EVALUATION OF SIGNAL PROCESSING METHODS FOR SPEECH ENHANCEMENT

BY

MAHIKA DUBEY

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Electrical and Computer Engineering
at the University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Professor Paris Smaragdis

ABSTRACT

This thesis explores some of the main approaches to the problem of speech signal enhancement. Traditional signal processing techniques including spectral subtraction, Wiener filtering, and subspace methods are very widely used and can produce very good results, especially in the cases of constant ambient noise, or noise that is predictable over the course of the signal. We first study these methods and their results, and conclude with an analysis of the successes and failures of each. Comparisons are based on the effectiveness of the methods of removing disruptive noise, the speech quality and intelligibility of the enhanced signals, and whether or not they introduce some new artifacts into the signal. These characteristics are analyzed using the perceptual evaluation of speech quality (PESQ) measure, the segmental signal-to-noise ratio (SNR), the log likelihood ratio (LLR), and weighted spectral slope distance.

Keywords: Signal Processing, Speech Enhancement

To my parents Smita and Abhay Dubey,
my sister Ambika, and my brother Akash.

ACKNOWLEDGMENTS

I would like to thank my adviser Professor Paris Smaragdis and my graduate student mentor Ramin Anushiravani for their guidance and expertise over the last year and a half. This thesis would not have been possible without their assistance. I would also like to acknowledge my family for their love and encouragement throughout my undergraduate years, and my classmates and friends for the company, advice, and many memories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

STFT	Short-Time Fourier Transform
VAD	Voice Activity Detection
LMS	Least Mean Squares
SVD	Singular Value Decomposition
EVD	Eigenvalue Decomposition
WGN	White Gaussian Noise
SNR	Signal-to-Noise Ratio
PESQ	Perceptual Evaluation of Speech Quality
LLR	Log Likelihood Ratio
WSS	Weighted Spectral Slope

CHAPTER 1

INTRODUCTION

Speech signal enhancement is performed in many systems used today. Speech recognition and speech-to-text services such as those found in smart phones require the ability to uncover clean speech from a signal that was recorded in a noisy environment. Music recognition softwares require high quality signals to be able to identify songs and artists, and so need to be able to filter out unnecessary ambient noise from a recording.

Figure 1.1 models the high level goal of speech enhancement; we want to be able to extract a high quality clean signal from a given noisy signal. Signal processing methods are commonly used to achieve this. Some of the most popular algorithms include spectral subtraction, Wiener filtering, and subspace enhancements. We will detail each of these methods in the following chapters and conclude with a discussion of the performance of each on various corrupted test signals.

While signal processing is often very effective, there are some issues that come about from its use, including the inability to remove non-stationary noise, and the inherent inability to respond to very harsh corruption in signals. As a result, machine learning approaches to this problem have been successfully applied, and we will discuss some of the theory and reasoning behind this.

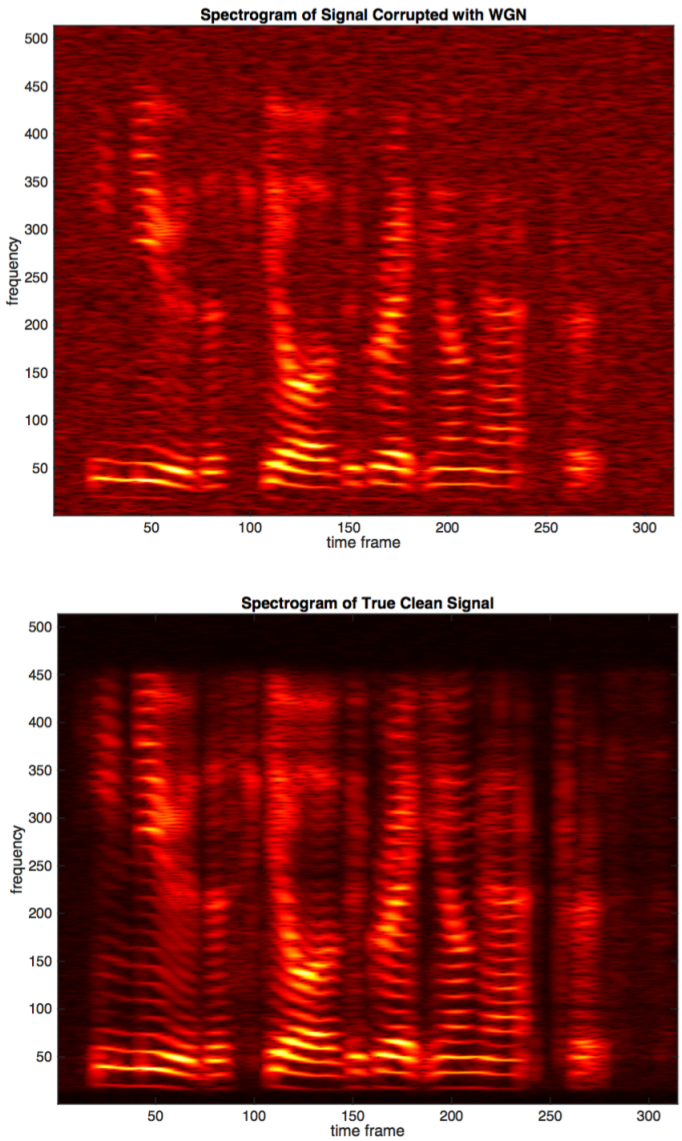

Figure 1.1: Spectrograms showing the STFTs of a signal with (top) and without (bottom) corruption

1.1 Short-Time Fourier Transform (STFT)

An important step in any frequency domain speech enhancement algorithm is finding the STFT of the noisy signal at hand. We divide the signal into multiple frames of the same size and find the Fourier transform of every frame. Each frame is windowed (usually using a Hamming window) and there is some overlap between frames used so as to ensure there is no information

loss during transformation and reconstruction. Below is the mathematical definition of STFT of a time domain signal $y(n)$.

$$Y(\omega) = \sum_{n=-\infty}^{\infty} y(n)w(n - mR)e^{-j\omega n}$$

where $y(n)$ is the time domain signal at time n, $w(n)$ is the windowing function, and R is the number of samples between each frame [1].

1.2 Noise Estimation

Every algorithm discussed in this thesis relies on some form of noise estimation to enhance a given signal. The method of spectral subtraction inherently requires some knowledge of the noise profile, as it must be subtracted from the noisy signal to receive the clean signal. In most situations, we are not given a noise profile and so must construct one of our own using the noisy signal. The most widely use approach is to average the first few frames of the noisy signal, as we can assume that the recording will contain a few milliseconds of ambient noise before the speaker starts speaking. Once we have taken the STFT of the noisy signal, we can simply take an average of the first few frames and keep the resulting signal information aside as the noise spectrum.

Similarly to spectral subtraction, most Wiener filtering algorithms choose to assume that the first few frames of a speech recording are a good estimate for ambient noise. These frames are averaged to construct a profile for the assumed noise. Some approaches to Wiener filter even update, or add to the noise profile by identifying segments of the signal where there is no speech while processing each frame. This is accomplished by estimating filter coefficients at every frame of the signal, allowing for a progressively more accurate filter [2]. Another way this can be accomplished is with voice activity detection (VAD), wherein the power of a signal is checked to differentiate between segments with high magnitude (usually this means a speaker is speaking), and regions with low magnitude (where ambient noise is most prevalent) [3].

This updating is one feature that makes the Wiener approach adaptive and helps improve the error over time.

Noise estimation in subspace enhancement is slightly different than the previous two algorithms. Subspace enhancement takes advantage of the matrix representation of signals. Matrices are divided into subspaces whereby noise is approximated by the smallest eigenvalues or singular values and speech is approximated by the rest [4].

1.3 Phase Estimation

A more recently explored pitfall of signal processing techniques is the application of the original noisy spectrum's phase information to the enhanced signal's spectrum before finding the time domain signal using the inverse STFT. As an ideal noise profile of a signal is not usually available, the phase information from the original signal is usually assumed to be valid for the cleaned signal as well [5]. However, this may not always be the case because the two signals can often be quite different due to the removal of noise. Geometric approaches to speech enhancement take this into account and perform some manipulation on the phase information as well as the magnitude information of the signal in order to produce a better quality enhancement [6].

1.4 Musical Noise and Reduction

One of the biggest issues with any speech enhancement algorithm is the introduction of musical noise as a result of the subtraction of the noise from a signal [7]. Specifically in spectral subtraction, when we subtract the noise spectrum from each frame of the noisy signal's STFT, there arises the possibility of creating of some negative numbers. Upon reconstruction of the enhanced signal using the inverse STFT, these negative values become random noises that are inconsistent with the overall signal. These introduced artifacts can be audibly disorienting and reduce the quality of the enhancement. Similar effects are seen in signals enhanced using Wiener filtering and

subspace methods as a result of overfiltering and too much removal. Figure 1.2 shows the spectrogram of an an enhanced signal showing the existence of musical noise. One basic method of handling musical noise is to manually alter the signal and set negative values (and optionally, very small values determined by some threshold) produced after the subtraction to zero before performing the inverse transform. While this may result in some information loss, it is usually trivial compared to the qualitative benefits. Other methods include creating filters that aim to remove musical noise from a processed signal, or applying weighting functions to different parts of a signal to minimize the effect of musical noise [8].

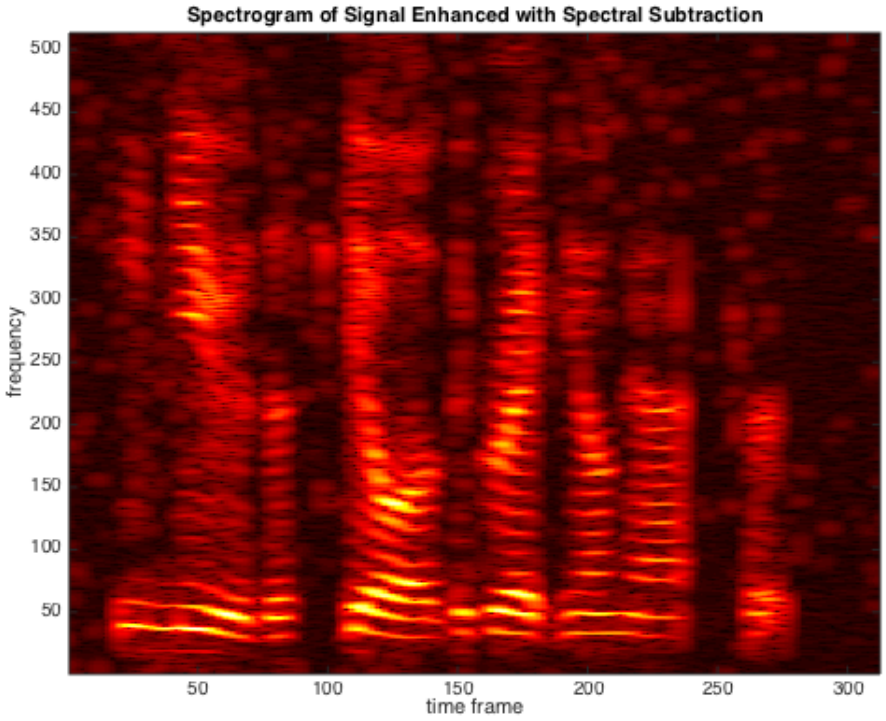

Figure 1.2: Spectrogram showing the STFT of a WGN corrupted signal enhanced with spectral subtraction. Musical noise is visible in the spectrum.

CHAPTER 2

SPECTRAL SUBTRACTION

One of the oldest and most popular signal processing algorithms for speech signal de-noising is spectral subtraction. While this algorithm is effective for most applications of speech enhancement, there are some inherent shortcomings with its ability to effectively remove noise, including the production of musical noise and issues with deleting noise that is dependent on the speaker. This process, at a high level, involves finding an estimate for assumed additive and uncorrelated noise, and subtracting it from the original signal to get a clean signal without any noise or unnecessary artifacts [4]. The model of the problem is this: We are given a time domain signal $y(n)$, which is the combination of speech $x(n)$ and some disruptive noise $d(n)$ at time frames n,

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

We want to extract the noiseless speech signal $x(n)$, but this is difficult in the time domain as the noise is not so easily distinguishable from the speech that we want to retrieve, so we take the Fourier transform to put the signal in frequency domain. The result is

$$Y(\omega) = X(\omega) + D(\omega)$$

Given this form, we can easily find the magnitude of the clean speech signal by subtracting the noise profile from the corrupted signal. This gives us the clean signal in frequency domain,

$$X(\omega) = Y(\omega) - D(\omega)$$

Now if we take the inverse transform of the clean spectrum we get $x(n)$. Before going into the details of the algorithm, we discuss some important background information.

2.1 Spectral Subtraction Algorithm

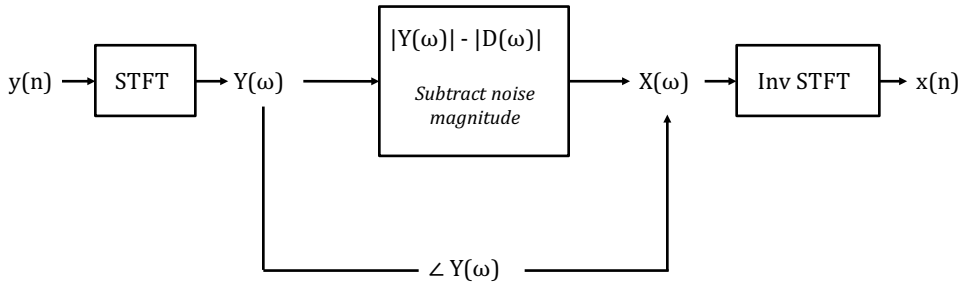

Figure 2.1: Block diagram of the spectral subtraction process.

Figure 2.1 outlines spectral subtraction at a high level. Below we outline the steps in detail for signal enhancement using Spectral Subtraction.

1. Find the STFT $Y(\omega)$ of noisy signal $y(n)$.
2. Save phase information $\angle Y(\omega)$ from STFT of noisy signal.
3. Estimate noise magnitude $D(\omega)$ from initial few frames of noisy signal spectrum.
4. Subtract noise from each frame of noisy spectrum to get clean signal $X(\omega)$.
5. Set negative values in $X(\omega)$ to zero to prevent musical noise.
6. Apply phase information $\angle Y(\omega)$ to cleaned signal $X(\omega)$.
7. Find the inverse STFT of the cleaned signal to get the cleaned signal $x(n)$ in the time domain.

CHAPTER 3

WIENER FILTERING

Wiener filtering uses a mathematical approach to decrease error between true clean speech and algorithmically enhanced speech signal. This approach aims to minimize the mean-square error to get a better estimate of the noise-free speech signal [4]. As such, the Wiener filter can be called an adaptive least mean squares (LMS) filter. This method is often more effective than spectral subtraction, especially in the cases where the assumptions of noise being constant and additive do not hold. However, this method does assume zero mean noise that is mostly uncorrelated with the signal of interest. The model of the problem is this: we are given a signal $y(n)$, and want to remove the noise $d(n)$ to recover the clean signal $x(n)$,

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

Wiener filtering is applicable in both the time and frequency domains. In the time domain, we construct a filter $h(n)$ from the autocorrelation matrix of the noise signal, and the cross-correlation vector of the noisy and clean signals. We now apply this filter to the noisy signal, as shown in figure 3.1.

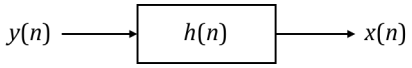

Figure 3.1: Wiener filtering in the time domain

The resulting signal, $x(n)$, is the enhanced signal, with noise removed. The process is mostly similar when performing Wiener filtering in the frequency domain. We first find the Fourier transform of the noisy signal,

$$Y(\omega) = X(\omega) + D(\omega)$$

We do not always have access to a clean signal, so we estimate the noise from segments of the signal without speech, and infer a filter from the estimated

noise and clean signals. Once we have this estimate, we construct a filter, $H(\omega)$, designed to remove this noise from the signal, and apply it to every frame to allow for an enhanced signal that is statistically closer to the true clean signal.

$$X(\omega) = H(\omega)\, Y(\omega)$$

The inverse Fourier transform can be applied to $X(\omega)$ to get the enhanced, or denoised, signal $x(n)$.

3.1 Time Domain Noise Removal Algorithm

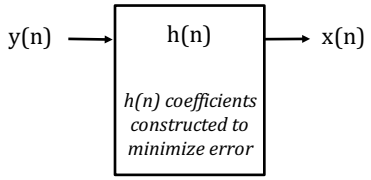

Figure 3.2: Block diagram of Wiener filtering process in the time domain.

Figure 3.2 shows the high level process of Wiener filtering in the time domain. Below we enumerate these steps, including details on how to construct the LMS filter $h(n)$.

1. Identify the error of approximation at time frame n as

   $$e(n) = x(n) \text{ - } \hat{x}(n)$$

   where $\hat{x}(n)$ can be replaced with $\mathrm{h}^T y(n)$, or the result of filtering the noisy signal.

2. Find the mean squared error value, which is ultimately to be minimized,

   $$J = E\left[e^2(n)\right]$$

   which we can expand by replacing $e^2(n)$ with $\mathrm{x}(n)$ - $\mathrm{h}^T y(n)$ to get

   $$J = E\left[x^2(n)\right] \text{ - } 2\mathrm{h}^T \mathrm{r}_{yx}^{-} + \mathrm{h}^T \mathrm{R}_{yy} \mathrm{h}$$

such that r_{yx}^{-} is the cross correlation vector between the noisy and clean signal, and R_{yy} is the autocorrelation matrix of the noisy signal.

3. Minimum error is reached when the derivative of J is zero, so we compute the derivative with respect to h to find the necessary filter,

$$\frac{\partial J}{\partial \mathrm{h}} = \text{-}2\mathrm{r}_{yx}^{-} + 2\mathrm{h}^{T}\mathrm{R}_{yy} = 0$$

4. Now we can construct the filter from the above,

$$\mathrm{h} = \mathrm{R}_{yy}^{-1}\mathrm{r}_{yx}^{-}$$

5. Apply the filter h to the noisy signal $y(n)$ using convolution to get the enhanced signal,

$$\hat{x}(n) = h(n) * y(n)$$

3.2 Frequency Domain Noise Removal Algorithm

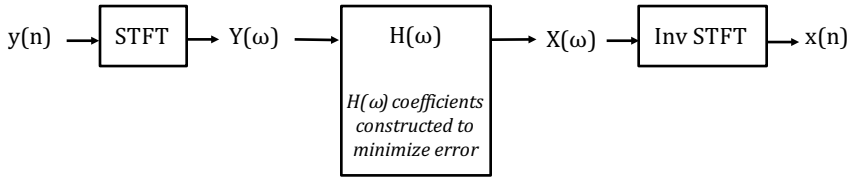

Figure 3.3: Block diagram of Wiener filtering process in the frequency domain.

Figure 3.3 shows the high level process of Wiener filtering in the frequency domain. Below we describe the steps for noise removal, including details on how to construct the LMS filter $H(\omega)$.

1. Find error of approximation at frequency ω as

$$E(\omega) = X(\omega) \text{ - } \hat{X}(\omega)$$

where $\hat{X}(\omega)$ can be replaced with $H(\omega)\,Y(\omega)$, or the result of filtering the noisy signal.

2. Find the mean squared error value, which is ultimately to be minimized,

$$J = E\left[E(\omega)^2\right]$$

which we can expand to

$$J = E\left[D(\omega)^2\right] \text{ - } H(\omega)P_{yx}(\omega) \text{ - } H^*(\omega)P_{yx}(\omega) + H(\omega)^2P_{yy}(\omega)$$

such that $P_{yy}(\omega)$ is the power of the noisy signal, $P_{yx}(\omega)$ is the cross power spectrum of the noisy and clean signals, and * indicates convolution.

3. Minimum error is reached when the derivative of J is zero, so we compute the derivative with respect to h to find the necessary filter,

$$\frac{\partial J}{\partial H(\omega)} = H(\omega)^2P_{yy}(\omega) \text{ - } P_{yx}(\omega) = 0$$

4. Now we can construct the filter from the above,

$$H(\omega) = \frac{P_{yx}(\omega)}{P_{yy}(\omega)}$$

5. Apply the filter h to the noisy signal $Y(\omega)$ to get the enhanced signal in the frequency domain,

$$\hat{X}(\omega) = H(\omega)\,Y(\omega)$$

and use the inverse Fourier transform to get back the time domain enhanced signal, $\hat{x}(n)$.

CHAPTER 4

SUBSPACE METHOD

Subspace methods take advantage of the characteristics of singular value decomposition (SVD) and eigenvalue decomposition (EVD) of matrices to remove noise from corrupted signals. The main idea behind this process is that given a noisy speech matrix, we can find two subspaces, the noisy subspace and the noise subspace. Given this information, removing noise from a signal can be accomplished by simply removing from the noisy speech matrix the values and vectors associated with the noise in the signal [4]. This method often works without the common side effects of spectral subtraction like musical noise production; however, it does assume that the noise is zero mean and uncorrelated with the speech signal we are trying to recover. The model of the problem is this: we are given a noisy signal $y(n)$, which is some clean speech $x(n)$ corrupted with noise $d(n)$,

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

In order to use subspace methods, we need to put these signals into matrix form, which can be accomplished through the formation of Toeplitz, Hankel, or cross correlation matrices,

$$Y = X + D$$

SVD and EVD analysis of the Y matrix can thus give us information about X and D that we can use to eliminate the D component from Y, and give us X [9]. Once we have this we can reconstruct a time domain signal from the matrix to get the enhanced signal that we want, $x(n)$.

4.1 SVD Based Noise Reduction

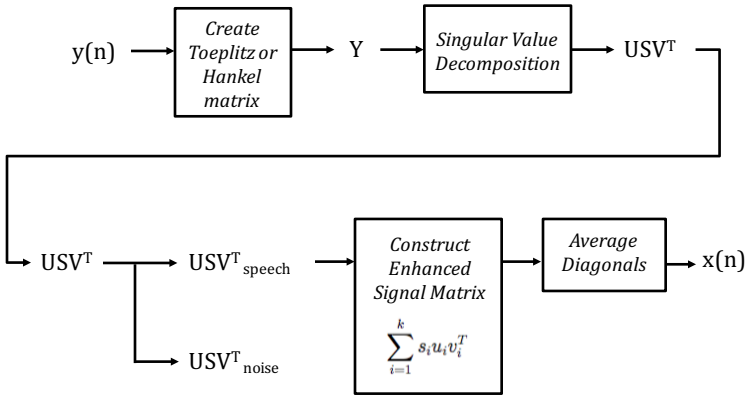

Figure 4.1: Block diagram of SVD-based subspace enhancement.

Figure 4.1 shows the high level process behind subspace enhancement. Below we detail these steps regarding speech enhancement using SVD-based subspace algorithms.

1. Separate the noisy time domain speech signal $y(n)$ into overlapping frames, and perform each of the following steps for each frame.

2. Form the Toeplitz or Hankel matrix Y.

3. Find the SVD decomposition of Y such that $Y = USV^T$. S contains singular values along the diagonal while U and V store the left and right singular vectors associated with the respective singular values.

4. Choose how many singular values to keep and how many to zero out as noise. The smallest singular values (and their associated vectors) correspond to the noise, while larger values (and associated vectors) correspond to speech. The number of singular values retained is some number k that is smaller than the actual number of singular values of Y.

5. Construct an enhanced signal matrix X by finding the low rank approximation of Y using only the largest k singular values and vectors of Y. The formula for finding this approximation is

$$X = \sum_{i=1}^{k} s_i u_i v_i^T$$

u_i and v_i^T are vectors from matrices U and V, which correspond to the largest singular values s_i we choose to use from S.

6. For every diagonal in X, find the average. These averages are the values of the cleaned signal $x(n)$.

4.2 EVD Based Noise Reduction

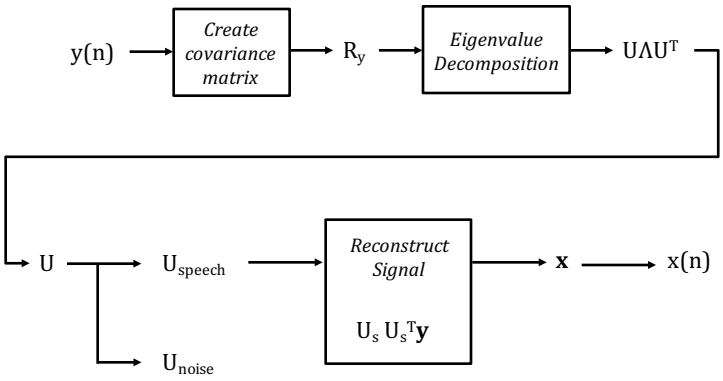

Figure 4.2: Block diagram of EVD-based subspace enhancement.

Figure 4.2 outlines the steps for noise reduction using EVD-based subspace algorithms at a high level. Below we detail the steps in this algorithm.

1. Given the noisy speech signal (in vector form), $\mathbf{y} = \mathbf{x} + \mathbf{d}$, construct the covariance matrix of $\mathbf{y}$, such that

$$R_y = R_x + R_d$$

2. Find the eigenvalue decomposition of R_y such that $R_y = U\Lambda U^T$. U is a matrix containing eigenvectors while Λ is a matrix containing eigenvalues on the diagonal.

3. The smallest eigenvalues correspond to noise while the larger ones, the principal eigenvalues, correspond to speech. Using eigenvalues corresponding to speech, construct a U_s matrix such that it contains eigenvectors relating only to the speech.

4. Reconstruct the clean signal vector by projecting $\mathbf{y}$ onto the speech subspace of the signal,

$$\mathbf{x} = U_s U_s^T \mathbf{y}$$

   This gives us the enhanced signal, $x(n)$.

CHAPTER 5

METRICS

We analyze the effectiveness of the discussed methods with a variety of metrics. The four main objective measures we will use are segmental SNR, PESQ measure, LLR measure, and WSS distance.

5.1 Segmental SNR

Signal-to-noise ratio (SNR) measures the ratio between the amount of important content and noise content in a signal. SNR can be calculated over an entire signal, but the segmental SNR is often better at providing a measure of the quality of a signal as it calculates the SNR frame by frame [10]. As segmental SNR is a ratio, a higher dB value denotes a better quality signal.

The SNR of an entire signal is calculated as follows, where $x(n)$ is the clean or enhanced signal and $d(n)$ is the noise signal at time n. The SNR calculates the ratio of the power of the signal-to-noise content. The noise signal is obtained as the difference between the clean and corrupted speech signals.

$$\text{SNR} = 10 \log_{10} \frac{\sum_{n=1}^{N} x(n)^2}{\sum_{n=1}^{N} d(n)^2}$$

where n is the current time frame and N is total number of samples. We can calculate segmental SNR by applying the above to single frames of the signal and doing some preprocessing during the process. This involves removing SNR values that may be too high or too low to indicate any change in quality. Segmental SNR is therefore calculated as follows:

$$\text{Segmental SNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Lm}^{Lm+L-1}x(n)^2}{\sum_{n=Lm}^{Lm+L-1}d(n)^2}$$

where L is the number of samples per frame and M is the number of frames in the signal.

5.2 PESQ Measure

The perceptual evaluation of speech quality (PESQ) measure is a widely used metric for judging signal intelligibility, and is often used as a standard for speech signal quality. This measure came up as a replacement for the traditional use of human listening tests to judge speech signal quality. The scores range from 1 to 5, with higher values indicating a better quality signal [11].

5.3 LLR Measure

The log likelihood ratio (LLR) is a measure of the distance of a corrupted (or enhanced) signal from the clean signal by comparing the linear predictive coding (LPC) vectors of the clean and corrupted speech [12]. This metric is calculated with a log function, so smaller values indicate signals closer to the true clean signal.

The LLR is calculated as

$$\text{LLR} = \log\left(\frac{\mathbf{a_d}\mathbf{R_c}\mathbf{a_d^T}}{\mathbf{a_c}\mathbf{R_c}\mathbf{a_c^T}}\right)$$

where a_d is the LPC vector of the corrupted signal, a_c is the LPC vector for the clean signal, and R_c is the auto correlation matrix for the clean signal.

5.4 WSS Distance

The weighted spectral slope (WSS) distance measures the difference in spectral slopes of difference frequency bands in each frame of the distorted or enhanced signal from that of the clean signal [13]. This way, the difference in actual signal intensity is given less importance, and speech quality is measured on the basis of the similarity of changes in signal intensity to that of the clean signal. A lower measure indicates a higher similarity, and thus a cleaner signal.

The WSS distance for a signal can be calculated as follows. The metric takes into account the different frequency bands present in each frame of the signal.

$$\text{WSS Distance} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m)(S_c(j,m) - S_d(j,m))^2}{\sum_{j=1}^{K} W(j,m)}$$

where K is the number of frequency bands, M is the number of frames in the signal, S_c is the spectral slope of the clean signal, S_d is the spectral slope of the corrupted signal, and W is the weight for a specific frequency band at a certain frame. The weights are calculated using characteristics of the spectra of both signals.

CHAPTER 6

RESULTS AND ANALYSIS

6.1 Noisy Signal Database

Each algorithm was used to enhance 1080 speech signals taken from the NOIZEUS database. The speech signals include short sentences spoken by both male and female speakers. There were 9 different noise types tested against, including an ideal case of white Gaussian noise (WGN) corruption. For each noise type, we had signals with 4 different amounts of noise corruption. White Gaussian noise (WGN) is the most assumed case for noise in a signal. WGN is stationary, uncorrelated, and fairly constant over the course of a signal. To create these noisy signals we added to the clean speech signals different scaled amounts of random white noise depending on what level of corruption (SNR) we wanted. Since signal processing algorithms are most effective on stationary noise, we should expect to see better performance when enhancing signals corrupted with WGN rather than colored noise.

We tested the three methods on 8 other noise types at different corruption levels. These signals were corrupted with ambient noise related to an airport, babble, car, exhibition, restaurant, station, street, and train. Depending on the characteristics of these noise types as being more varied than WGN, as well as their characteristics relative to each other, we can expect to see different results.

6.2 Spectral Subtraction

Below we explore the results of enhancing various corrupted signals with spectral subtraction. First we look at the most ideal case of stationary, zero

mean, white gaussian noise (WGN).

6.2.1 White Gaussian Noise

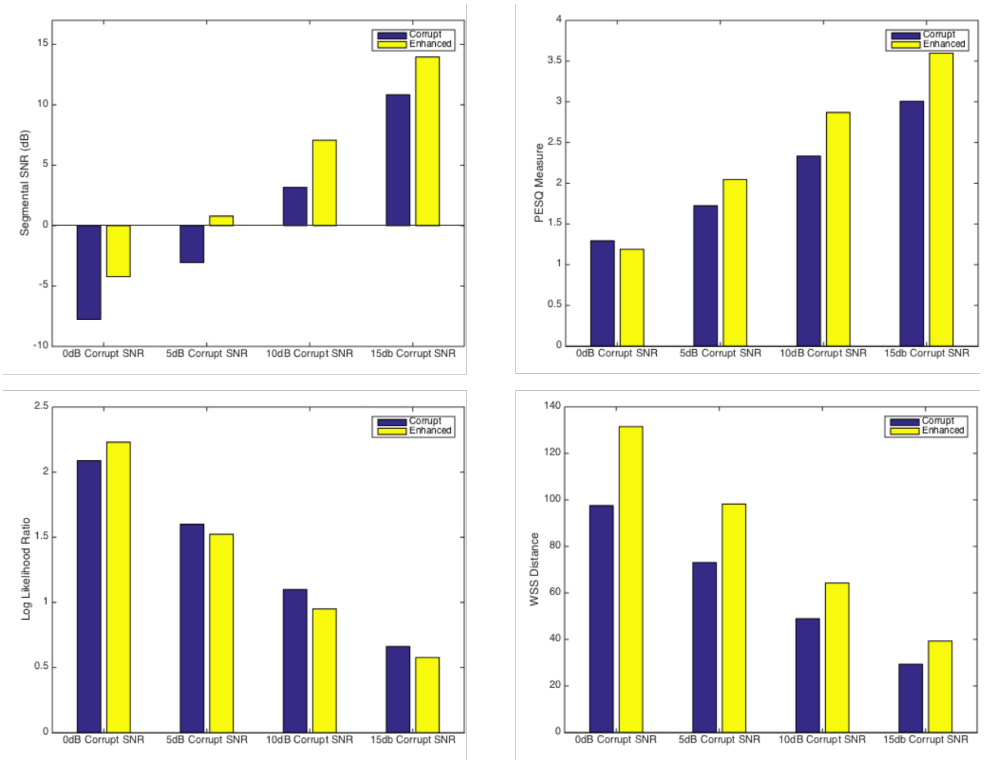

Figure 6.1: Comparison of segmental SNR (top left), PESQ measure (top right), LLR (bottom left), and WSS distance (bottom right) for WGN corrupt signals and spectral subtraction enhanced signals. In all plots, the dark purple bars refer to the corrupt signals, while the light yellow bars refer to the enhanced signals.

As seen in figure 6.1, spectral subtraction improved the segmental SNR of the signal by quite a lot. Even the case of highest corruption (0 dB SNR) resulted in an improvement in overall quality.

The PESQ is a measure of signal quality, looking more at whether or not the signal is understandable. From the PESQ plot in figure 6.1, we see that in the case of highest level of corruption (0 dB SNR), spectral subtraction does not do enough to improve the quality of the signal, mainly due to garbling and musical noise (as showed by listening tests). However, in all other cases, there is an improvement, and the amount increases as the level of corruption in the original signal decreases.

LLR is a measure of intelligibility of a signal. Looking at the LLR plot in figure 6.1, in the case of highest level of corruption (0 dB SNR), the enhanced signal is slightly less understandable, probably due to musical noise and overlapping. In the other cases, we have improvement in LLR (lower is better), but the improvement is not too great, indicating that the enhanced signal is not of very high quality.

WSS compares spectra of the noisy and enhanced signals with that of the clean signal. As discussed earlier, a smaller distance indicates closer values. As seen from the previous three plots, spectral subtraction has been removing a good amount of noise from the signal. However, from the WSS plot in figure 6.1 we see that the WSS measure indicates an increase the spectral distance. This would thus indicate the introduction of random artifacts distorting the signal spectra. While the perceived quality of the signal may be better, and the level of noise may be reduced, the modifications to the signal are clearly compromising the quality of the enhanced signal.

6.2.2 Other Noise Types

Given the results for signals corrupted with the ideal case of white Gaussian noise, it follows that the performance of spectral subtraction on various other noise types (as described in the introduction to this chapter) will be lesser in quality. In figure 6.2, we can see that segSNR and PESQ measure are improved on average, while LLR and WSS reflect a drop in signal quality.

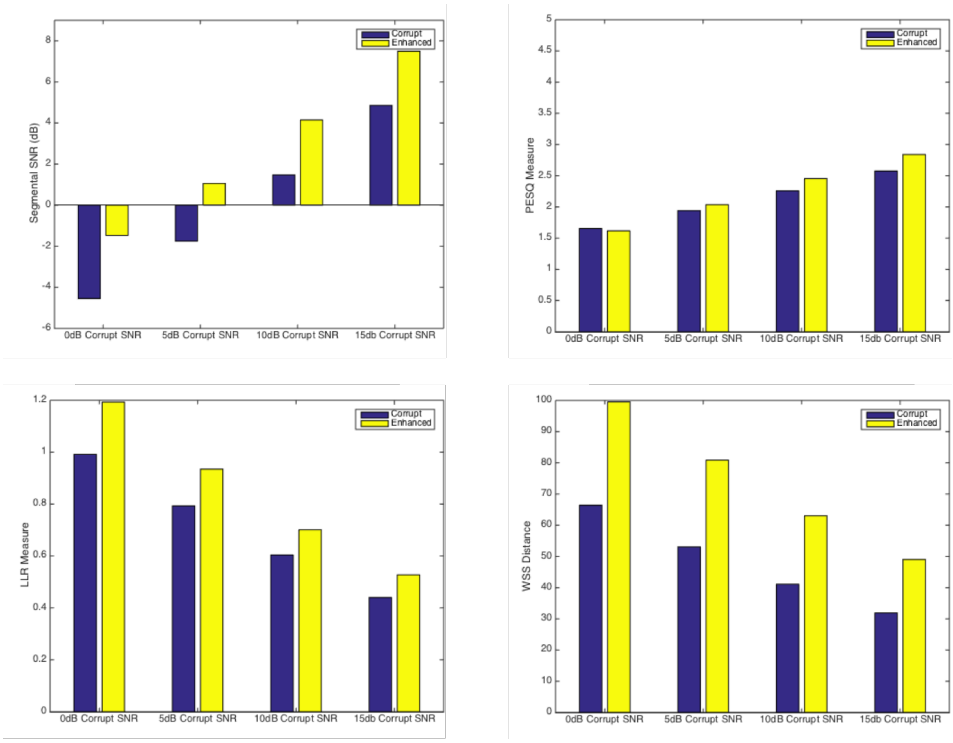

Figure 6.2: Comparison of segmental SNR (top left), PESQ measure (top right), LLR (bottom left), and WSS distance (bottom right) for corrupt signals and spectral subtraction enhanced signals. In all plots, the dark purple bars refer to the corrupt signals, while the light yellow bars refer to the enhanced signals. Results are averaged values of signals corrupted with 8 different noise types and their respective enhanced signals.

6.3 Wiener Filtering

Below we explore the results of enhancing various corrupted signals with Wiener filtering. First we look at the most ideal case of stationary, zero mean, white Gaussian noise (WGN).

6.3.1 White Gaussian Noise

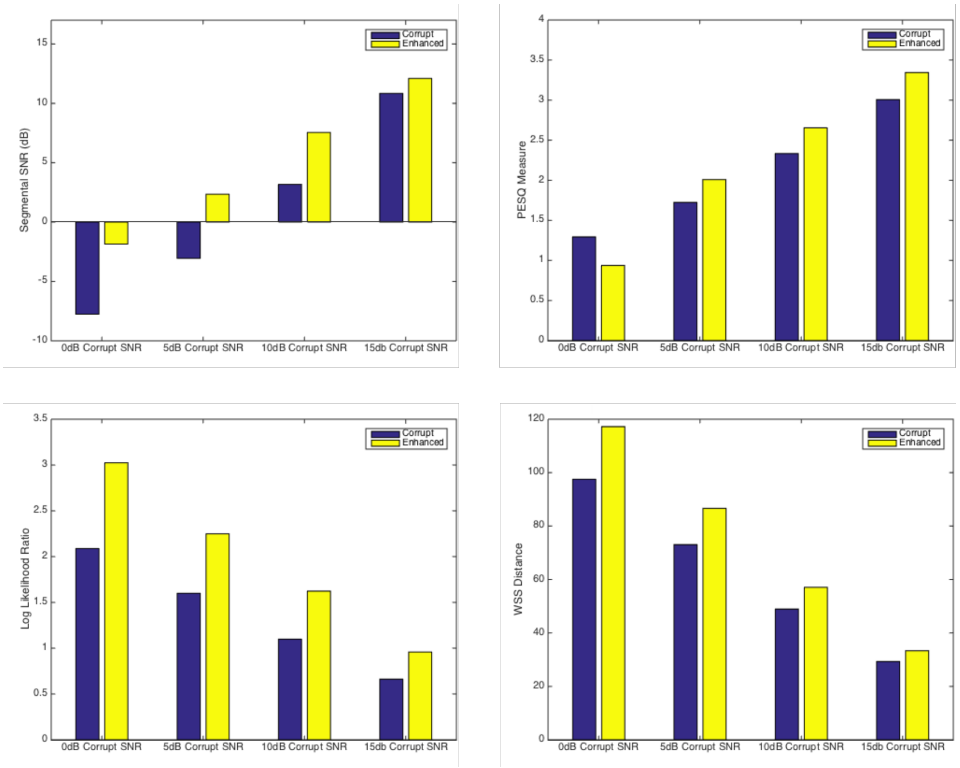

Figure 6.3: Comparison of segmental SNR (top left), PESQ measure (top right), LLR (bottom left), and WSS distance (bottom right) for WGN corrupt signals and Wiener filtered signals. In all plots, the dark purple bars refer to the corrupt signals, while the light yellow bars refer to the enhanced signals.

Figure 6.3 shows an improvement in segmental SNR for all levels of corruption when enhancing with Wiener filtering. This does not say too much about quality, but it shows that there is a decrease in the amount of random noise present in the enhanced signal from the corrupt signal.

From figure 6.3, we can see that in the cases of less corruption (5 dB to 15 dB SNR), the PESQ measure is appropriately improved, indicating that the subjective quality of the signal improves after filtering. However, the most corrupt signal's quality actually decreased, possibly due to overfiltering, and thus removal of some speech.

Further quality decrease is visible in the LLR plot in figure 6.3, showing that the LLR of the enhanced signals are actually higher than those of the corrupt signals. It would appear that the Wiener filter used to enhance the

signal fails to improve the intelligibility of the signal, possibly removing too much, or introducing some new artifacts that are skewing the LPC coefficients by a large amount. This assumption is further backed by the increase in WSS distances, as seen in the WSS plot in figure 6.3.

6.3.2 Other Noise Types

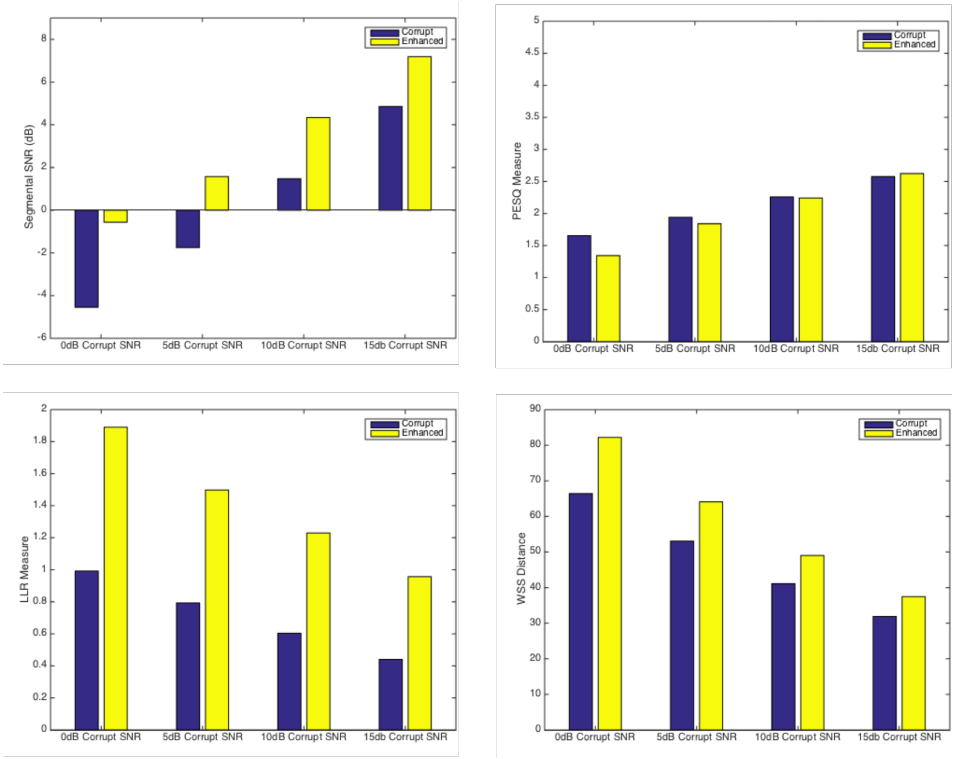

Figure 6.4: Comparison of segmental SNR (top left), PESQ measure (top right), LLR (bottom left), and WSS distance (bottom right) for corrupt signals and Wiener filtered signals. In all plots, the dark purple bars refer to the corrupt signals, while the light yellow bars refer to the enhanced signals. Results are averaged values of signals corrupted with 8 different noise types and their respective enhanced signals.

The results displayed in figure 6.4 reflect a similar trend to those for the WGN corrupted signals. The PESQ measure is more or less unchanged, though there is a more significant drop in the case of most corruption. However from the large increases in LLR and WSS, it follows that the Wiener filtering would have removed noise but introduced extra artifacts into the signal that resulted in some garbling.

6.4 Subspace Enhancement

Below we explore the results of enhancing various corrupted signals with subspace enhancement. First we look at the most ideal case of stationary, zero mean, white Gaussian noise (WGN).

6.4.1 White Gaussian Noise

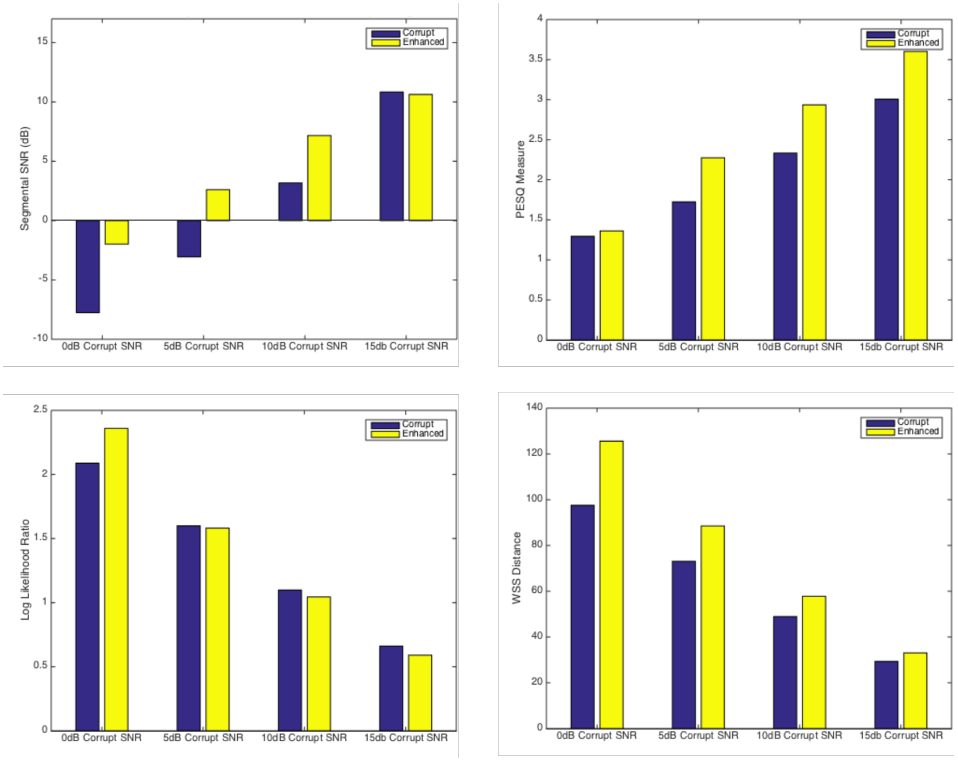

Figure 6.5: Comparison of segmental SNR (top left), PESQ measure (top right), LLR (bottom left), and WSS distance (bottom right) for WGN corrupt signals and Subspace enhanced signals. In all plots, the dark purple bars refer to the corrupt signals, while the light yellow bars refer to the enhanced signals.

We see from figure 6.5 that segmental SNR is almost always improved by subspace enhancement. There is a slight drop in the segSNR value in the case of least corruption (15 dB SNR), which could be from discarding too many singular values due to over-assuming the amount of noise that is present. We also see a trend of increasing improvements in PESQ measure as the level of corruption goes down. Clearly, the quality of the signal is improving with the application of the subspace method.

LLR is worsened in the case of highest level of corruption, as seen in the LLR plot from figure 6.5, but the LLR is reduced in all other cases, though the amount of improvement is small. We also see that WSS distance, however, does not improve, but slightly worsens. These results could reveal a shortcoming in the algorithm, demonstrating either too much removal, or too little.

6.4.2 Other Noise Types

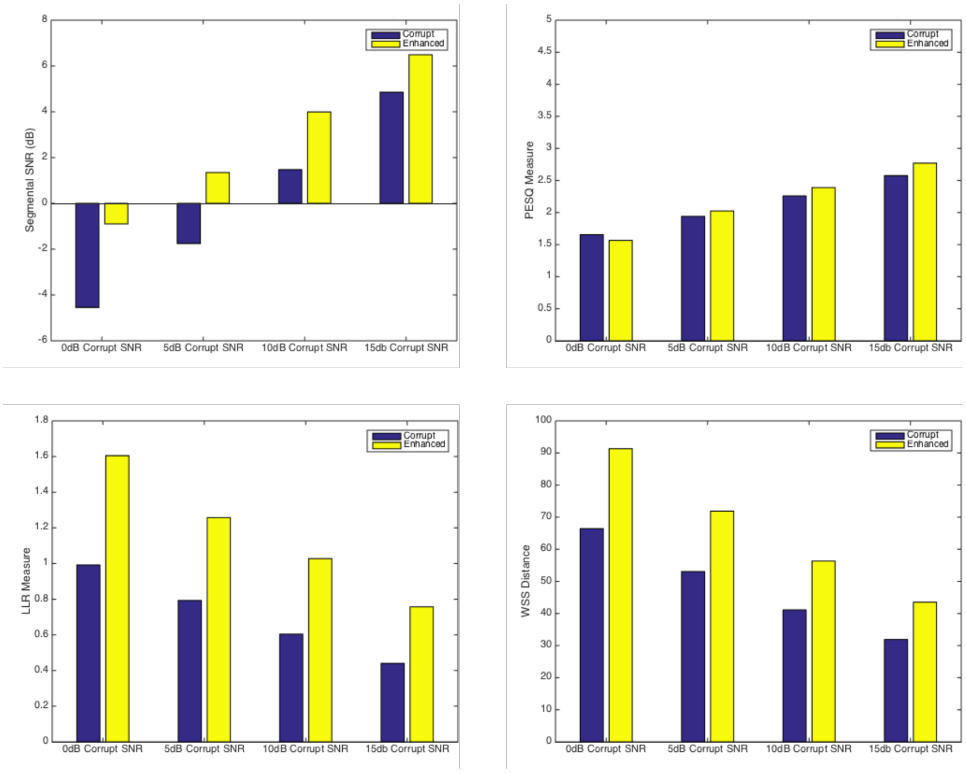

Figure 6.6: Comparison of segmental SNR (top left), PESQ measure (top right), LLR (bottom left), and WSS distance (bottom right) for corrupt signals and Subspace enhanced signals. In all plots, the dark purple bars refer to the corrupt signals, while the light yellow bars refer to the enhanced signals. Results are averaged values of signals corrupted with 8 different noise types and their respective enhanced signals.

The results displayed in figure 6.6 follow the trend seen in the enhancement of WGN corrupted signals. On average, segSNR and PESQ measure are either increased, or stay around the same, and LLR and WSS distance are worsened. This shows that the algorithm is canceling out either too much or

not enough noise, and that it is not robust enough to produce high quality results in response to non-ideal noises.

6.5 Comparison of Algorithm Performance

In this section we analyze the performance of each algorithm by comparing their abilities to improve each metric.

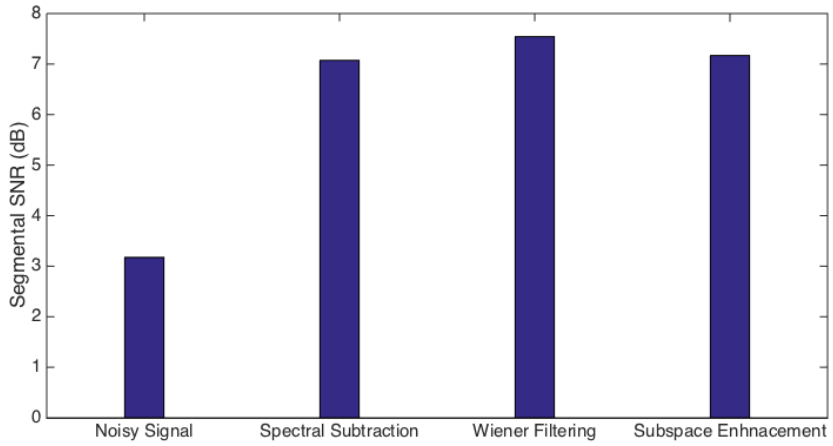

Figure 6.7: Comparison of the effect of Spectral Subtraction, Wiener Filtering, and Subspace Enhancement on segmental SNR. These results are averaged over signals corrupted at 10 dB WGN.

Segmental SNR is the best metric for checking removal of noise content in a signal. The higher the segmental SNR, the less noise there is present. Figure 6.7 shows that all three algorithms are able to almost double the segmental SNR of a signal with respect to the original corrupt signal. Evidently, the algorithms are successful at noise removal.

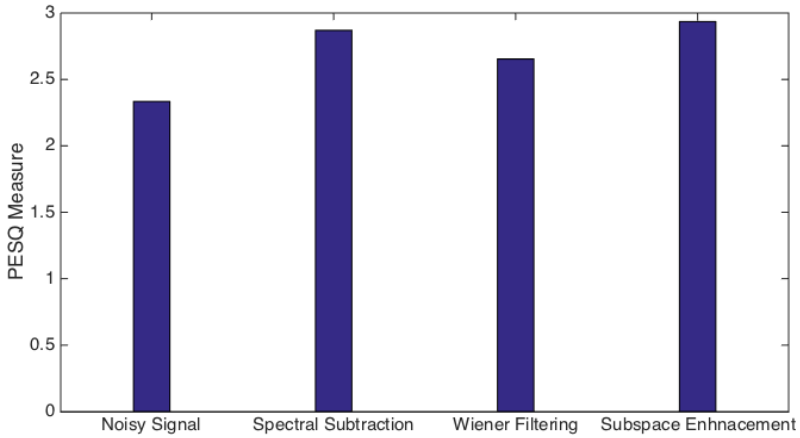

Figure 6.8: Comparison of the effect of Spectral Subtraction, Wiener Filtering, and Subspace Enhancement on segmental SNR. These results are averaged over signals corrupted at 10 dB WGN.

The PESQ measure attempts to replicate the results of human listening tests, so can be said to estimate intelligibility of a signal. Figure 6.8 summarizes the effect of the algorithms on corrupted signals, and shows a small improvement in the PESQ measure. Previous analysis also showed similar results of moderate improvement in this measure. Since we know that noise is removed as shown by figure 6.7, clearly the enhanced signals are still not qualitatively much better than the corrupted signals. This shows that there must be some different disruption introduced.

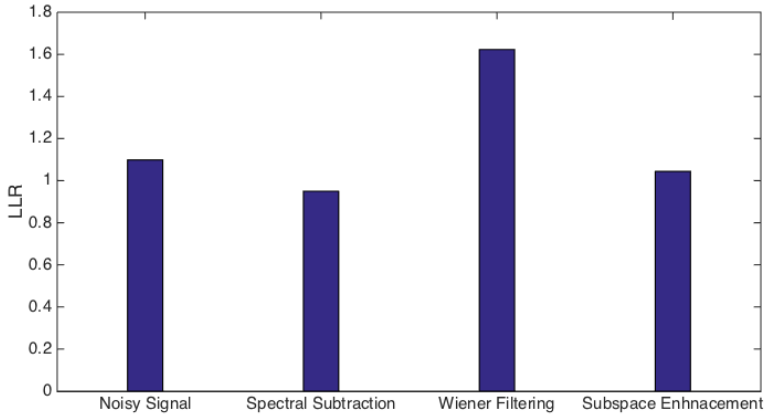

Figure 6.9: Comparison of the effect of Spectral Subtraction, Wiener Filtering, and Subspace Enhancement on segmental SNR. These results are averaged over signals corrupted at 10 dB WGN.

Figure 6.9 supports the claim that some new corruption is introduced to the enhanced signals that reduce the qualitative improvements of the algorithms. Spectral subtraction and subspace enhancement succeed in slightly reducing the LLR, showing that the intelligibility is improved by a small amount. However Wiener filtering is unable to improve this metric and actually worsens it. We saw in figure 6.7 that Wiener filtering was most successful at improving segmental SNR, showing that it is very effective at noise removal, but figures 6.8 and 6.9 that the speech signal quality is compromised, indicating possible overfiltering of the signal. With spectral subtraction and subspace enhancement however, PESQ measure and LLR are improved but not significantly. This shows that the noise removal results in the introduction of musical noise, which serves to reduce the quality of the signal despite removing the initial noise.

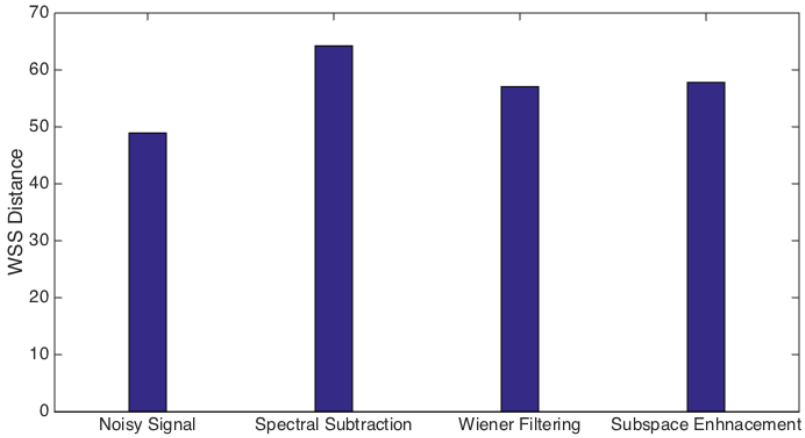

Figure 6.10: Comparison of the effect of Spectral Subtraction, Wiener Filtering, and Subspace Enhancement on segmental SNR. These results are averaged over signals corrupted at 10 dB WGN.

The last metric we discuss is WSS distance. Figure 6.10 indicates that none of the algorithms were able to minimize this distance. As explained in the previous chapter, the WSS distance measures the difference in the spectra of two signals of interest. The fact that none of the algorithms were able to minimize this distance shows that the enhanced signals' spectra are not more similar to those of the clean signals than those of the noisy signals. This supports the notion of introduced musical noise affecting the smoothness of the spectra of the enhanced signals.

Overall, we notice that the algorithms discussed in this thesis are quite effective at removing noise from a signal, but are not very successful at improving the signal quality. While numerical measures are not the best means for measuring signal intelligibility, listening tests conducted confirmed the results shown by the objective metrics. There is definitely reduction in noise, but musical noise affects the smoothness of the spectrum and the subjective quality of the signal is not completely preserved [14]. While such results may not be extremely useful in trying to improve quality of speech signals, for the purposes of feature extraction, they are quite reliable.

CHAPTER 7

CONCLUSION AND FUTURE WORK

As we discussed in the previous chapters, signal processing approaches can be quite effective at canceling out noise in a signal, but do not always remove all types of noise, and often fail when noise is correlated to the speech in the signal or is not constant and hence difficult for a static model to predict. As we saw in the previous chapter, signals with high levels of corruption (and therefore lower SNRs) were not always improved with the methods used. And even if they were improved, it was not by a great amount. For relatively lower levels of corruption, and where ambient noise was much more stationary over the course of the entire signal, noise removal improved a good amount after running enhancement algorithms on them, as indicated by the increases in PESQ measures and segmental SNR values, but not without comprising the intelligibility, as demonstrated by the LLR and WSS distance results. In these cases, listening tests also showed decreases in noise content, but not necessarily improvement in speech quality.

Spectral subtraction, Wiener filtering, and subspace enhancement have been some of the more popular algorithms used for decades to achieve noise removal, and are still used in many applications of signal processing and not just in speech and sound research. However, there is still the problem of efficiently solving the problem of removing non-ideal unwanted noise from a signal, especially in a way that does not create new artifacts that only reduce the quality of the enhancement [15]. There are also the problems of assuming that the first few frames of a signal are noise, and assuming the phase information of the noisy signals is also that of the clean signal. Speech enhancements and denoising using deep learning, however, is a more recent approach that is generally more accurate at creating very close estimates to true clean speech.

Neural networks are modeled after neural systems found in the brain. They are adaptive and robust models that make them ideal for many machine learning or big data tasks that require handling and large amounts of data. Speech enhancement tasks can be easily applied to neural networks [16]. Given a large amount of available noisy signals $y(n)$ and their associated clean signals $x(n)$, a network could be trained to identify noise in a signal and then remove it, with minimal difficulties in removing non-stationary noise. Performing enhancements on the signals used to generate the results in chapter 6 would most probably result in cleaner signals.

REFERENCES

[1] J. O. Smith, *Spectral Audio Signal Processing*. W3K Publishing, 2011.

[2] M. A. El-fattah, M. I. Dessouky, S. M. Diab, and F. E. El-samie, "Speech enhancement using an adaptive wiener filtering approach," *Progress In Electromagnetics Research*, vol. 4, pp. 167–184, 2008.

[3] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

[4] P. C. Loizou, *Speech Enhancement Theory and Practice*, 2nd ed. Boca Raton: CRC Press, Taylor & Francis Group, 2013.

[5] Y. Ghanbari, M. R. Karami-Mollaei, and B. Amelifard, "Improved multi-band spectral subtraction method for speech enhancement," *6th IASTED International Conference, Signal and Image Processing, Honolulu, Hawaii*, pp. 225–230, 2004.

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

[7] Y. H. Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, 2004.

[8] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 4409–4412, 2009.

[9] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[10] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," *Proc. Int. Conf. on Spoken Language Processing (ICSLP), Sydney, Australia*, pp. 2819–2822, 1998.

[11] K. Kondo, "Subjective quality measurement of speech," *Signals and Communication Technology*, 2012. [Online]. Available: http://www.springerlink.com/index/10.1007/978-3-642-27506-7

[12] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4389058

[13] H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*, vol. 7, pp. 1278–1281, 1982.

[14] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[15] P. C. Loizou, S. Member, and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2011.

[16] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," *Interspeech 2014, Singapore*, pp. 1–5, 2014.

[17] B. Milner and I. Almajai, "Noisy audio speech enhancement using Wiener filters derived from visual speech," *Proc. AVSP, Hilvarenbeek, Holland*, 2007.

[18] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners." *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–94, 2009.

[19] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech Enhancement. A Signal Subspace Perspective.* Waltham, MA: Elsevier Inc, 2014.

[20] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 2080–2090, 2010.

[21] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, pp. 629–632, 1996.