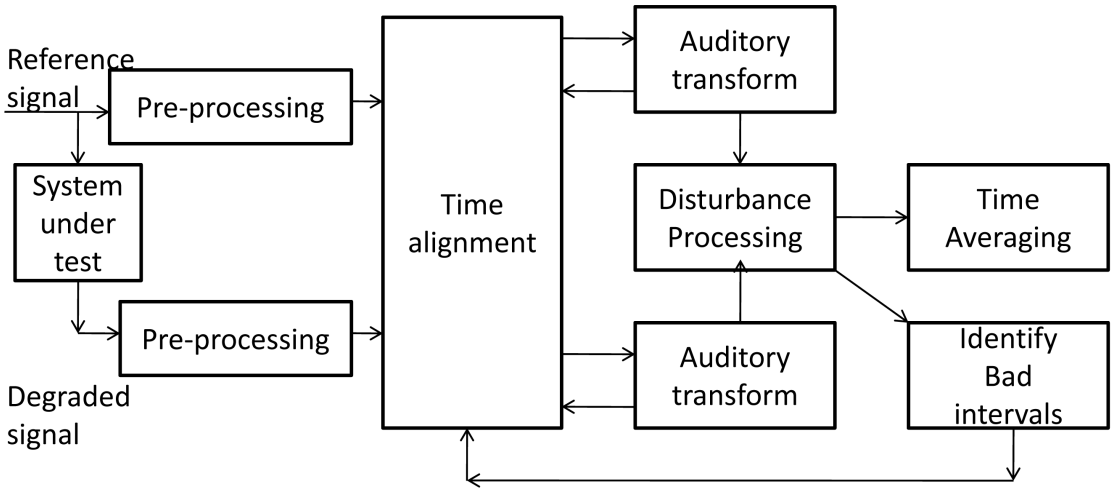


Figure 3.1: Block diagram of PESQ measure computation.Taken from [23]

The structure of the PESQ computation system is shown in the above figure . The original (clean) and degraded signals are first level-equalized to a standard listening level, and processed by a filter whose response is similar to a standard telephone handset. The signals are aligned in time to correct for time delays, and then processed through an auditory transform (this consists of the short time Fourier transform followed by Bark scale transformation of the power spectrum) to obtain the loudness spectra. The difference termed as disturbance between the loudness spectrum of clean speech and the degraded speech is computed and averaged over time and frequency to get the PESQ measure [23]. The range of PESQ is: 0.5 - 4.5. Higher values indicate higher resemblance of the loudness spectra of clean speech and degraded speech.

Chapter 4

USING BASEBAND PHASE DIFFERENCE FOR NON-STATIONARY NOISE ESTIMATION

In this chapter, we discuss the use of *Baseband Phase Difference* to identify the frequency bins dominated by noise in the voiced frames, and these are used to update the noise estimate to track the non-stationary noise accurately. Noise estimation is the most important step in a speech enhancement system and accurate noise estimation can help to reduce the annoying artifacts introduced by speech processing. Depending on the environment, the noise corrupting the speech can be quite non-stationary like noise originating from a train passing by, from passing cars or from people walking on the street or in a restaurant. Most speech enhancement algorithms try to reduce the amount of noise by applying a gain function in the spectral domain. This gain function is generally a function of noisy speech power, clean speech power and noise power. Inaccurate noise estimation can result in speech and noise distortion including annoying artifacts in the enhanced speech. If the noise is under-estimated then residual noise or musical noise will be audible, while over-estimation of noise will cause speech distortion resulting in loss in speech quality and intelligibility. In [22], phase

enhancement is carried out assuming the sinusoidal model for the voiced speech. Although, this results in reduction of noise between the speech harmonics, the processed speech sounds unnatural due to inaccurate speech modeling. Also, only voiced frames in the speech are enhanced. We propose to use this harmonic modeling to identify the noise dominated frequency bins to obtain better noise estimates. These noise estimates can be integrated with existing speech enhancement algorithms to improve the performance in non-stationary noise. This chapter is organized as follows. Section 4.1 briefly discusses the existing noise estimation approaches. Section 4.2 explains proposed noise estimation algorithm and Section 4.3 demonstrate the usage of the proposed noise estimation algorithm.

4.1 Review of Existing Noise Estimation Algorithms

The most widely used approach in noise estimation involves *voice activity detection* (VAD) based algorithms. VAD algorithms typically extract some feature/features (e.g., short time energy, zero crossing rate) from the input signal that is in turn compared against a threshold value, usually determined during speech-absent periods. VAD algorithms generally output a binary decision per frame, where frames may last for 20-40 msec. A frame is declared to contain voice activity (VAD=1) if the measured feature value exceeds a threshold, otherwise it is considered to be noise (VAD=0) . So, this algorithm estimates and updates the noise spectrum only in speech inactive periods. Although a VAD based algorithm works well for stationary noises (like white noise), it might fail for the case of non-stationary noise [59]. Several VAD based noise estimation algorithms have been proposed based on the extracting features from the input speech [60, 61, 62, 63]. Some VAD algorithms are used in the commercial applications including audio-conferecing, cellular networks and digital cordless telephone systems. VAD algorithms exploit the fact that there can be silence not only at the end and beginning of the sentence, but also in the middle of sentence. These silence segments correspond to the closures of the stop consonants, primarily the unvoiced stop consonants i.e., /p/,/t/,/k/, etc. For example, the VAD based classification of speech and silence periods is shown in the following figure.

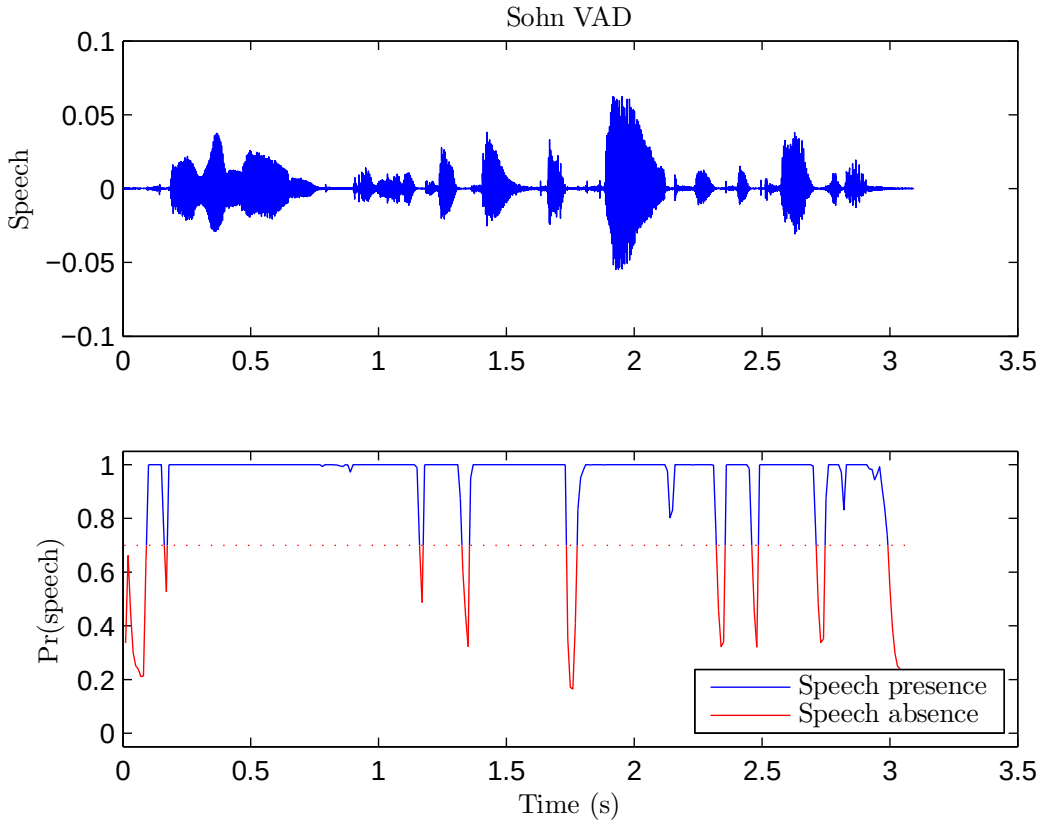


Figure 4.1: Speech and noise classification using VAD [64]. Time domain speech is shown in top figure. Speech detection as indicated by speech presence probability is shown in bottom figure.

VAD based noise estimation works well only for stationary noises and in high SNR conditions. Also, it is not able to track the noise during speech activity. Various noise estimation algorithms are proposed to track the non-stationary noise even during speech activity and low SNR. Those algorithm includes *Minimum Statistics Noise Estimation* [5], *Moving Controlled Recursive Averaging* [4], *histogram based noise estimation* [65], *MMSE noise estimation* [66] etc. Those algorithms are based on following facts:

1. Power of the noisy speech signal in individual frequency bands often decays to the power level of the noise, even during speech activity. Hence, by tracking the minimum of the noisy power in each frequency band, a rough estimate of the noise can be obtained. The minimum statistics algorithm is based on this fact. This algorithm tracks the minimum of the noisy power spectrum within a finite window.

2. Noise affects the signal spectrum non-uniformly. Some regions are affected more than others.

Hence, the noise is estimated by averaging the noise estimates at each frequency bin depending upon the effective SNR at each frequency bin. *Moving Controlled Recursive Averaging* algorithm is based on this fact.

3. Histogram based noise estimation is based on the fact that most frequent values of the energy levels at given frequency band correspond to the noise at that frequency band.

All of these algorithms do not consider the fact that speech is composed of voiced speech and unvoiced speech. Voiced speech presence can be detected even in low SNR due to its robust harmonic structure. Using this additional information, noise estimate can be improved further. In the following section, we propose a noise estimation algorithm which estimates the noise even during voiced frames. This algorithm makes use of the harmonic structure of the voiced speech.

4.2 Baseband Phase Difference as a Clue for Noise Estimation

4.2.1 Motivation

As discussed in section 2.5.1, in baseband STFT (Short Time Fourier Transform) the phase difference from one frame to another is highly correlated to the magnitude spectrum of voiced speech. Here, the harmonic model is used to represent voiced speech as given in the following equation:

$$\tilde{s}(n) = \sum_{h=0}^{H} A_h cos(\Omega_h n + \psi_h). \tag{4.1}$$

To compute the phase (in baseband domain) for voiced speech we use the following two equations derived from the above voiced speech model.

$$\phi_{\tilde{S}_B}(k,n) = \phi_{\tilde{S}_B}(k,n-1) + (\Omega_h^k - \Omega_k)L. \tag{4.2}$$

where, $\phi_{\tilde{S}_B}(k,n)$ stands for phase for voiced speech Fourier coefficient at index k and frame n and L is the window shift in number of samples. This equation is used recursively to find the phase values at the frequency coefficient directly associated with the harmonic component [22]. Also, Ω_h^k , the

angular frequency of the harmonic closest to current DFT bin 'k', is given by:

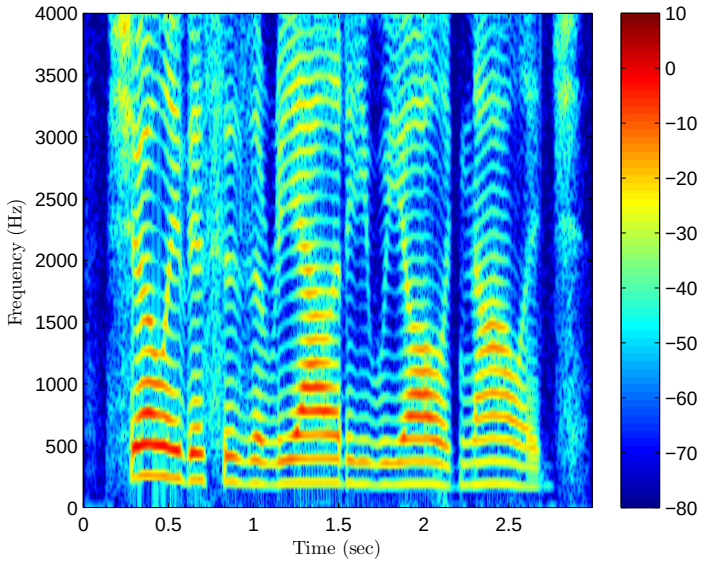
$$\Omega_h^k = \underset{\Omega_h}{\operatorname{argmin}}(|\Omega_k - \Omega_h|),$$

where Ω_k is angular frequency corresponding to current DFT bin 'k'.

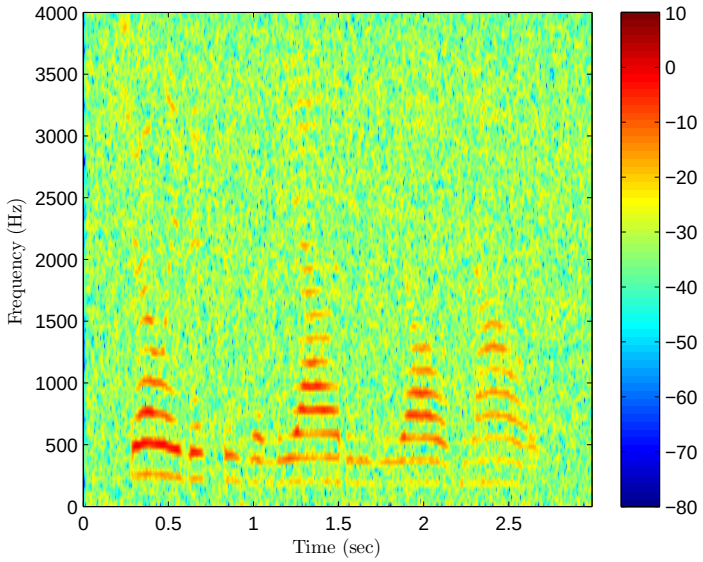
To estimate the phase between the harmonics in a voiced frame following equation is used

$$\phi_{\tilde{S}_B}(k+i,n) = \phi_{\tilde{S}_B}(k,n) + i\pi - i\frac{2\pi nL}{N} \tag{4.3}$$

where $i \in [\lceil \frac{-f_0/2}{f_s}N \rceil, ..., \lceil \frac{f_0/2}{f_s}N \rceil]$. Once the clean speech phase difference is estimated, it can be used to detect the frequency bins dominated by noise. This can be seen from the following figures. An enhanced speech spectrogram is obtained from speech reconstructed after phase enhancement. Correlation between the *enhanced speech spectrogram* and *estimated clean speech phase difference* indicates the use of estimated clean speech phase difference to estimate noise between harmonics during voiced speech frames. This algorithm uses the YIN [67] algorithm to estimate the pitch frequency.

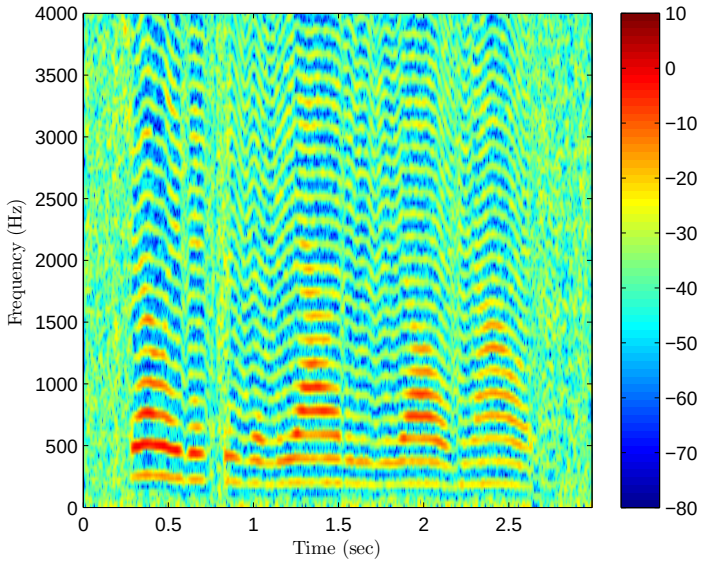


(a) Clean speech spectrogram

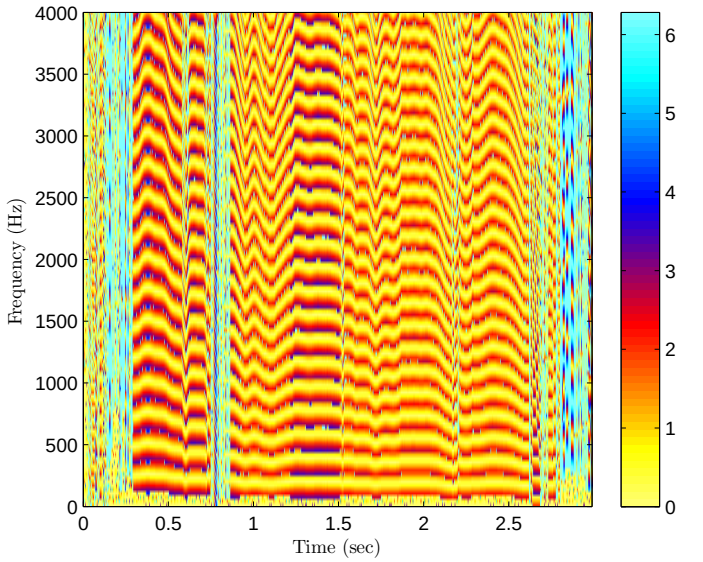


(b) Noisy speech spectrogram

Figure 4.2: Clean, noisy and enhanced speech spectrogram are shown.



(c) Enhanced speech spectrogram using phase enhancement



(d) Estimated clean speech phase difference i.e. $\phi_{\tilde{S}_B}(k,n) - \phi_{\tilde{S}_B}(k,n-1)$

Figure 4.2: (Continued).

4.3 Proposed Noise Estimation Algorithm

4.3.1 Determination of Noise Dominant Frequencies

In [22], the estimated clean speech phase given by (4.2) and (4.3) is used to reconstruct the speech, and the reconstructed speech is shown to be enhanced in the voiced segments. We used this phase estimation method to identify the noise dominant frequency bins in the voiced frames. These values are then used to further refine the final noise estimation. We compute the frame to frame phase difference from the above estimated clean phase as $\Delta\phi_{\tilde{S}_B}(k,n) = \phi_{\tilde{S}_B}(k,n) - \phi_{\tilde{S}_B}(k,n-1)$. This phase difference is highly correlated with the magnitude of the underlying clean speech in the voiced frames as shown in Fig. 4.2a and Fig. 4.2d. Clean speech is corrupted by adding babble noise at 0dB global SNR (See Fig. 4.2b). Estimated frame to frame phase difference for clean speech, i.e., $\Delta\phi_{\tilde{S}_B}(k,n) = \phi_{\tilde{S}_B}(k,n) - \phi_{\tilde{S}_B}(k,n-1)$, is represented in Fig. 4.2d. Here, we have plotted the absolute value of the phase difference in the range from 0 to 2π. From Fig. 4.2d, it can be noted that phase difference can be used to determine the frequencies dominated by the harmonics and the frequencies containing high amount of noise in the voiced frames. Those noise dominant frequencies correspond to the gaps between the harmonics.

From (4.2) and (4.3), it can be noted that in voiced frames the phase difference is close to zero for frequencies associated with the harmonics, and this phase difference deviates from zero for other frequencies. Thus, we use a threshold(ϕ_T) based test to separate such frequencies as described below: Let H be the total number of harmonics in a voiced frame, let F_h be the set of frequencies dominated by harmonic h, and let F_{nh} be the set of frequencies considered to be valid noise candidates in the neighboring of harmonic h. If k_h is the DFT bin corresponding to harmonic h then we apply the following bin selecting rule in the range of frequencies $k_h + i$, where $i \in [\lceil \frac{-f_0/2}{f_s} N \rceil, ..., \lceil \frac{f_0/2}{f_s} N \rceil]$, for each harmonic:

$$k \in \begin{cases} F_{nh}, \text{if } \Delta\phi_{\tilde{S}_B}(k,n) > \phi_T. \\ F_h, \text{otherwise.} \end{cases} \tag{4.4}$$

4.3.2 Computation of Noise PSD

For all frequencies in the frequency sets F_h and F_{nh}, the noise power is assumed to be constant and is given as the average of spectral magnitudes over F_{nh}. The noise estimate is calculated

as:

$$N_{Fnh}(n) = \sum_{j=1}^{|F_{nh}|} \frac{|Y(F_{nh}(j), n)|^2}{|F_{nh}|} \text{.....for } k \in k_h + i. \tag{4.5}$$

This is repeated for each harmonic in a voiced frame, n. Final noise PSD is obtained by combining the individual noise estimates and can be represented as:

$$|\hat{W}_\phi(n)|^2 = \{N_{Fn1}(n), N_{Fn2}(n), N_{Fn3}(n)....., N_{FnH}(n)\}. \tag{4.6}$$

This noise estimation is valid only for voiced frames. In the unvoiced frames, noise estimation is carried out using standard VAD based noise estimation [23, 68]. When a voiced frame is detected, the noise estimate is updated with the proposed noise PSD as:

$$|\hat{W}(k, n)|^2 = 0.8|\hat{W}(k, n-1)|^2 + 0.2|\hat{W}_\phi(k, n)|^2. \tag{4.7}$$

4.4 Use of Noise Estimation for Speech Enhancement

In this section, we describe the use of the previously discussed noise estimation algorithm for the speech enhancement in presence of stationary and non-stationary noises. We combine this noise estimation algorithm with the *spectral subtraction* and *MMSE STSA* algorithms. The spectral subtraction over-attenuation factor is adjusted to further improve the quality of the enhanced speech. The use of baseband phase difference as a means for detecting the noise dominant frequency components in the voiced frames results in more accurate estimation of the noise spectrum, and can be combined with any speech enhancement algorithm for noise estimation. But, this requires accurate estimation of pitch frequency in presence of noise, hence a robust pitch detection algorithm like the YIN algorithm [67] is used to detect the pitch frequency in each voiced frame. Also, *aperiodicity* measure of the YIN algorithm is set to 0.5 to detect the voiced frames.

4.4.1 Spectral Subtraction with Proposed Noise Estimation

Here, we explain in detailed how spectral subtraction is modified to exploit the estimated clean speech phase difference. With this phase difference, it becomes easier to detect the spectral sparsity in the voiced frame facilitating the non-stationary noise estimation. The basic spectral

subtraction rule is given as:

$$|\hat{S}(n,k)|^2 = \begin{cases} |Y(n,k)|^2 - \alpha|\hat{W}(n,k)|^2, & \text{if} |Y(n,k)| > (\alpha+\beta)|\hat{W}(n,k)| \\ \beta|\hat{W}(n,k)|^2, & \text{otherwise.} \end{cases} \tag{4.8}$$

where $\hat{S}(n,k)$ is the estimated clean speech, $Y(n,k)$ is noisy speech, $\hat{W}(n,k)$ is estimated noise, n is STFT frame index, k is FFT bin index, α is the over-subtraction factor determined using [25] (This factor is a constant number for all the frequency bins in the frame, and it is calculated by comparing the SNR of the present frame against some threshold as mentioned in [25]). The parameter β is the floor parameter to reduce the amount of musical noise in the enhanced speech. We extend the basic spectral subtraction algorithm to take the new noise estimation algorithm into account. The overall algorithm is described in the following steps:

1. Noisy speech $y(n)$ (sampled at 8000 Hz) is divided into the frames of 32 msec. with 4 msec. shift using the Hamming window. This small shift is used as it gives higher correlation between the phase difference of the clean speech and the magnitude spectrum.

2. For each frame, we take a 256 point DFT (modulated STFT) and transform into baseband STFT. We decide whether a frame is voiced or not using YIN algorithm [67]. For voiced frames, baseband phase difference is determined by using the algorithm described in section 2.5.

3. Noise estimation (on the power spectrum) is carried out differently in the voiced and non-voiced frames. It is assumed that the first 30 frames (as frame shift is just 4 msec) are noise-only frames, and those are averaged to obtain the initial noise estimate. In the non-voiced frames we use VAD to detect the noise-only frame by comparing the current SNR to some threshold (in this case it is set to 3dB). If the current SNR is less than this threshold then the frame is taken as noise, and the noise estimate is updated accordingly. In each voiced frame, we use the algorithm described in section4.3 to estimate the noise and running noise estimate is again updated. This all process is described in the following set of equations.

Let

$$Y(n,k) = S(n,k) + W(n,k). \tag{4.9}$$

be the noisy speech frame where n is the frame index and k is the DFT bin index.

Let $\hat{W}(n,k)$ be the noise estimate for frame n.

Assuming that the first 30 frames as noise-only we have initial noise estimate:

$$\hat{W}(n,k) = \frac{\sum_{n=1}^{30} Y(n,k)}{30}. \tag{4.10}$$

If a non-voiced frame is detected and SNR $>$ 3dB, we update the noise estimate using

$$\hat{W}(n,k) = 0.9\hat{W}(n-1,k) + 0.1Y(n,k). \tag{4.11}$$

When a voiced frame is encountered, the noise estimate $\hat{W}_{Voiced}(n,k)$ determined using 4.7 is used to update the running noise estimate as:

$$\hat{W}(n,k) = 0.8\hat{W}(n-1,k) + 0.2\hat{W}_{Voiced}(n,k). \tag{4.12}$$

4. In addition to incorporating this new noise estimate, we also make the over-attenuation factor α frequency dependant in the voiced frames. For test purpose, we set $\alpha = 8$ if $\Delta\phi(\omega,n) > \phi_T$ else α = 2.7. This results in less attenuation for the harmonic dominant frequencies and more attenuation for noise dominant frequencies in the voiced frame.

5. The new noise estimation algorithm and the adaptive over-attenuation factor α are used in equation (4.4) to obtain the estimate of the clean speech. Due to this new noise estimation method and adaptive over-attenuation factor low energy voiced speech is maintained resulting in higher speech quality.

4.4.2 MMSE STSA with Proposed Noise Estimation

We also verify the effectiveness of this new noise estimation algorithm for the *MMSE STSA* noise reduction algorithm. The MMSE STSA parameters are kept as it is (except the frame shift is changed to 4msec to exploit the baseband phase difference clue) as mentioned in section 2.3 but the noise is estimated using the proposed algorithm. It is observed that due to this noise estimation algorithm, the performance of the MMSE STSA is improved significantly for the non-stationary noise. This will be discussed further in the next chapter where we discuss the performance of this

method.

4.4.3 Combined MMSE STSA and Spectral Subtraction

As we have discussed previously, the spectral subtraction algorithm suffers from introducing annoying musical noise though it suppresses the noise effectively. It is observed that due to our proposed noise estimation algorithm which exploits the spectral sparsity for updating the noise estimate, the amount of the musical noise is reduced significantly at low SNR($<$ 5 dB). Several approaches exists to minimize the effect of musical noise [26, 28, 69]. On the other hand, the MMSE noise reduction algorithm eliminates the musical noise due to its decision-directed based *a prior SNR* estimation. We verify the effectiveness of the combination of those two algorithms to minimize the effect of musical noise and obtain significant noise reduction in the voiced period of the speech.

The fusion of MMSE STSA and spectral subtraction is performed in the short-time spectral domain by combining the magnitude spectra of these two speech enhancement algorithms. The fusion is performed by following set of rules:
Let U and V denote the unvoiced and voiced frame detected by YIN algorithm respectively, $|\hat{S}_{MMSE}(n,k)|$ and $|\hat{S}_{SpecSub}(n,k)|$ be the magnitude spectra of speech enhanced by MMSE STSA and spectral subtraction rule.

$$|\hat{S}_{Fusion}(n,k)|^2 = \begin{cases} |\hat{S}_{LMMSE}(n,k)|^2 & \text{if } |Y(n,k)| = U \\ & or \Delta\phi(n,k) < \phi_T \\ \hat{S}_{Comb} & \text{otherwise.} \end{cases} \tag{4.13}$$

where $\hat{S}_{Comb} = 0.8 * |\hat{S}_{SS}(\lambda,\mu)|^2 + 0.2 * |\hat{S}_{LMMSE}(\lambda,\mu)|^2$. i.e., we are using the contribution of MMSE STSA enhanced spectra in the unvoiced and harmonic dominant speech to reduce the effect of annoying musical noise with minimum speech distortion. We use spectral subtraction in the noise dominant speech for effective noise reduction in the voiced frame.

Chapter 5

RESULTS

We have evaluated the performance of the proposed noise estimation algorithm in this chapter. This algorithm is combined with spectral subtraction and MMSE STSA, and the performance is evaluated on 500 phonetically balanced sentences from the TIMIT database. The speech is degraded by adding white, babble, restaurant and subway noises with global SNRs ranging from -5 dB to 10 dB. White noise is an example of stationary noise while the remaining noises are non-stationary. The segment length is 32 ms with a 4 ms shift. With a sampling frequency of 8 kHz, this corresponds to frame length of 256 samples with a shift of 32 samples. PESQ is employed as an objective measure for speech quality. The fundamental frequency is estimated using the YIN [67] algorithm with a threshold set to 0.5 and segment shift of 4 ms. The aperiodicity measure of the YIN algorithm is set to 0.7 to classify each speech frame as voiced/unvoiced. For an analysis of the upper bound, we also present the results when the fundamental frequency is estimated from clean speech.

5.1 Spectral Subtraction with the Proposed Noise Estimation Algorithm

In the following tables, performance of the proposed noise estimation algorithm is evaluated by combining it with the traditional spectral subtraction, which is denoted as 'SpecSub'. Pitch estimation is carried out on both noisy speech and clean speech and results are presented separately. SpecSub, combined with the proposed noise estimation algorithm, using pitch estimation on noisy

speech, is denoted as 'SpecSub-NPE'. When the pitch estimation is based on clean speech, the resulting combined method is denoted as 'SpecSub-CPE'.

5.1.1 Results and Analysis of Results

Table 5.1: PESQ evaluation of the proposed algorithm against standard spectral subtraction for white noise.

	PESQ			
Global SNR(in dB)	Noisy	SpecSub	SpecSub-NPE	SpecSub-CPE
-5	1.22	1.32	1.63	1.76
0	1.43	1.84	2.05	2.12
5	1.72	2.21	2.41	2.43
10	2.05	2.45	2.67	2.68

Table 5.2: PESQ evaluation of the proposed algorithm against standard spectral subtraction for babble noise.

	PESQ			
Global SNR(in dB)	Noisy	SpecSub	SpecSub-NPE	SpecSub-CPE
-5	1.32	1.21	1.47	1.76
0	1.66	1.73	1.99	2.17
5	2.02	2.17	2.38	2.47
10	2.38	2.57	2.66	2.72

Table 5.3: PESQ evaluation of the proposed algorithm against standard spectral subtraction for restaurant noise.

	PESQ			
Global SNR(in dB)	Noisy	SpecSub	SpecSub-NPE	SpecSub-CPE
-5	1.35	1.12	1.45	1.78
0	1.66	1.64	1.91	2.11
5	2.00	2.07	2.31	2.42
10	2.34	2.46	2.64	2.68

Table 5.4: PESQ evaluation of the proposed algorithm against standard spectral subtraction for subway noise.

	PESQ			
Global SNR(in dB)	Noisy	SpecSub	SpecSub-NPE	SpecSub-CPE
-5	1.22	1.13	1.58	1.73
0	1.49	1.56	1.90	2.07
5	1.81	2.01	2.29	2.37
10	2.16	2.40	2.59	2.62

In the above tables, the first column, 'Global SNR(in dB)', represents the signal-to-noise ratio after speech is degraded. The second column, 'Noisy', gives the value of the objective measure 'PESQ' for the degraded speech. The third column indicates the PESQ value for speech enhanced using the traditional spectral subtraction algorithm. Similarly, the fourth and fifth columns give the values of the PESQ measure for the enhanced speech using proposed approach with pitch estimation on noisy and clean speech, respectively. The upper bound due to pitch estimation on clean speech can be observed from the data in the tables. We also give the graphical representation of the above tabulated performance comparison in the figures 5.1, 5.2, 5.3 and 5.4, which follow.

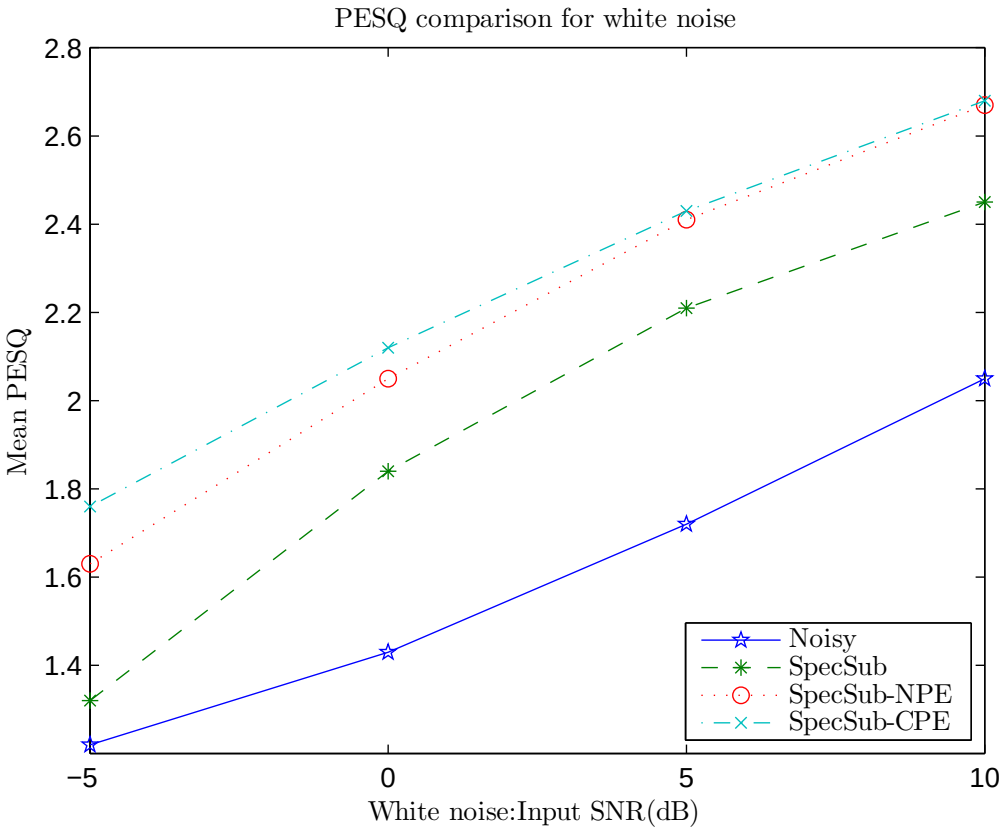


Figure 5.1: Results of the proposed spectral subtraction speech enhancement algorithm for white noise.

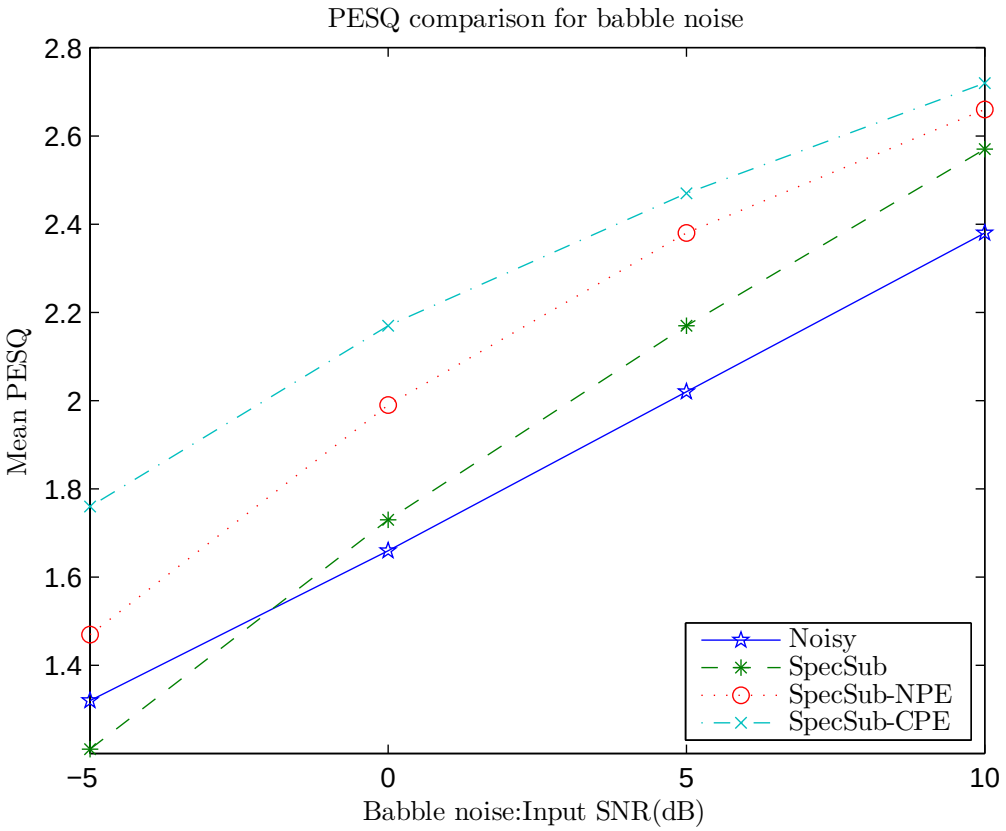


Figure 5.2: Results of the proposed spectral subtraction speech enhancement algorithm for babble noise.

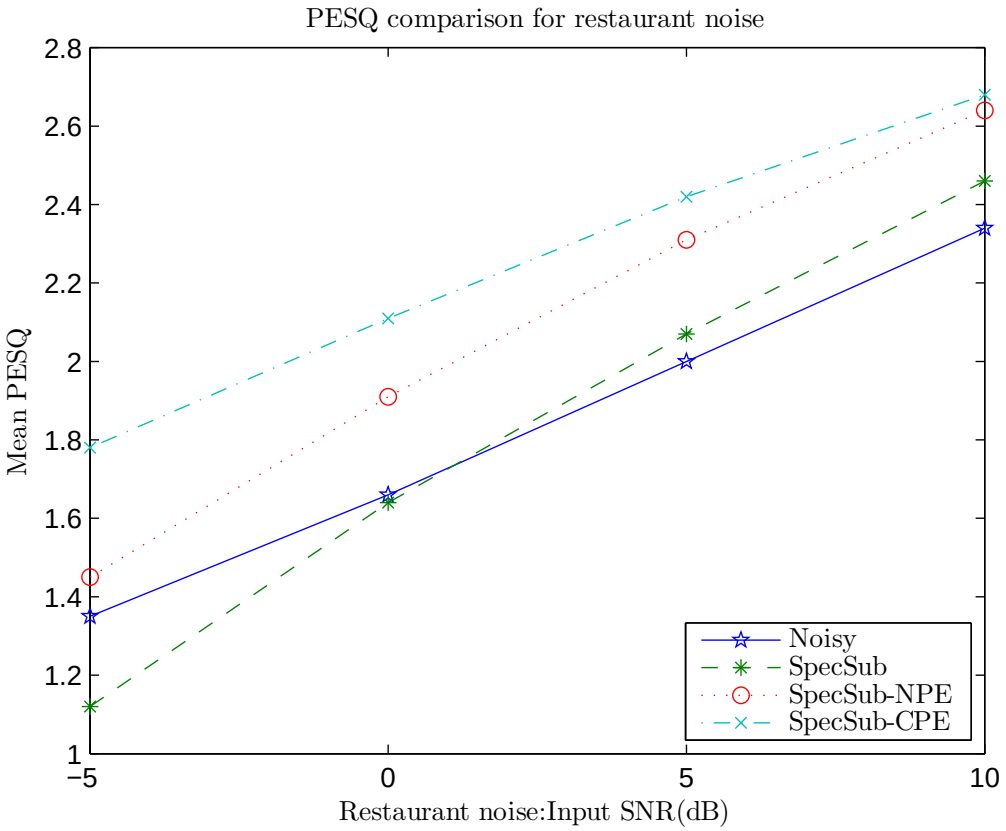


Figure 5.3: Results of the proposed spectral subtraction speech enhancement algorithm for restaurant noise.

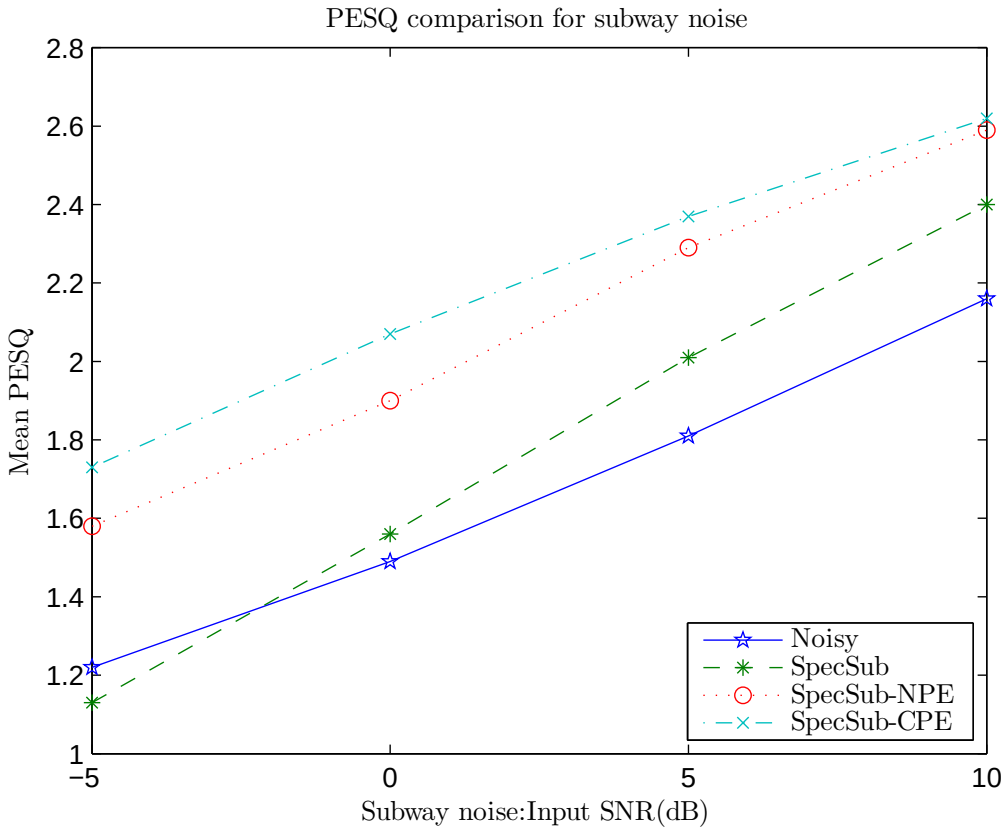


Figure 5.4: Results of the proposed spectral subtraction speech enhancement algorithm for subway noise.

From the above results, the effectiveness of the proposed noise estimation algorithm can be confirmed for the mentioned types of noises. For stationary noises like white noise, though initial noise estimation (noise estimation obtained by averaging first few silent frames of speech) might be sufficient for noise reduction in all future frames, the improvement in the speech quality with our algorithm for stationary noises is mainly due to less distortion of the dominant harmonic bins in the voiced frames. For other non-stationary noises like babble noise, the noise estimation even in the voiced frames results in effective noise tracking which provides further improvement of the speech quality. Also, it should be noted that for low SNR, the YIN algorithm detects only few voiced frames [70] and this limits the performance of proposed speech enhancement algorithm. Pitch estimation on clean speech improves the quality further.

5.2 MMSE STSA with the Proposed Noise Estimation Algorithm

The proposed noise estimation algorithm is combined with the traditional MMSE STSA algorithm and performance is evaluated in the following tables. The proposed noise estimation algorithm combined with MMSE algorithm is denoted as 'MMSE-NE'. Pitch estimation is carried out on both noisy speech and clean speech, and results are presented separately. MMSE-NE using pitch estimation on noisy speech is denoted as 'MMSE-NPE' and MMSE-NE using pitch estimation on clean speech is denoted as 'MMSE-CPE'.

5.2.1 Results and Analysis of Results

Table 5.5: PESQ evaluation of the proposed algorithm against the standard MMSE for white noise.

	PESQ			
Global SNR(in dB)	Noisy	MMSE	MMSE-NPE	MMSE-CPE
-5	1.22	1.56	1.60	1.64
0	1.43	2.01	2.03	2.02
5	1.72	2.47	2.36	2.34
10	2.05	2.83	2.67	2.62

Table 5.6: PESQ evaluation of the proposed algorithm against the standard MMSE for babble noise.

	PESQ			
Global SNR(in dB)	Noisy	MMSE	MMSE-NPE	MMSE-CPE
-5	1.32	1.41	1.56	1.69
0	1.66	1.85	1.99	2.11
5	2.02	2.26	2.34	2.41
10	2.38	2.59	2.63	2.70

Table 5.7: PESQ evaluation of the proposed algorithm against the standard MMSE for restaurant noise.

	PESQ			
Global SNR(in dB)	Noisy	MMSE	MMSE-NPE	MMSE-CPE
-5	1.35	1.42	1.56	1.68
0	1.66	2.81	1.99	2.08
5	2.00	2.16	2.34	2.42
10	2.34	2.46	2.63	2.69

Table 5.8: PESQ evaluation of the proposed algorithm against the standard MMSE for subway noise.

	PESQ			
Global SNR(in dB)	Noisy	MMSE	MMSE-NPE	MMSE-CPE
-5	1.22	1.36	1.64	1.76
0	1.49	1.68	2.01	2.12
5	1.81	2.05	2.37	2.46
10	2.16	2.40	2.67	2.74

In the above tables, the first column, 'Global SNR(in dB)', represents the signal-to-noise ratio after speech is corrupted. The second column, 'Noisy', gives the value of objective measure 'PESQ' for the corrupted speech. The third column indicates the PESQ value for speech enhanced using the traditional MMSE STSA algorithm. Similarly, the fourth and fifth columns give the values of PESQ measure for enhanced speech using the proposed approach with pitch estimation based on noisy and clean speech, respectively. The upper bound due to pitch estimation on clean speech can be observed from data in the tables for babble noise. We also give the graphical representation of the above tabulated performance comparison in figures 5.5, 5.6, 5.7 and 5.8, which follow.

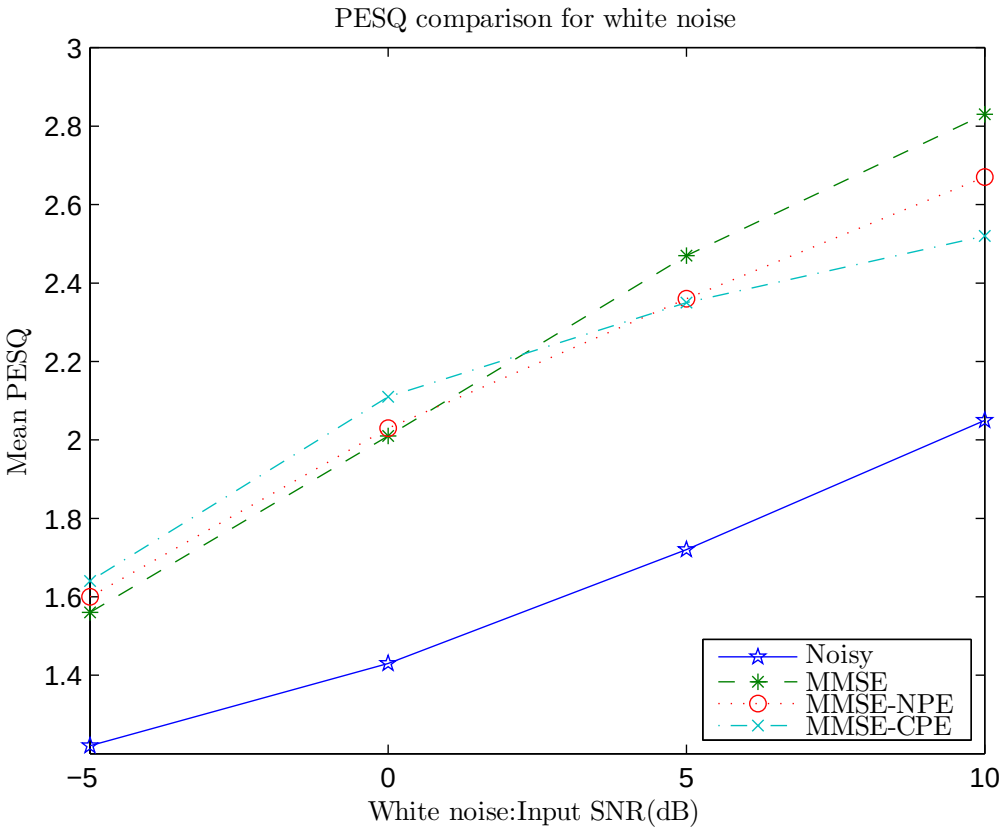


Figure 5.5: Results of the proposed MMSE STSA speech enhancement algorithm for white noise.

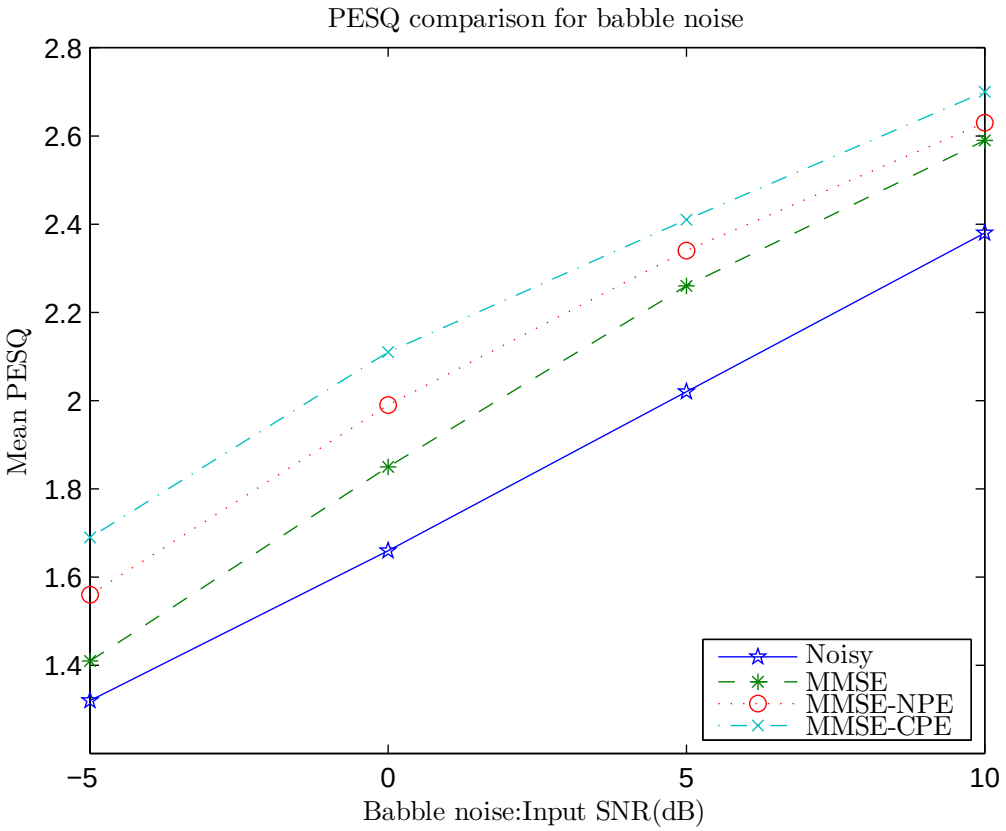


Figure 5.6: Results of the proposed MMSE STSA speech enhancement algorithm for babble noise.

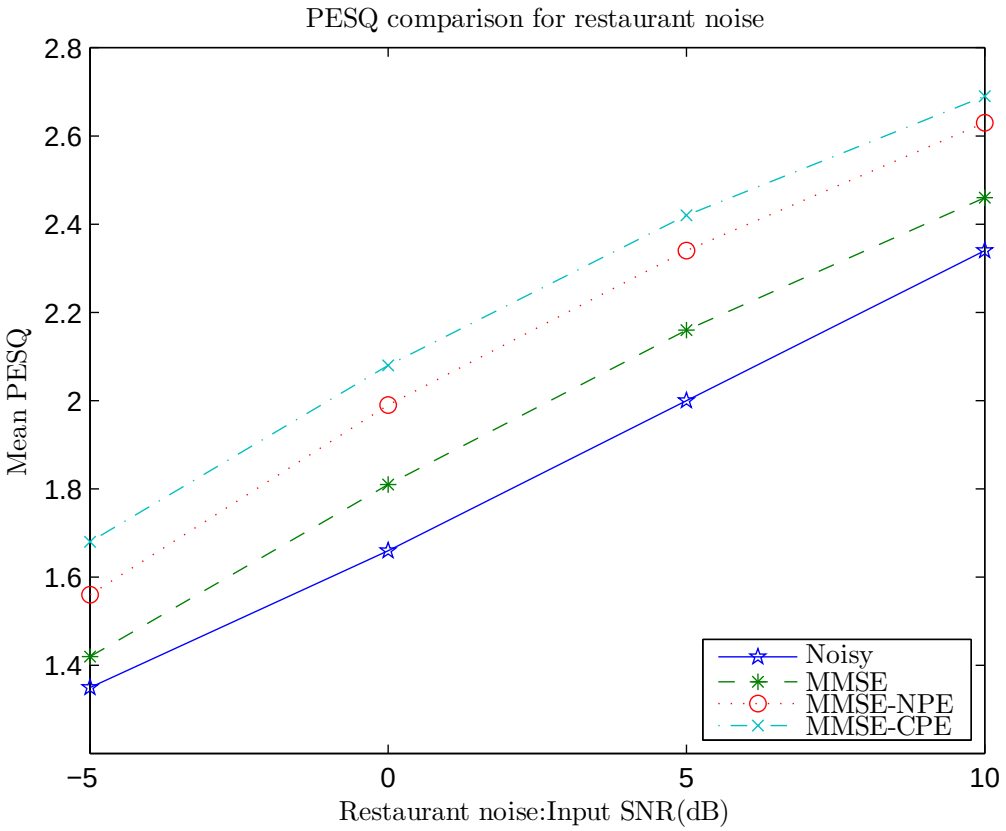


Figure 5.7: Results of the proposed MMSE STSA speech enhancement algorithm for restaurant noise.

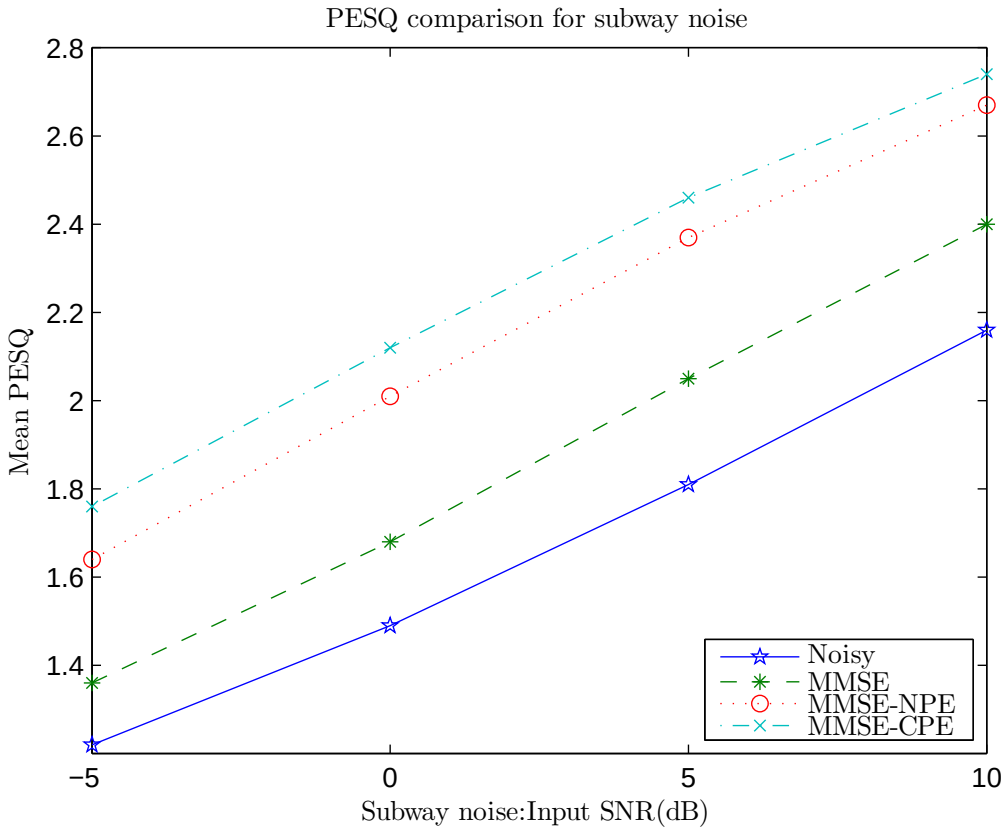


Figure 5.8: Results of the proposed MMSE STSA speech enhancement algorithm for subway noise.

Improvement is obtained for non-stationary noises as seen from figures 5.6, 5.7 and 5.8. However, white noise is a stationary noise type, and hence using proposed noise estimation does not result in improvement over the traditional MMSE noise reduction algorithm as seen in figures 5.5. We think this is because estimating the noise in the voiced frames further results in suppression of unvoiced speech, and overall speech quality decreases for stationary noises. While traditional MMSE can not respond to the non-stationary changes in the noise due to a decision-directed approach, since a priori SNR is averaged over successive frames [15], the proposed noise estimation results in better speech quality for highly non-stationary noises like babble noise. The MMSE algorithm is effective for eliminating the annoying musical noise artifact in the unvoiced frames, while spectral subtraction combined with the proposed noise estimation removes the noise in the voiced frames effectively and consistently. This motivates the combination of MMSE and the proposed spectral subtraction algorithm to improve the speech quality further with minimum musical noise. We present the result

of this fusion in the next section.

5.3 Combined Spectral Subtraction and MMSE STSA with the Proposed Noise Estimation Algorithm

Spectral subtraction provides high attenuation of background noise but with annoying musical noise effect. On the other hand, the MMSE STSA algorithm effectively eliminates the musical noise by smoothing a priori SNR across frames. Due to this averaging, the noise attenuation is lower as compared to spectral subtraction. Also, the MMSE STSA algorithm causes less speech distortion. These two contradictory behaviors of the spectral subtraction and MMSE STSA algorithms are combined to achieve maximum noise suppression in the low SNR periods during voiced frames and minimum musical noise in the non-voiced frames. In this fusion, non-voiced frames are processed by the basic MMSE-NE algorithm to minimize musical noise and in the voiced frames MMSE-NE and SpecSub-NE are combined to suppress the noise between harmonics with minimal speech distortion. As we have shown in the last section, MMSE STSA with the proposed noise estimation algorithm works well only for non-stationary noises. Therefore, this combination provides better speech quality only for non-stationary noises. The formulation of this combination is given below.

Let U and V denote the unvoiced and voiced frame detected by YIN algorithm, respectively, $|\hat{S}_{MMSE}(n,k)|$ and $|\hat{S}_{SpecSub}(n,k)|$ be the magnitude spectra of speech enhanced by the MMSE STSA and spectral subtraction rules:

$$|\hat{S}_{Fusion}(n,k)|^2 = \begin{cases} |\hat{S}_{MMSE}(n,k)|^2 & \text{if } |Y(n,k)| = U \\ & or \Delta\phi(n,k) < \phi_T \\ \hat{S}_{Comb} & \text{otherwise.} \end{cases} \tag{5.1}$$

where $\hat{S}_{Comb} = 0.8 * |\hat{S}_{SpecSub}(\lambda,\mu)|^2 + 0.2 * |\hat{S}_{MMSE}(\lambda,\mu)|^2$. We are using the contribution of the MMSE STSA enhanced spectra in the unvoiced speech and harmonic dominant bins in voiced speech to reduce the effect of annoying musical noise with minimum speech distortion. We use the spectral subtraction in the noise dominant speech for effective noise reduction in the voiced frame.

5.3.1 Results and Analysis of Results

Table 5.9: PESQ evaluation of the proposed algorithm for white noise when pitch is estimated from noisy speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-NPE	MMSE-NPE	Fusion-NPE
-5	1.22	1.07	1.56	1.63	1.60	1.57
0	1.43	1.48	2.01	2.05	2.03	1.96
5	1.72	1.97	2.47	2.41	2.36	2.36
10	2.05	2.45	2.83	2.67	2.67	2.69

Table 5.10: PESQ evaluation of the proposed algorithm for white noise when pitch is estimated from clean speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-CPE	MMSE-CPE	Fusion-CPE
-5	1.22	1.07	1.56	1.76	1.64	1.76
0	1.43	1.48	2.01	2.12	2.11	2.07
5	1.72	1.97	2.47	2.43	2.35	2.41
10	2.05	2.45	2.83	2.68	2.52	2.71

Table 5.11: PESQ evaluation of the proposed algorithm for babble noise when pitch is estimated from noisy speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-NPE	MMSE-NPE	Fusion-NPE
-5	1.32	1.21	1.41	1.47	1.56	1.50
0	1.66	1.73	1.85	1.99	1.99	1.96
5	2.02	2.17	2.26	2.38	2.34	2.32
10	2.38	2.57	2.59	2.66	2.63	2.60

Table 5.12: PESQ evaluation of the proposed algorithm for babble noise when pitch is estimated from clean speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-CPE	MMSE-CPE	Fusion-CPE
-5	1.32	1.21	1.41	1.76	1.69	1.98
0	1.66	1.73	1.85	2.17	2.11	2.30
5	2.02	2.17	2.26	2.47	2.41	2.58
10	2.38	2.57	2.59	2.72	2.70	2.84

Table 5.13: PESQ evaluation of the proposed algorithm for restaurant noise when pitch is estimated from noisy speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-NPE	MMSE-NPE	Fusion-NPE
-5	1.35	1.12	1.42	1.45	1.55	1.49
0	1.66	1.64	1.81	1.91	1.99	1.92
5	2.00	2.07	2.16	2.31	2.34	2.24
10	2.34	2.46	2.46	2.64	2.63	2.53

Table 5.14: PESQ evaluation of the proposed algorithm for restaurant noise when pitch is estimated from clean speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-CPE	MMSE-CPE	Fusion-CPE
-5	1.35	1.12	1.42	1.78	1.68	1.97
0	1.66	1.64	1.81	2.11	2.08	2.27
5	2.00	2.07	2.16	2.42	2.42	2.53
10	2.34	2.46	2.46	2.68	2.69	2.73

Table 5.15: PESQ evaluation of the proposed algorithm for subway noise when pitch is estimated from noisy speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-NPE	MMSE-NPE	Fusion-NPE
-5	1.22	1.13	1.36	1.58	1.64	1.66
0	1.49	1.56	1.68	1.90	2.01	2.01
5	1.81	2.01	2.05	2.29	2.37	2.36
10	2.16	2.40	2.40	2.59	2.67	2.64

Table 5.16: PESQ evaluation of the proposed algorithm for subway noise when pitch is estimated from clean speech.

	PESQ					
Global SNR(in dB)	Noisy	SpecSub	MMSE	SpecSub-CPE	MMSE-CPE	Fusion-CPE
-5	1.22	1.13	1.36	1.73	1.76	1.93
0	1.49	1.56	1.68	2.07	2.12	2.24
5	1.81	2.01	2.05	2.37	2.46	2.51
10	2.16	2.40	2.40	2.62	2.74	2.73

For better comparison of data in the above tables, results are shown separately for pitch estimation on noisy speech and on clean speech. In the above tables, the first column, 'Global SNR(in dB)', represents the signal-to-noise ratio after speech is corrupted. The second column, 'Noisy', gives the value of objective measure 'PESQ' for the corrupted speech. The remaining columns indicate the PESQ measure when noisy speech is processed by the mentioned algorithms. For each row in the above table, the value in the right-hand column, for Fusion-CPE, is the highest. We also give the graphical representation of the above tabulated performance comparison in figures 5.9-5.14 which follow.

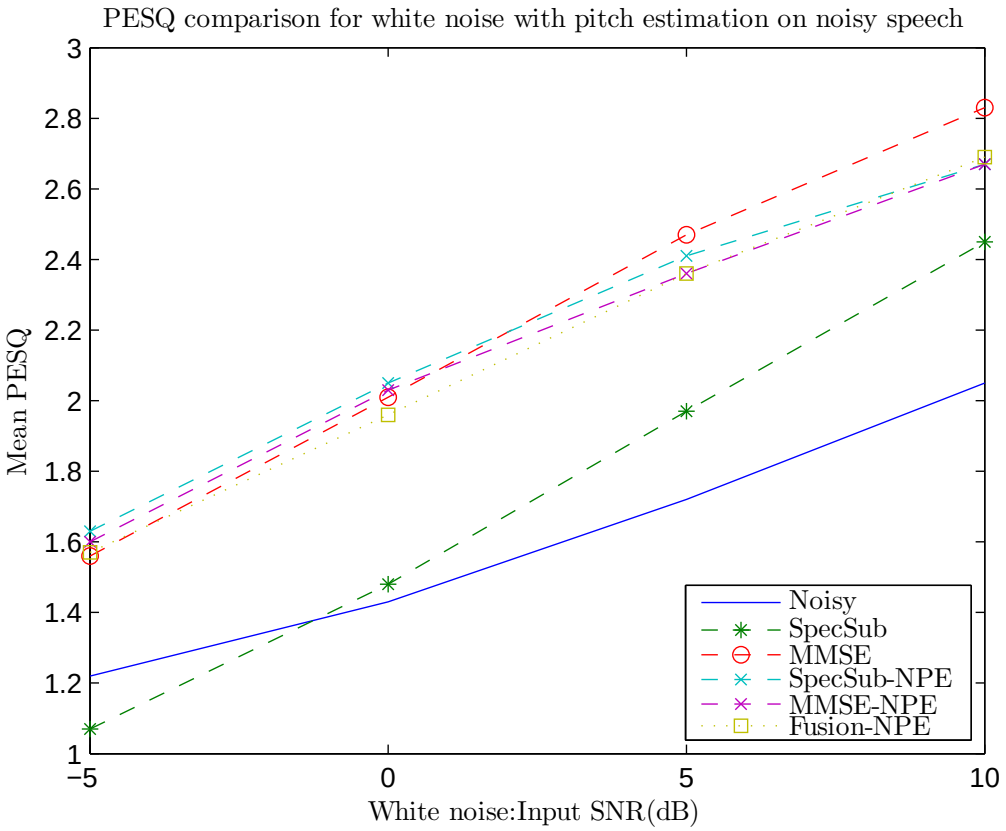


Figure 5.9: Results of the proposed fusion algorithm for white noise with pitch estimation on noisy speech.

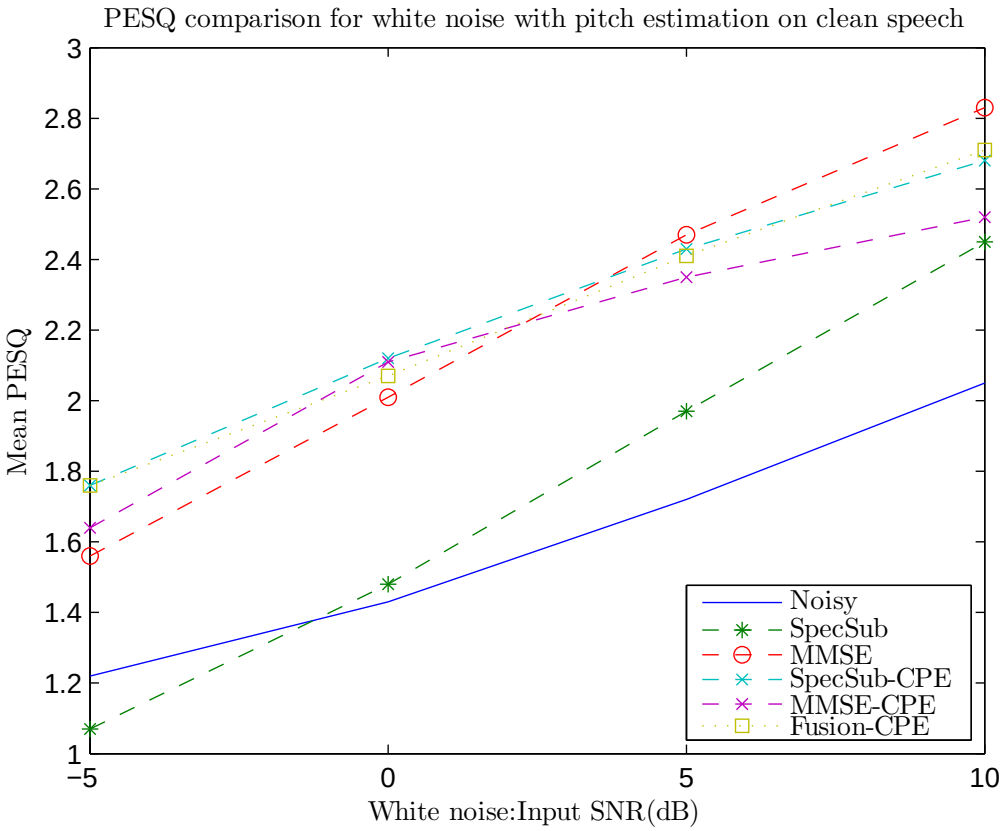


Figure 5.10: Results of the proposed fusion algorithm for white noise with pitch estimation on clean speech.

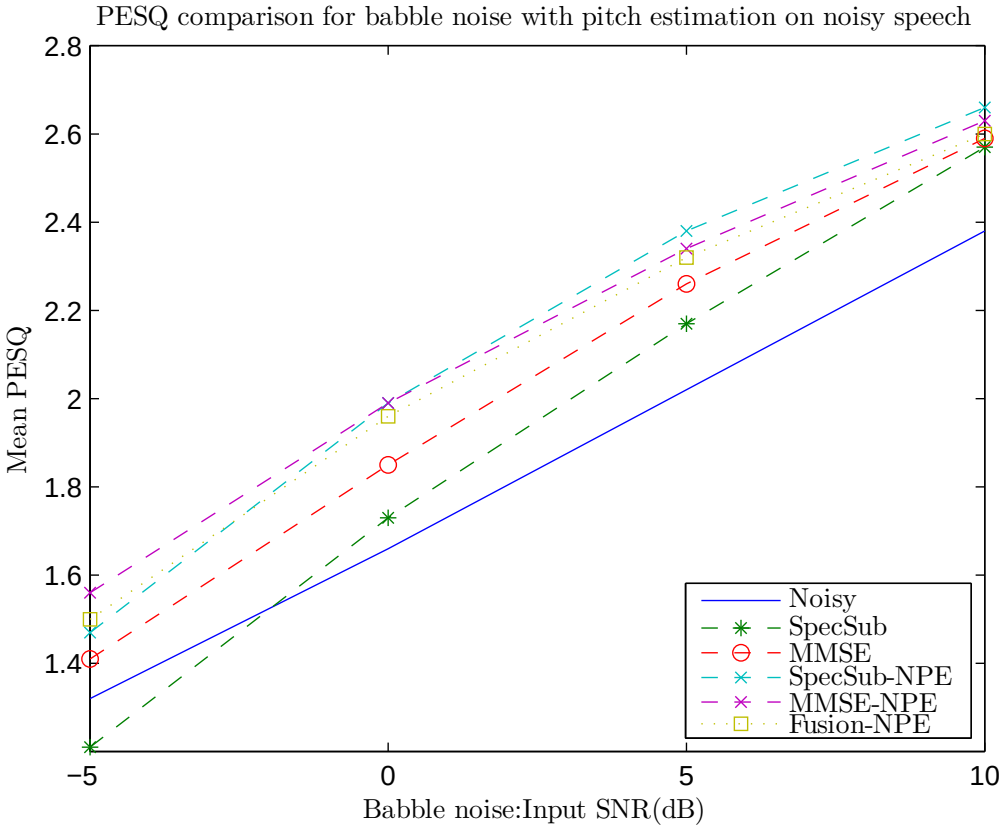


Figure 5.11: Results of the proposed fusion algorithm for babble noise with pitch estimation on noisy speech.

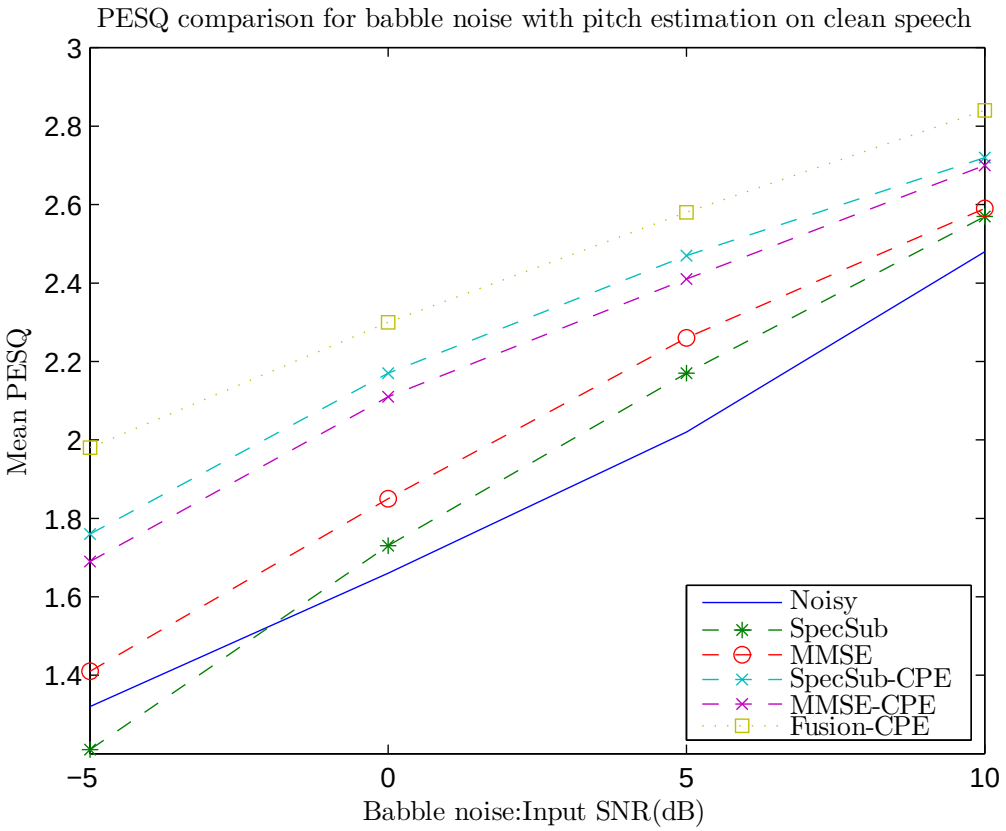


Figure 5.12: Results of the proposed fusion algorithm for babble noise with pitch estimation on clean speech.

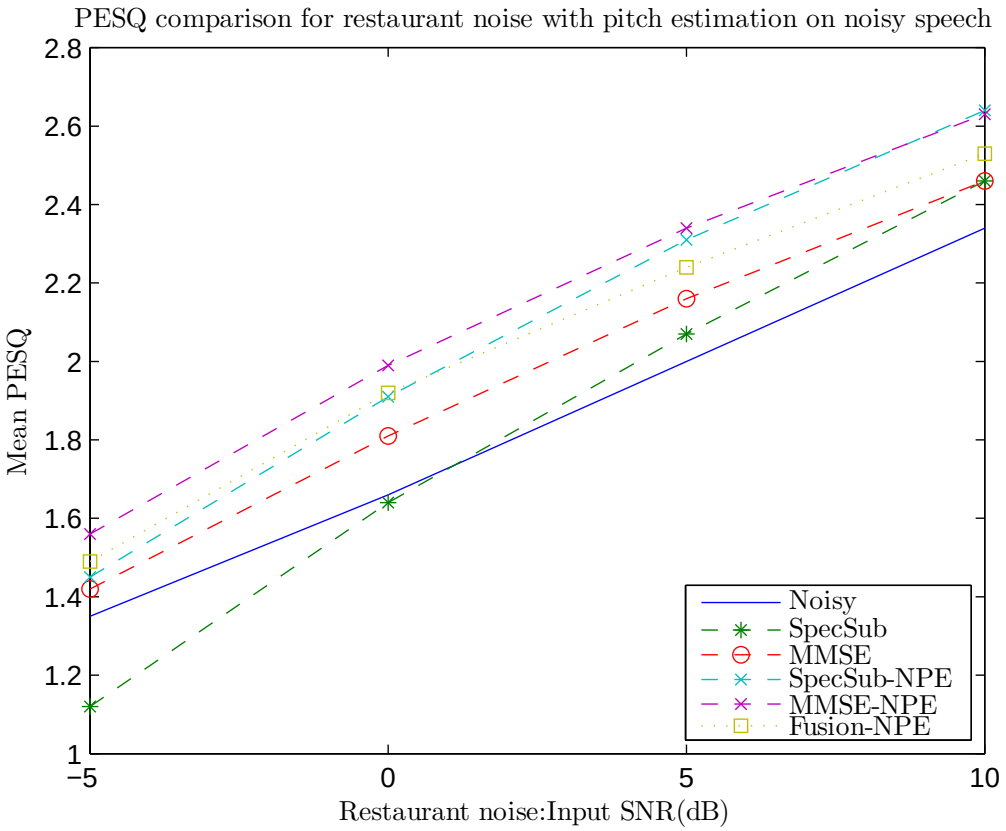


Figure 5.13: Results of the proposed fusion algorithm for restaurant noise with pitch estimation on noisy speech.

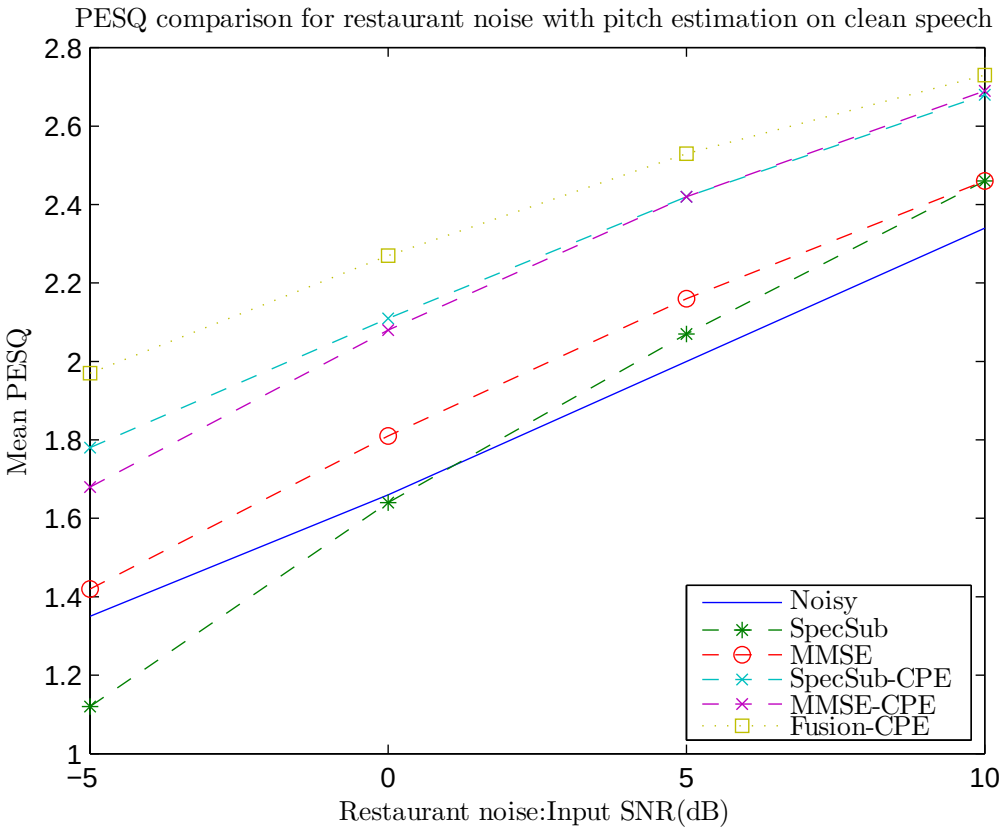


Figure 5.14: Results of the proposed fusion algorithm for restaurant noise with pitch estimation on clean speech.

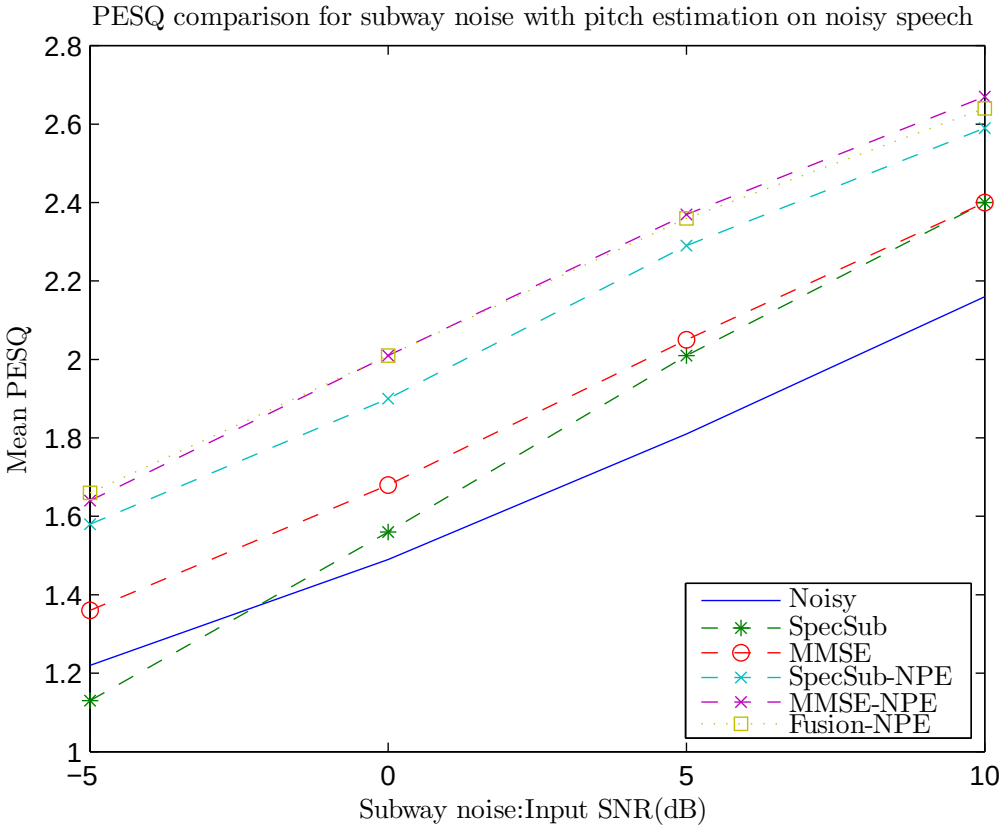


Figure 5.15: Results of the proposed fusion algorithm for subway noise with pitch estimation on noisy speech.

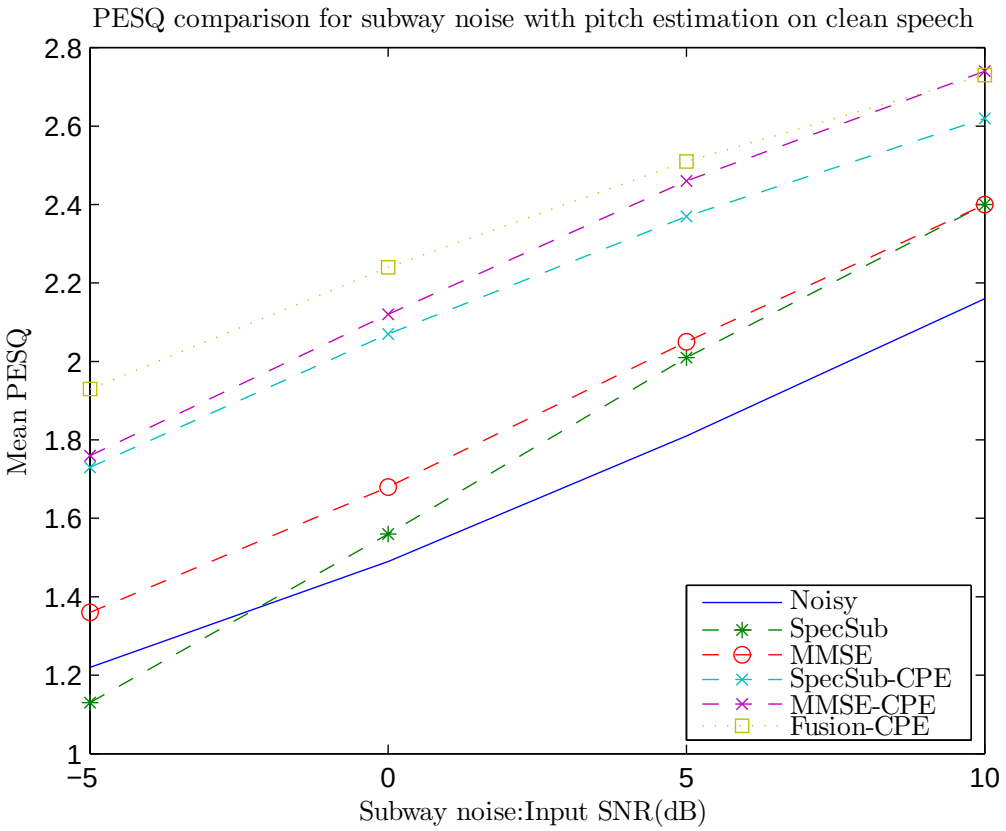
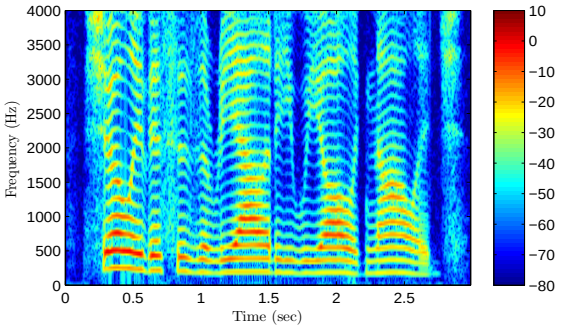


Figure 5.16: Results of the proposed fusion algorithm for subway noise with pitch estimation on clean speech.

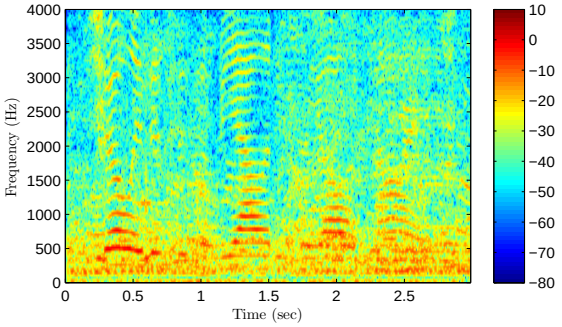
As indicated in the above figures, good performance of this combined algorithm is dependent on good estimation of pitch for the voiced speech and non-stationary noise. As inaccurate pitch estimation will remove excessive amount of signal due to spectral subtraction in the voiced region, and overall speech quality is reduced as seen in figures 5.11, 5.13 and 5.15. However, improvement is significant when pitch is estimated using the clean speech for non-stationary noise as seen in figure 5.10 and 5.12. Accurate pitch estimation using some advanced pitch estimation algorithm would result in better speech quality.

5.4 Spectrogram Based Comparison

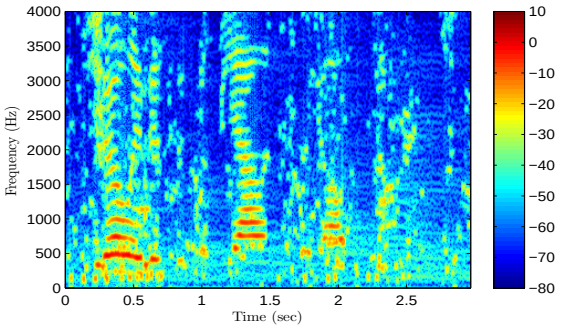
Below, we have shown the spectrograms for all of the above mentioned algorithms. Clean speech (See in Fig. 5.17a) is degraded by adding babble noise at 0 dB as shown in Fig. 5.17b .



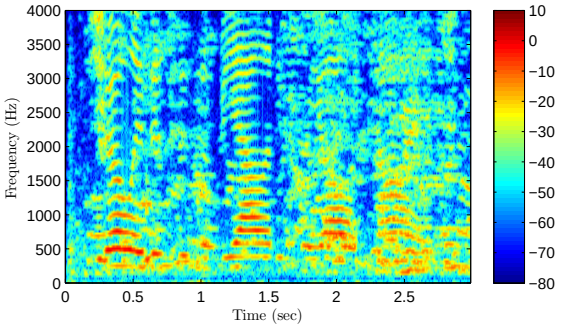
(a) Clean speech spectrogram



(b) Noisy speech spectrogram

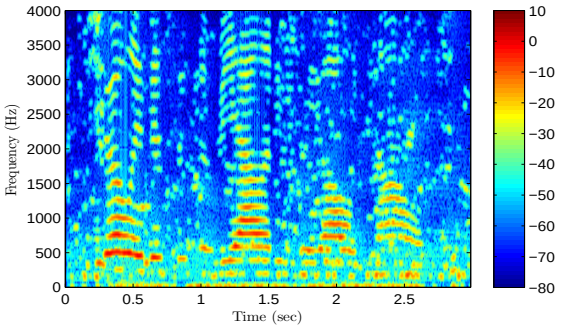


(c) Spectrogram for SpecSub processed speech.

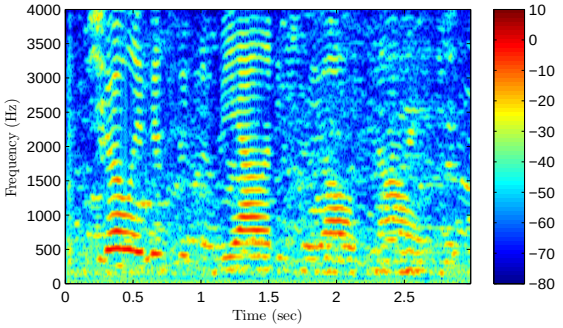


(d) Spectrogram for MMSE processed speech.

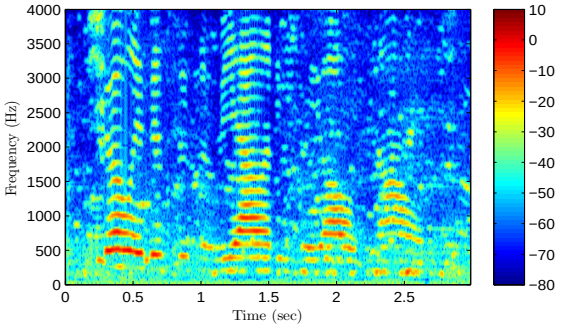
Figure 5.17: Spectrograms of enhanced speech processed by the discussed algorithms.



(e) Spectrogram for SpecSub-CPE processed speech.



(f) Spectrogram for MMSE-CPE processed speech.



(g) Spectrogram for Fusion-CPE processed speech.

Figure 5.17: (Continued).

The effectiveness of the proposed approach can also be confirmed by the spectrograms of the processed speech as shown in Fig. 5.17. SpecSub-CPE processed speech has less speech distortion as compared to SpecSub processed speech as shown in Fig. 5.17c and Fig. 5.17e. Also, MMSE-CPE results in better noise reduction than the standard MMSE algorithm as shown in Fig. 5.17d and Fig. 5.17f. Fusion-CPE suppresses the noise present between the harmonics effectively as shown in Fig. 5.17g. Fusion-CPE utilizes the noise suppression properties of the spectral subtraction rule and the minimum musical noise reduction capability of MMSE.

Chapter 6

CONCLUSIONS AND FUTURE WORK

The purpose of this chapter is two-fold, first is to draw some conclusions based on our discussion, in previous chapters, and then to propose some future areas of research.

6.1 Conclusions

In chapter 2, we discussed some of the existing speech enhancement algorithms including the spectral subtraction, MMSE STSA, modulation domain based speech enhancement and phase enhancement using the harmonic model for voiced speech. These algorithms attempt to improve the quality of speech with minimum speech distortion and maximum possible noise reduction. Due to inaccurate noise estimation, performance of these algorithms is limited and also artifacts are introduced in the processed speech. Various measures to quantify speech quality are discussed in chapter 3.

In this work, we have used the harmonic model for voiced speech to estimate the noise even in voiced frames. The harmonic model is used to estimate the frame to frame phase difference for the clean speech, and this knowledge is exploited to track the noise in the voiced speech. This approach for noise estimation has been shown to improve the performance of traditional spectral subtraction significantly for white, babble, restaurant and subway noises. We also showed the effectiveness of

this technique when used with MMSE STSA for non-stationary noise reduction. Thus, the proposed technique of noise estimation can be combined with any of the existing amplitude enhancement algorithms to further improve the performance in presence of non-stationary noises. Combined spectral subtraction and MMSE further improved the quality of speech with minimum musical noise and maximum possible noise reduction when good estimate of pitch is available. For non-stationary noises, average PESQ improvement of spectral subtraction with new noise estimation is 0.3 when pitch is estimated on noisy speech. When pitch estimation is based on clean speech, PESQ is increased to 0.5. MMSE with new noise estimation gives an average PESQ improvement of 0.2 when pitch is estimated on noisy speech and 0.3 when it is estimated on clean speech. With fusion of these two algorithms average PESQ improvement is pushed further to 0.4 (over traditional MMSE) when pitch is estimated on clean speech.

We have used the YIN fundamental frequency tracking algorithm to estimate the pitch for the voiced frames. The performance of this algorithm degrades in the low SNR conditions, resulting in less number of voiced frames detected. Better results can be obtained by using some more advanced pitch estimation algorithms. Though we are estimating the frame to frame phase difference for clean speech, this knowledge has not been used to carry out the actual phase enhancement for the noisy speech. Phase difference is just used as an additional means to estimate the noise.

6.2 Future Work

In this section, we discuss some of the drawbacks of the proposed approach and further scope of research to improve it.

It has been shown that phase enhancement using the harmonic model for the voiced speech results in improved speech quality [22]. However, this also results in additional artifacts in the processed speech due to inaccurate harmonic modeling for voiced speech. If the harmonic model is improved further then the proposed noise estimation algorithm can be used to enhance the amplitude spectrum along with the phase estimation. This combination of amplitude and phase enhancement should result in better speech quality.

We have implemented noise estimation using the phase difference as an additional means in the acoustic domain. Integrating this approach in the modulation domain might result in better speech quality, as modulation domain speech enhancement is already superior to even MMSE STSA.

Modulation domain speech enhancement consists of two STFTs, namely acoustic STFT and modulation STFT. The technique we used in this work can be used to estimate the phase in the acoustic STFT, but phase estimation in modulation STFT is still challenging and it is very important for speech perception in our auditory system [24].

Also, the knowledge of phase difference to detect the noise-dominant frequency bins in the voiced frames can be used to improve the performance of the existing noise estimation algorithms, as most of them do not use the noise-robust harmonicity property of the voiced speech.

BIBLIOGRAPHY

[1] T. S. Gunawan, E. Ambikairajah, and J. Epps, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," *Speech Communication*, vol. 52, no. 5, pp. 381–393, 2010.

[2] G. Kim and P. Loizou, "Why do speech-enhancement algorithms not improve speech intelligibility?," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4738–4741, March 2010.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing.*, vol. 27, no. 2, pp. 113–120, 1979.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters.*, vol. 9, no. 1, pp. 12–15, 2002.

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

[6] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1875–1878, June 2000.

[7] S. Rangachari, P. Loizou, and Y. Hu, "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 305–308, May 2004.

[8] A. Milani, G. Kannan, I. Panahi, and R. Briggs, "A multichannel speech enhancement method for functional mri systems using a distributed microphone array," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6946–6949, 2009.

[9] A. Borowicz and A. Petrovsky, "Incorporating the human hearing properties into multichannel speech enhancement," in *Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA)*, pp. 1–6, 2011.

[10] J. Jensen, J. Benesty, M. Christensen, and S. Jensen, "Non-causal time-domain filters for single-channel noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1526–1541, 2012.

[11] J. Benesty, M. Souden, and C. Jingdong, "A study of multichannel noise reduction linear filters in the time domain," in *IEEE International Conference on Signal Processing, Communications and Computing*, pp. 1–6, 2011.

[12] N. Mohammadiha and A. Leijon, "Nonnegative hmm for babble noise derived from speech hmm: Application to speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 998–1011, 2013.

[13] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.

[14] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

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

[16] S. Chehresa and M. Savoji, "Mmse speech enhancement using gmm," in *International Symposium on Artificial Intelligence and Signal Processing*, pp. 266–271, 2012.

[17] B. Fodor and T. Fingscheidt, "Mmse speech enhancement under speech presence uncertainty assuming (generalized) gamma speech priors throughout," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4033–4036, 2012.

[18] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Transactions onSpeech and Audio Processing,*, vol. 8, no. 5, pp. 497–507, 2000.

[19] J. Jensen and R. Heusdens, "Improved subspace-based single-channel speech enhancement using generalized super-gaussian priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 862–872, 2007.

[20] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[21] K. Paliwal, K.Wojcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, 2010.

[22] M. Krawczyk and T. Gerkmann, "Stft phase improvement for single channel speech enhancement," in *International Workshop on Acoustic Signal Enhancement*, pp. 1–4, September 2012.

[23] P. Philipos, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[24] K. Paliwal and L. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 52, pp. 153–170, February 2005.

[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, 1979.

[26] Z. Goh, K. Tan, and T. Tan, "Postprocessing method for suppressing noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing.*, vol. 6, no. 3, pp. 287–292, 1998.

[27] P. Crozier, B. Cheetham, C. Holt, and E. Munday, "Speech enhancement employing spectral subtraction and linear predictive analysis," *Electronics Letters*, vol. 29, no. 12, pp. 1094–1095, 1993.

[28] J. Seok and K. Bae, "Reduction of musical noise in spectral subtraction method using subframe phase randomisation," *Electronics Letters*, vol. 35, no. 2, pp. 123–125, 1999.

[29] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 1, pp. 648–651, 2003.

[30] J. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795–805, 1991.

[31] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[32] L. Zadeh, "Frequency analysis of variable networks," *Proceedings of the IRE*, vol. 38, no. 3, pp. 291–299, 1950.

[33] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[34] S. Bacon and D. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2575–2580, 1989.

[35] S. Sheft and W. Yost, "Temporal integration in amplitude modulation detection," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 796–805, 1990.

[36] C. Schreiner and J. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. i. the anterior auditory field (aaf)," *Hearing Research*, vol. 21, no. 3, pp. 227–241, 1986.

[37] N. M'Sirdi and J. Zarader, "Adaptive comb filters for enhancement of quasi periodic signals," in *International Conference on Acoustics, Speech, and Signal Processing.*, pp. 1461–1464, 1990.

[38] W. Jin, X. Liu, M. Scordilis, and H. Lu, "Speech enhancement using harmonic emphasis and adaptive comb filtering," *IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 18, no. 2, pp. 356 – 368, 2010.

[39] H. Kasuya, S. Ogawa, and Y. Kikuchi, "An adaptive comb filtering method as applied to acoustic analyses of pathological voice," in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 11, pp. 669–672, 1986.

[40] D. Malah and R. Cox, "A generalized comb filtering technique for speech enhancement," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, pp. 160–163, 1982.

[41] T. Yoshioka, T. Nakatani, and H. Okuno, "Noisy speech enhancement based on prior knowledge about spectral envelope and harmonic structure," in *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 4270–4273, 2010.

[42] Y. A.-T. Yu and H. Wang, "New speech harmonic structure measure and it application to post speech enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 29–32, 2004.

[43] E. Zavarehei, S. Vaseghi, and Y. Qin, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1194–1203, 2007.

[44] Y. Stark and J. Tabrikian, "Mmse-based speech enhancement using the harmonic model," in *IEEE 25th Convention of Electrical and Electronics Engineers*, pp. 626–630, 2008.

[45] T. Selvi and J. Pragatheeswaran, "Efficient speech enhancement technique by exploiting the harmonic structure of voiced segments," in *International Conference on Recent Trends in Information Technology*, pp. 764–769, 2011.

[46] C. Eunjoon, J. Smith, and B. Widrow, "Exploiting the harmonic structure for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4569–4572, 2012.

[47] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement; unimportant, important, or impossible," in *IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, pp. 1–5, 2012.

[48] H. Yi and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions onAudio, Speech, and Language Processing.*, vol. 16, no. 1, pp. 229–238, 2008.

[49] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T*, 2003.

[50] W. A. Munson and J. E. Karlin, "Isopreference method for evaluating speech-transmission circuits," *The Journal of the Acoustical Society of America*, vol. 34, no. 6, pp. 762–774, 1962.

[51] M. Hecker and C. Williams, "Choice of reference conditions for speech preference tests," *The Journal of the Acoustical Society of America*, vol. 39, no. 5A, pp. 946–952, 1966.

[52] P. Combescure, A. Guyader, and A. Gilloire, "Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, pp. 988–991, 1982.

[53] D. Goodman and R. Nash, "Subjective quality of the same speech transmission conditions in seven different countries," *IEEE Transactions on Communications*, vol. 30, no. 4, pp. 642–654, 1982.

[54] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 2, pp. 204–207, 1977.

[55] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments," *IEE Proceedings in Communications, Speech and Vision*, vol. 136, no. 5, pp. 317–324, 1989.

[56] R. Kubichek, D. Atkinson, S. Voran, J. Lansford, H. Li, and J. Schroeder, "Advances in objective voice quality assessment," in *IEEE 42nd Vehicular Technology Conference.*, vol. 1, pp. 155–158, 1992.

[57] D. Richards, "Speech-transmission performance of p.c.m. systems," *Electronics Letters*, vol. 1, no. 2, pp. 40–41, 1965.

[58] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, 2001.

[59] P. Yong, S. Nordholm, and H. Dam, "Noise estimation with lowcomplexity for speech enhancement," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 109–112, 2011.

[60] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *IEEE Workshop on Speech Coding for Telecommunications*, pp. 85–86, 1993.

[61] K. El-Maleh and P. Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems," in *IEEE Canadian Conference on Electrical and Computer Engineering Innovation: Voyage of Discovery*, vol. 2, pp. 470–473, 1997.

[62] D. Freeman, G. Cosier, C.B.Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 369–372, 1989.

[63] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on*, vol. 3, pp. 321–324, 1993.

[64] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[65] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 153–156, 1995.

[66] R. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing.*, pp. 4266–4269, 2010.

[67] A. Cheveign and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[68] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[69] K. Sorensen, V.Karsten, and S. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," *EURASIP Journal on Advances in Signal Processing.*, vol. 2005, no. 18, pp. 2954–2964, 2005.

[70] D. Sharma and P. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *European Signal Processing Conference*, 2009.