Master Thesis

Post-Processing Method for Single Channel Speech Enhancement Systems

Thesis Committee:
Prof.dr.ir. J. Biemond
Dr.ir. R. Heusdens
Dr. J. Erkelens
Dr.ir. Pascal Wiggers

Author	**Vasileios Stasinopoulos**
Email	beilfft@gmail.com
Student number	1399462
Thesis supervisor	Dr.ir. R. Heusdens
Date	September 24, 2009 - 11:00

Information and Communication Theory Group [I,C]^T

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Preface

This thesis is the final work I have carried out as a master student at the Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, at Delft University of Technology. I want to thank especially my thesis supervisor, Richard Heusdens, for being a very good teacher with inexhaustible sense of humor, especially when the things were going wrong. I also want to thank Jan Erkelens, who advised me on several subjects within this thesis. A lot of ideas are inspired by the conversations we had. I am grateful to all the rest of the guys from the audio group. Cees, Christian, Jorge and Richard Hendricks have contributed a lot to my work with their unique experience. Finally, a lot of thanks go to my family and to my friends.

Vasileios Stasinopoulos
Delft, the Netherlands
September 18, 2009

Contents

List of Figures

Chapter 1

Introduction

Speech is one of the fundamental means of human communication. Beginning with limited distance fixed-line telephone networks, recent developments have made feasible high quality mobile communication across the globe. Speech processing devices like cellular phones, hands-free equipment and every kind of human-to-machine speech processing systems are an integral part of everyday life.

While the freedom and flexibility provided by mobile technology has made it possible to communicate outside controlled environments, it has also introduced new challenges. Mobile users communicate in different environments with varying levels and types of background noise such as traffic noise, car engine noise, babble noise as in cafeterias, machine noise as in a factory, etc. Suppression of the background noise is a relevant and challenging problem. Apart from reducing listener fatigue and improving the quality and intelligibility of the speech, noise reduction is also crucial to obtain good performance of the speech coding algorithms used in mobile communications. Moreover, environmental noise has remained a limiting factor in the widespread deployment of speech enabled services, such as speech recognition and speech identification systems. Although these technologies have been originally developed to work in noise free environment, there has been an increasing effort to make them efficient under noisy conditions as well. Noise reduction is becoming also an increasingly important feature in digital hearing aids.

For all the above mentioned reasons, much effort has been devoted over the last few decades towards developing efficient speech enhancement algorithms. The term speech enhancement in fact refers to a large group of methods that all aim at improving the quality of speech signals. Some examples are bandwidth extension of narrowband signals, packet loss concealment, noise reduction, etc. In this thesis we use the term "speech enhancement" to describe additive noise reduction.

Noise reduction can be viewed as an estimation problem, where an unknown signal (speech) is to be estimated in the presence of noise, where only the noisy observation is available. The vast family of speech enhancement algorithms may be broadly classified into two categories: single and multi-channel enhancement. Single-channel methods operate on the input obtained from only one microphone. They have been attractive due to low cost and size factors, especially in mobile communications. Multi-channel enhancement on the other hand uses two or more microphones to record the noisy signal and can as such exploit also spatial information.

In this thesis we focus on single-channel speech enhancement. Single-channel enhancement

systems achieve noise reduction by exploiting the spectral diversity between the speech and noise signals and the high degree of the nonstationarity of the speech signal. Consequently, it is natural to perform enhancement in the frequency domain. Since the frequency spectra of speech and noise often overlap, speech enhancement systems generally achieve noise reduction at the expense of speech distortion. In the next section we discuss a general frequency domain based single-channel speech enhancement scheme.

1.1 Single-Channel Speech Enhancement

Processing is done in short segments of the signal (frame-by-frame basis), typically of the order of 10-30 ms, to ensure that the speech signal satisfies assumptions of wide-sense stationarity (WSS) [1]. The segmentation is done using a sliding window of finite support. The applied window is called analysis window and is often a Hann or Hamming window. The frames have a length of K samples and a frame shift of P samples. Typical values at a sampling frequency of 8 kHz are $K = 256$ $(32ms)$ and $P = 128$ (50% overlap between consecutive frames). The windowed signal is transformed to the frequency domain by applying a discrete Fourier transform (DFT). We consider the DFT coefficients of a signal as zero-mean complex random variables.

We now introduce some notation and terminology. We refer to the DFT coefficients $X(k,i)$ as the complex spectrum of a signal and to $|X(k,i)|$ as the magnitude or amplitude spectrum. Let k be the frequency-bin index and i the time-frame index. The variance of the signal's DFT coefficients is denoted as $\sigma_{XX}^2(k,i) = E\{|X(k,i)|^2\}$. The quantity $\frac{1}{K}|X(k,i)|^2$ denotes the periodogram. The periodogram is an asymptotically unbiased estimate of the power spectral density (PSD). That is $P_{XX}(k,i) = \frac{1}{K}E\{|X(k,i)|^2\}$ while the frame length K approaches infinity. Note that the PSD and the autocorrelation function of a signal form a Fourier transform pair. We consider an additive noise model

$$y_i(n) = x_i(n) + w_i(n) \tag{1.1}$$

where $y_i(n), x_i(n)$ and $w_i(n)$ represent the noisy speech, the clean speech and the noise respectively of the time-frame i. Let n be the discrete-time index and $\mathbf{x}_i = [x_i(0)x_i(1)...x_i(K-1)]^T$ denotes a segment of length K of the clean speech signal. $\mathbf{y}_i$ and $\mathbf{w}_i$ are defined analogously. Due to the linearity of the Fourier transform the additive model can be expressed in the frequency domain as

$$Y(k,i) = X(k,i) + W(k,i) \tag{1.2}$$

where $Y(k,i), X(k,i)$ and $W(k,i)$ are DFT coefficients obtained at frequency index k and in time-frame i from the noisy speech, clean speech and noise process, respectively. We assume that X and W are independent. As a consequence they are also uncorrelated, i.e.,

$$E\{X(k,i)W(k,i)\} = 0, \forall k,i \tag{1.3}$$

and the following relations holds between the corresponding PSDs:

$$P_{YY}(k,i) = P_{XX}(k,i) + P_{WW}(k,i) \tag{1.4}$$

[1] A wide-sense stationarity process $X(t)$ is a week form of stationary process in which the 1st and the 2nd moments don't vary with respect to time. In other words the mean function is constant $(E\{X(t)\} = m_X(t) = m_X(t+\tau), \forall \tau, t \in \mathbf{R})$ and the correlation function depends only on the difference between two time instances $(E\{X(t_1)X(t_2)\} = R_X(t_1,t_2) = R_X(t_1+\tau, t_2+\tau) = R_X(t_1-t_2,0), \forall \tau, t_1, t_2 \in \mathbf{R})$.

In the noise reduction problem we wish to obtain an estimate $\hat{x}(k,i)$ of the clean speech from the noisy observation $y(k,i)$ [2]. It turns out that, in general, the estimate $\hat{x}(k,i)$ is a function of the noise PSD, the clean speech PSD and the noisy observation, that is

$$\hat{x}(k,i) = g(P_{WW}(k,i), P_{XX}(k,i), y(k,i)) \tag{1.5}$$

An alternative notation that is usually used for the above equation is in terms of the *a-priori* signal-to-noise ratio (SNR) $\xi(k,i)$ and the *a-posteriori* SNR $\gamma(k,i)$, that is

$$\hat{x}(k,i) = g(\xi(k,i), \gamma(k,i), y(k,i)) \tag{1.6}$$

where

$$\xi(k,i) = \frac{P_{XX}(k,i)}{P_{WW}(k,i)} \tag{1.7}$$

and

$$\gamma(k,i) = \frac{|y(k,i)|^2}{P_{WW}(k,i)} \tag{1.8}$$

respectively.

The *a-posteriori* SNR $\gamma(k,i)$ is dependent on the noisy magnitude realization and the noise PSD. The noisy magnitude realization is known while the noise PSD is an expected value which is unknown and has to be estimated. The estimation of the noise statistics is a challenging task as the estimates has to be estimated from the noisy complex spectrum realization. A common method is to use voice activity detectors (VAD) to identify time segments where speech is absent and thus the signal consist only of the background noise [22]. While this method has the advantage of low computational complexity, its performance degrades in low SNR values and in environments with nonstationary noise. One of the most popular noise estimation schemes that adapts also during speech activity is the so called minimum statistics approach ([31], Appendix B). The minimum statistic approach tracks the minimum power level in a particular frequency bin seen across a sufficiently long time interval and compute the noise PSD from the minimum. Other recent advancements for noise PSD estimation comprise data-driven noise power estimation [15], DFT domain subspace decompositions [18, 21] and low complexity noise PSD tracking using high resolution PSDs [20]. The last advancements show excellent noise tracking capabilities for a variety of non-stationary noise sources.

On the other hand, the *a-priori* SNR $\xi(k,i)$ is completely defined in terms of expected values, which means that in practice besides the PSD of the noise, also the PSD of the speech has to be estimated. According to Eq. 1.4 the PSD of the speech can be estimated by subtracting an estimate of the noise PSD from the noisy PSD [2]. Since the latter is unknown as well, it is often estimated by the noisy periodogram. Often, the estimated speech PSD shows variation due to random fluctuations of the noisy realization. Since these variations can lead to perceptually annoying artifacts, other methods have been proposed that lead to smoother speech PSD estimates over time. The decision-directed approach discussed in [10] is one of the most commonly used techniques for this purpose. It makes use of the clean speech magnitude of the previous frame in combination with the noise PSD and the periodogram of the noisy PSD in order to obtain smooth estimates of the PSD (see section 2.3.1 on page 11 for details).

[2] We use upper case letters to denote random variables and the corresponding lower case letters to denote their realizations

To sum up, a block diagram of a generic frequency domain single-channel speech enhancement scheme is shown in Fig.1.1. An estimate of the noise PSD is obtained from the noisy speech. At this step any available prior knowledge of the noise signal can be exploited. This is done, for instance, by ascribing a particular form to the probability density function (pdf) of the noise complex spectrum, e.g. Gaussian priors. A more accurate method, though computational more demanding, is to use more sophisticated statistical models, e.g. Hidden Markov Models (HMM) [9] or codebooks [37] which have been trained on representative databases. Using the noise estimate, an estimate of the complex spectrum of clean speech is obtained. Again, any prior knowledge about the speech signal and about the human auditory system can be exploited. In some cases, the speech and the noise PSD are jointly estimated [39]. The enhanced speech $\hat{\mathbf{x}}_i$ is reconstructed in the time domain through the inverse discrete Fourier transform (IDFT). Possibly $\hat{\mathbf{x}}_i$ is windowed again, using a so called synthesis window and the estimated clean speech is reconstructed using an overlap-add technique. Often, the analysis and synthesis window are chosen such that when no processing is done in the frequency domain, a perfect reconstruction of the input signal is given at the output.

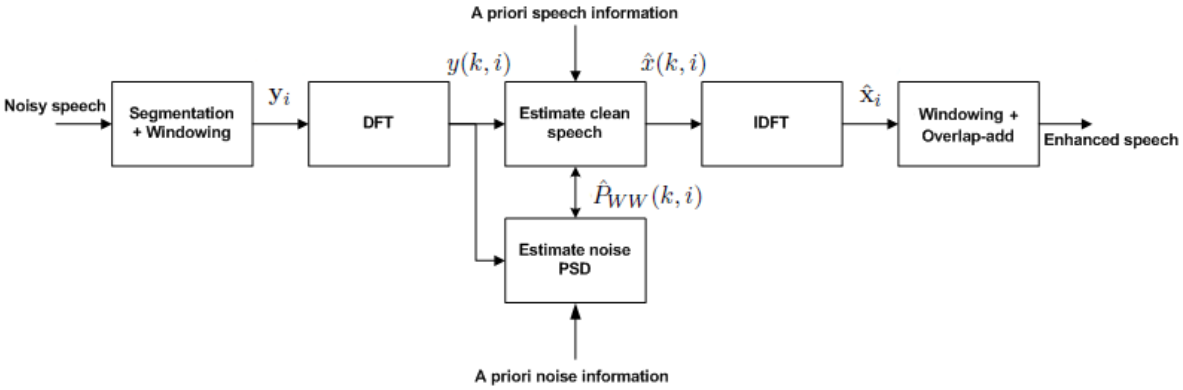

Figure 1.1: Block diagram of frequency domain single-channel speech enhancement.

Several different single-channel speech enhancement techniques have been developed over the last decades. In chapter 2 we provide a brief overview of the most important of them.

1.2 Problem definition

Generally, speech enhancement methods suffer from a number of deficiencies. These are errors in the estimates of speech and noise parameters (e.g. noise PSD estimation, *a priori SNR* estimation etc.), inaccuracies in some assumptions such as stationarity of the signals and the mismatch of the models to data (e.g. the probability distribution models to the actual distributions of speech or noise). In fact speech enhancement methods result in some loss of speech information, the severity of which depends on the SNR and the speech enhancement scheme. Furthermore, some level of noise remains in the output spectrum (*residual* noise).

Fig.1.2 presents the spectrograms of a speech signal. Comparing the clean (left) and noisy (middle) spectrograms, it is evident that parts of the speech spectrum structure have been totally masked by noise. The right spectrogram shows the enhanced signal. The method used for noise reduction is the Wiener estimator (section 2.2). As it can be seen this method

successfully suppresses the background noise. However, comparing the left and right spectrograms, it is observed that parts of the spectrum structure of speech are lost.

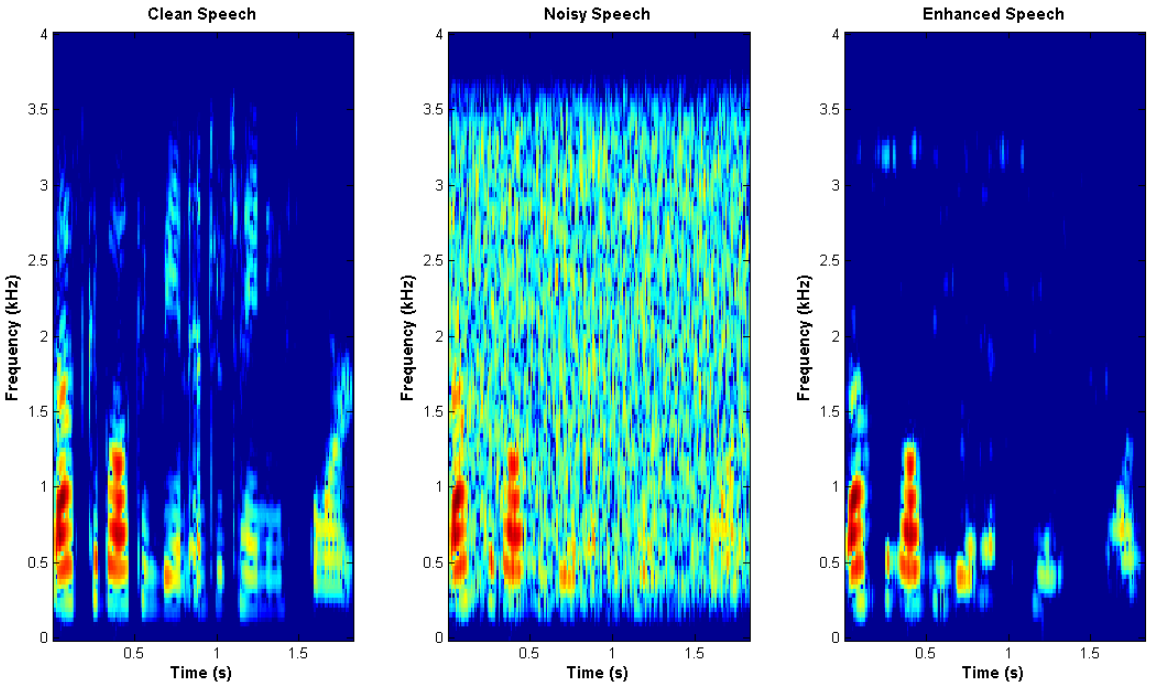

Figure 1.2: Spectrograms of clean speech, noisy speech (white Gaussian noise at 0db global SNR) and enhanced (Wiener filter - Section 2.2).

In this thesis we focus on enhancing the performance of a conventional speech enhancement system by applying post-processing restoration modules. In other words, we try to restore the parts of the speech spectrum that were lost due to noise or removed by the noise reduction algorithm. We choose to model the speech production process with linear prediction analysis (LPA - Appendix A), since LPA is widely used in many relevant speech processing applications, such as speech enhancement, speech coding and bandwidth extension. By doing so, speech is completely modeled by an excitation signal and a spectral envelope (modeling the filtering of the excitation signal by the vocal tract). This yields to a two step problem

- enhancement of the spectral envelope obtained after conducting LPA to the output signal of a conventional speech enhancement method (from now on, we call it "enhanced envelope")
- enhancement of the excitation signal obtained after conducting LPA to the output signal of a conventional speech enhancement method (from now on, we call it "enhanced excitation")

For enhancing the speech spectral envelope we use a priori knowledge of speech stored in an line spectrum frequency (LSF - Appendix A) codebook, pre-trained on clean speech. For enhancing the speech excitation we consider two different approaches. The first one, based on [33] (section 3.2), uses nonlinearity to preserve speech excitation harmonics that have been degraded by the conventional speech enhancement method, while the second one, inspired

by [44] (section 3.1), uses a priori knowledge of speech stored in codebooks, in order to restore the lost or suppressed excitation harmonics. The idea behind considering harmonic based excitation restoration methods is that in most of the spoken languages voiced sounds (where the harmonic structure is dominant) represent a large amount of the pronounced sounds. Fig 1.3 shows the block diagram of our post-processing speech enhancement approach.

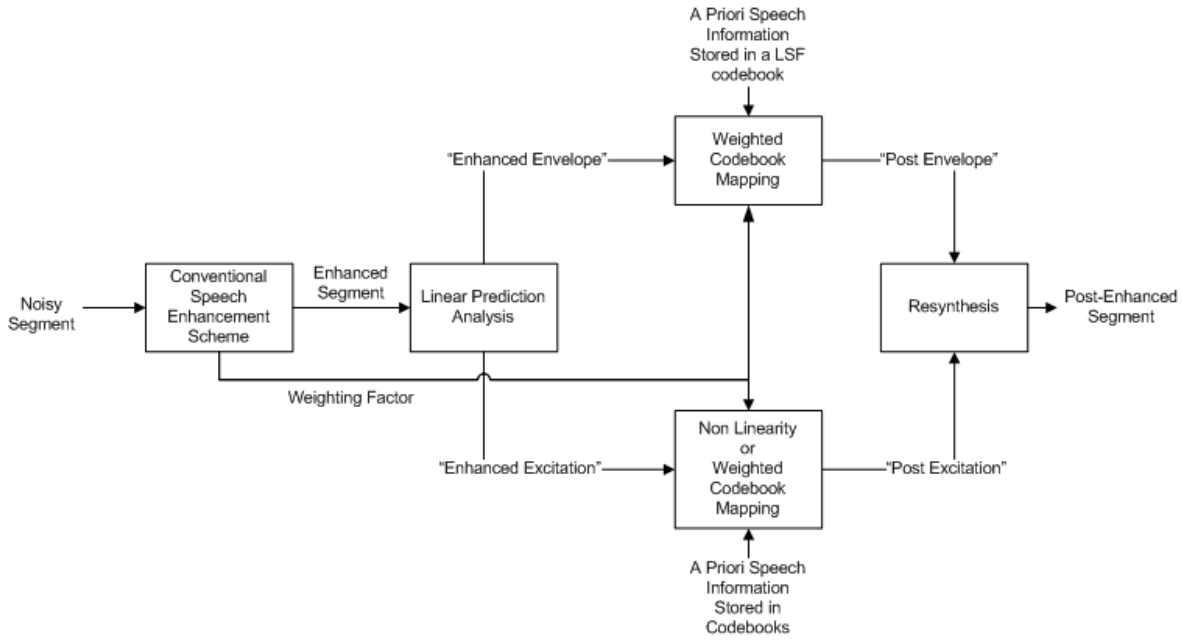

Figure 1.3: Block diagram of our post-processing speech enhancement system

The remainder of this thesis is organized as follows. In chapter 2 we give a brief overview of existing speech enhancement methods. Two resent methods that try to enhance the performance of a conventional speech enhancement method are discussed in chapter 3. In chapter 4 we present a new post-processing speech spectral envelope restoration approach for enhancing the speech spectral envelope and in chapter 5 we discuss two different approaches for enhancing the excitation signal. Experiments and results are described in chapter 6 and the thesis is concluded in chapter 7.

Chapter 2

Background

In this chapter we give a brief overview of existing speech enhancement methods. We do not provide here a complete historical overview, but we discuss the methods most relevant to the work presented in this thesis. For a comprehensive and up-to-date coverage of all major speech enhancement algorithms proposed in the last few decades see [27].

2.1 Spectral Subtraction

One of the first methods used for noise reduction of noisy speech signals was spectral subtraction [2]. It is based on a direct estimation of the short-time spectral magnitude of clean speech while maintaining the phase spectrum of the noisy signal [42]. This can be written as

$$\hat{X}(k,i) = max(|Y(k,i)| - \overline{|W(k,i)|}, 0)\frac{Y(k,i)}{|Y(k,i)|} \tag{2.1}$$

where $\overline{|W(k,i)|}$ is an estimate of the average magnitude spectrum of the noise signal. Since the magnitude spectrum cannot be negative, the negative values resulting from the subtraction are set to zero. Assuming $|Y(k,i)| \geq \overline{|W(k,i)|}$ the spectral error resulting from the above estimator is given by

$$\epsilon(k,i) = \hat{X}(k,i) - X(k,i) = W(k,i) - \overline{|W(k,i)|}\frac{Y(k,i)}{|Y(k,i)|} \tag{2.2}$$

Since the above spectral error equals the difference between the noise spectrum and its average estimate, local average of the spectral magnitudes can be used to reduce the error variance [2]. Thus replacing $|Y(k,i)|$ in Eq.2.1 with a time-averaged magnitude spectrum $\overline{|Y(k,i)|}$ gives

$$\hat{X}(k,i) = max(\overline{|Y(k,i)|} - \overline{|W(k,i)|}, 0)\frac{Y(k,i)}{|Y(k,i)|} \tag{2.3}$$

The obvious problem with this modification is that the speech is nonstationary and therefore only limited time averaging is allowed.

The aforementioned spectral subtraction algorithm is called amplitude spectral subtraction. In another variant of the spectral subtraction scheme, called power spectral subtraction, an estimate of the periodogram of the clean speech signal is obtained as

$$|\hat{X}(k,i)|^2 = max(\overline{|Y(k,i)|^2} - P_{WW}(k,i), 0) \tag{2.4}$$

which results in the following estimate of the clean speech:

$$\hat{X}(k,i) = \sqrt{max(\overline{|Y(k,i)|^2} - P_{WW}(k,i), 0)} \frac{Y(k,i)}{|Y(k,i)|} \tag{2.5}$$

A rather general formulation of the spectral subtraction estimators is given by

$$\hat{X}(k,i) = (\max(1 - b\frac{P_{WW}(k,i)^a}{|Y(k,i)|^a}, 0)^{\frac{1}{a}} Y(k,i) \tag{2.6}$$

The parameter b determines the amount of subtraction, i.e. $b > 1$ leads to an over subtraction and thus an aggressive noise reduction, while $b < 1$ leads to an under subtraction of the noise and will lead to a higher noise floor. Parameter a determines the type of subtraction, e.g. $a = 1$ and $a = 2$ for which we obtain amplitude spectral subtraction and power spectral subtraction respectively.

One of the main drawbacks of the spectral subtraction scheme is that the enhanced signal suffers from musical noise, which is especially audible in speech pauses. Random fluctuations in the periodogram result in randomly spaced narrow bands of magnitude spikes in the enhanced spectrum. In between these peaks, the spectral values are strongly attenuated since they are close or below the estimated noise spectrum ($\overline{|W(k,i)|}$) or the noise PSD, $P_{WW}(k,i)$. Transformed back in time domain this residual noise sounds like the sum of tone generators with random fundamental frequencies and is hence referred as *musical noise*. The musical noise phenomenon is common to many frequency domain speech enhancement algorithms, e.g., the Wiener filter (section 2.2) also suffers from this problem, especially when the estimated speech PSD is obtained in a subtractive fashion (Eq. 1.4)

2.2 The Wiener Filter

Another well known estimator that has been applied for noise reduction in noisy speech signals is the Wiener filter [41]. Taking into account the independence of speech and noise signals and under the assumption of large frame size K, the Wiener filter can be written in frequency domain as

$$H(k,i) = \frac{P_{XX}(k,i)}{P_{YY}(k,i)} = \frac{P_{XX}(k,i)}{P_{XX}(k,i) + P_{WW}(k,i)} \tag{2.7}$$

However, $P_{XX}(k,i)$ is not known, and in practice an estimate $\hat{P}_{XX}(k)$ of $P_{XX}(k)$ is used. This estimate is sometimes obtained in a subtractive fashion from Eq. 1.4 using an estimate $\hat{P}_{YY}(k,i)$ of $P_{YY}(k,i)$ and an estimate $\hat{P}_{WW}(k,i)$ of $P_{WW}(k,i)$. The negative values are set to zero since the PSD cannot be negative:

$$\hat{P}_{XX}(k,i) = max(\hat{P}_{YY}(k,i) - \hat{P}_{WW}(k,i), 0) \tag{2.8}$$

so that the clean speech spectrum is estimated according to

$$\hat{X}(k,i) = \frac{max(\hat{P}_{YY}(k,i) - \hat{P}_{WW}(k,i), 0)}{\hat{P}_{YY}(k,i)} Y(k,i) \tag{2.9}$$

In practice the $\hat{P}_{YY}(k,i)$ may be estimated through the periodogram or a smooth version thereof and $\hat{P}_{WW}(k)$ using noise tracking methods.

Dividing the numerator and the denominator of Eq. 2.7 by the noise PSD $P_{WW}(k,i)$ it yields

$$H(k,i) = \frac{\xi(k,i)}{\xi(k,i)+1} \tag{2.10}$$

From Eq. 2.10, the following interpretation of the Wiener filter frequency response $H(k,i)$ in terms of the *a priori* SNR can be deduced. Consider the two limiting cases of a noise-free signal $\xi(k,i) = \infty$ and an extremely noisy signal $\xi(k,i) = 0$. At very high SNR, $H(k,i) \approx 1$, and the filter applies little or no attenuation to the noise-free frequency bin k. At the other extreme, when $\xi(k,i) = 0$, $H(k,i) = 0$ and the filter completely suppresses these frequency-bins. Therefore, for additive noise, the Wiener filter attenuates each frequency component in proportion to an estimate of the *a priori* SNR. Fig.2.1 shows an example of the Wiener estimator for a voiced segment of speech degraded by car noise at 15db overall SNR. Since we are using the modified decision directed approach (Eq. 2.18) to estimate the *a priori* SNR, the Wiener gain function is bounded to values between 0.01 and 1. Notice that in the figure the Wiener gain function is scaled to make it clearer.

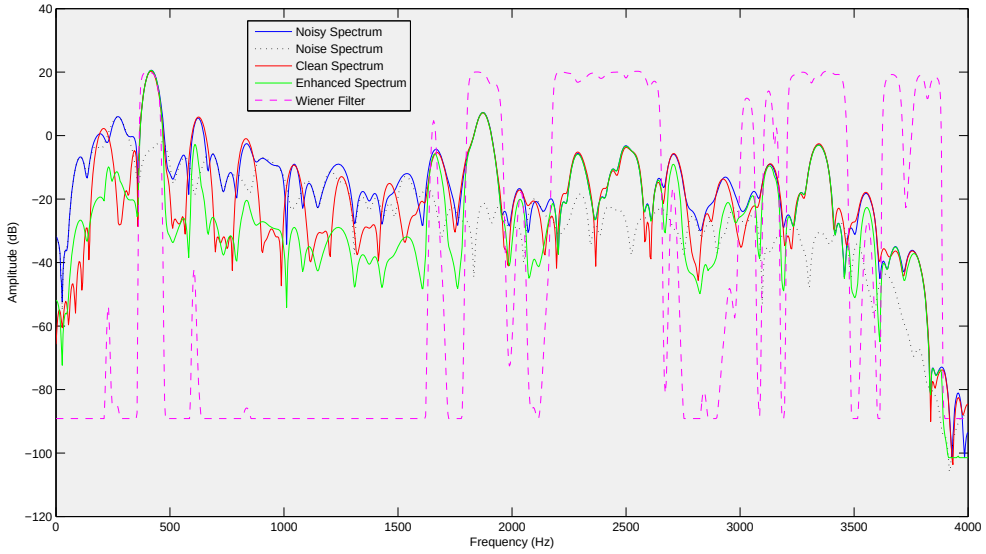

Figure 2.1: Wiener estimator for a voiced segment of speech degraded by car noise at 15db overall SNR. The noise level was estimated by the minimum statistics approach (Appendix B) and the a priori SNR by the decision directed approach (section 2.3.1).

Among the linear estimators, the wiener filter is the optimal one in terms of mean-square error (MSE). When the clean speech and the noise DFT coefficients are both complex Gaussian distributed the Wiener filter is also the optimal nonlinear estimator.

2.3 Statistical Model Based Systems

2.3.1 Gaussian models

In [10] an minimum mean-square-error (MMSE) magnitude estimator was proposed under the same statistical model as for the Wiener filter, i.e. both the speech and the noise DFT coefficients were assumed Gaussian distributed. This implies that the clean speech magnitude follows a Rayleigh distribution. The reason to consider a magnitude estimator instead of a complex was based on the argument that the phase of speech is perceptually less relevant than the magnitude [2, 42]. The MMSE short-time spectral amplitude (STSA) estimate is obtained by applying the following gain function to the noisy spectral magnitude:

$$H_{STSA}(k,i) = \frac{\sqrt{\pi u(k,i)}}{2\gamma(k,i)} \exp(-\frac{u(k,i)}{2})[(1+u(k,i))I_0(\frac{u(k,i)}{2}) + u(k,i)I_1(\frac{u(k,i)}{2})] \quad (2.11)$$

where $I_0(.)$ and $I_1(.)$ are the modified Bessel functions of order zero and one respectively and $u(k,i) = \frac{\xi(k,i)}{\xi(k,i)+1}\gamma(k,i)$.

Motivated by the observation that the MSE of the log-spectral amplitude (LSA) is subjectively a more meaningful distortion measure than the MSE of the spectral amplitude, in [11], using the same statistical model as in [10], an MMSE LSA is derived:

$$H_{LSA}(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)} \exp(\frac{1}{2}\int_{u(k,i)}^{\infty} \frac{e^{-t}}{t} dt) \quad (2.12)$$

This approach was found to result in lower residual noise than when the MSE was minimized in the spectral domain, which can be explained by the higher suppression provided by the LSA scheme. In Fig. 2.2 suppression curves for different values of $\xi(k,i)$, are plotted as a function of the *instantaneous* SNR $\frac{|y(k,i)|^2 - P_{WW}(k,i)}{P_{WW}(k,i)} = \gamma(k,i) - 1$, for the Wiener filter, the MMSE STSA and the MMSE LSA. Note that the Wiener gain function does not depend on the *a posteriori* SNR.

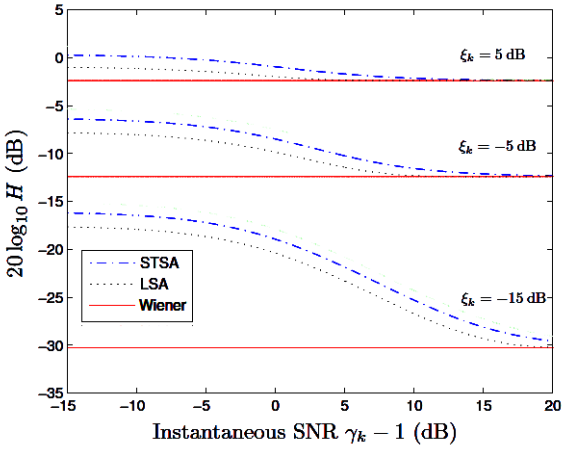

Figure 2.2: Suppression curves, source [38].

Estimation of the *A priori* SNR

Most of the frequency domain speech enhancement schemes are defined in terms of the *a priori* SNR. In practice $\xi(k,i)$ is unknown and has to be estimated for each frequency bin and for each frame index from the noisy observation.

The decision directed approach discussed in [10] is one of the most commonly used technique for *a priori* SNR estimation. It is based on the definition of $\xi(k,i)$ and its relationship with the *a posteriori* SNR $\gamma(k,i)$, as given below:

$$\xi(k,i) = \frac{E\{|X(k,i)|^2\}}{P_{WW}(k,i)} \tag{2.13}$$

and

$$\begin{aligned}\xi(k,i) &= \frac{E\{|Y(k,i)|^2\} - E\{|W(k,i)|^2\}}{P_{WW}(k,i)} \\ &= E\{\gamma(k,i) - 1\}\end{aligned} \tag{2.14}$$

Combining linearly the above two expression for $\xi(k,i)$ we get:

$$\xi(k,i) = E\{\alpha\frac{|X(k,i)|^2}{P_{WW}(k,i)} + (1-\alpha)[\gamma(k,i) - 1, 0]\} \tag{2.15}$$

with $0 \leq \alpha \leq 1$. In practice the estimator $\hat{\xi}(k,i)$ is deduced from Eq. 2.15, and is given by

$$\hat{\xi}(k,i) = \alpha\frac{|\hat{x}(k,i-1)|^2}{P_{WW}(k,i)} + (1-\alpha)max[\gamma(k,i) - 1, 0] \tag{2.16}$$

A typical value for α is 0.98. By comparing Eq. 2.15 and Eq. 2.16 we see that $\hat{\xi}(k,i)$ is obtained by dropping the expectation operator, using the amplitude estimator of the clean speech of the previous frame $\hat{x}(k,i-1)$ and using the *max* operator to ensure the positiveness in case $\xi(k,i) - 1$ is negative. The parameter α determines how smooth the estimate will be and is therefore called smoothing factor. The closer α is to one, the smoother across time the estimator will be. Since the attenuation of the noisy spectral amplitude, e.g. Eq.2.11 and Eq.2.12, or of the complex spectrum, e.g. Eq.2.10, depends on the *a priori* SNR estimation, its smooth behavior eliminates large variations across successive frames, resulting in reduced musical noise [6]. In return for this decrease in variance, the price to be payed is a delay in the estimation. In other words, α controls the trade-off between the degree of smoothing of $\xi(k,i)$ during speech-absent frames and the level of transient distortion incurred during signal onsets and offsets. Eq. 2.16 needs initials conditions for the first frame,i.e. $i = 0$. In [10] they propose:

$$\hat{\xi}(k,0) = \alpha + (1-\alpha)max[\xi(k,i) - 1, 0] \tag{2.17}$$

since it minimizes initial transition effects in the enhanced signal.

Following the work presented in [10], several improvements were made in the decision directed approach in order to reduce the bias and to improve the speed of adaptation which is controlled by the smoothing parameter α. In this work we use the modified decision directed approach proposed by Erkelens et. al [12] which corrects a bias at low SNR values

$$\hat{\xi}(k,i) = max[\alpha\frac{4}{\pi}\frac{|\hat{x}(k,i-1)|^2}{\sigma_W^2(k,i)} + (1-\alpha)(\xi(k,i) - 1), \xi_{min}] \tag{2.18}$$

where ξ_{min} was set to -19 dB. For an analytical derivation of the above equation see [12]. Fig 2.3 shows an example.

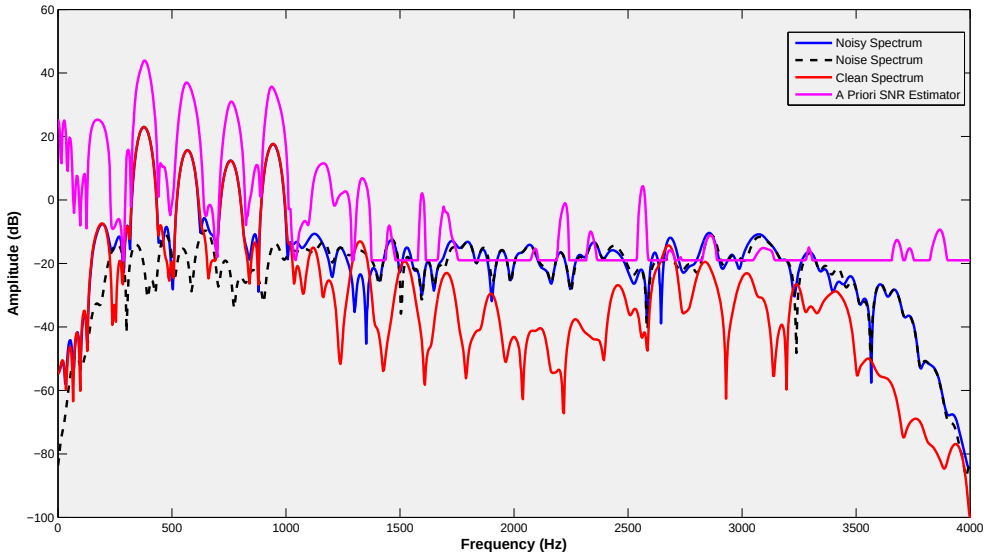

Figure 2.3: Modified decision directed approach (Eq. 2.18). The noise level was estimated by the minimum statistics approach (Appendix B).

2.3.2 Super-Gaussian Models

The use of Gaussian density to model speech is debatable. In [32] the density of speech DFT coefficients has been thoroughly investigated. It was concluded, by measuring histograms of speech DFT coefficients conditioned on *a priori* SNR values, that the observed density of speech DFT coefficients is more super-Gaussian, i.e. Laplace and Gamma distributions. Super-Gaussian pdfs are more peaky than the Gaussian density and possess heavier tails. It is important to mention that the preference for these super-Gaussian densities is influenced by conditioning on the *a priori* SNR and on the method used to estimate it [18].

Because of the observed super-Gaussian priors for speech DFT coefficients, there has been an increased interest over the last years to derive estimators for the clean DFT coefficients under these densities. In [32] MMSE estimators for the clean complex spectrum are derived under Laplace and Gamma Densities. In [13] it was shown that all known frequency domain MMSE estimators (for both the clean complex spectrum and the clean magnitude) can be derived as special cases under the generalized-Gamma speech prior density.

The complex spectrum estimators were derived by assuming that real and imaginary parts of DFT coefficients are independent and are both distributed as a double-sided generalized-Gamma density. The magnitude estimators were derived by assuming that the magnitude has a single-sided generalized-Gamma density and the phase has a uniform density. However, there do exist inconsistencies between the densities assumed in the cartesian domain and in the polar domain. The uniform phase distribution seems to be the most realistic assumption, since measured histograms show that phase is uniform and that real and imaginary parts of speech DFT coefficients are uncorrelated, but not independent [13, 28]. In [14] a new theoretical framework is presented that eliminates these inconsistencies by adopting the uniform phase distribution and dropping the independence assumption of real and imaginary parts. In [19]

comparison results between the estimators of the [14] and the conventional ones are presented. For a more detailed discussion of statistical-model based systems see the aforementioned papers.

2.3.3 Trained Statistical Models - Parametric Models

The methods discussed in the previous sections are optimal only within the framework of the statistical models that they assume. Rather than describing complex signals such as speech with models with few parameters, a more accurate method is to use more sophisticated statistical models such as hidden Markov models [9], Gaussian mixture models (GMM) [4] and codebooks [37,39] that have been trained using a representative database. The improved accuracy is at the expense of increased computational complexity and memory usage.

In these methods, the pdfs of the speech and noise processes are estimated from corresponding training sequences. They often apply certain constraints on the estimation process by using the fact that speech can be very well modeled as an autoregressive (AR) process (see Appendix A).As such these methods can exploit certain a priori information and can ensure that the enhanced speech signal satisfies certain spectral-temporal constraints. However, they need a parametric model of the noise process as well. Clearly modeling the noise process with HMMs or codebooks restricts the system to work only for certain noise types [18]. Moreover, not all noise types can be described well with a low order AR model (e.g. siren noise).

For a more detailed discussion of speech enhancement methods based on parametric models see the aforementioned papers and [27]. In the next section we discuss the codebook based approach described in [39], since it is relevant to the work presented in this thesis.

Codebook based approach

For speech enhancement Srinivasan et al. [39] assume that both the speech and the noise are described by independent AR processes. They use a priori information contained in speech and noise Linear Prediction Coefficients (LPC - Appendix A) codebooks. The problem is then one of estimating both the speech and the noise models (LPC and the corresponding scaling factors, i.e., excitation variances), using the observed noisy complex spectrum. The speech ss_i^* and the noise nn_i^* codebook indexes and the excitation variances $\sigma_{x,i}^{2*},\ \sigma_{w,i}^{2*}$ corresponding to the vectors that the indexes represent are obtained in a unified maximum likelihood (ML) framework according to:

$$\{ss_i^*, nn_i^*, \sigma_{x,i}^{2*}, \sigma_{w,i}^{2*}\} = \arg\max_{ss,nn} \max_{\sigma_{x,i}^2, \sigma_{w,i}^2} p(\mathbf{y_i}|\mathbf{a_x^{ss}}, \mathbf{a_w^{nn}}; \sigma_{x,i}^2, \sigma_{w,i}^2) \tag{2.19}$$

where $\sigma_{x,i}^2$ and $\sigma_{w,i}^2$ are the excitation variances of the clean speech and noise respectively, $\mathbf{a_x^{ss}} = [1\, a_{x_1}^{ss} \ \dots a_{x_p}^{ss}]$ is the LPC of the ss^{th} entry of the speech codebook and $\mathbf{a_w^{nn}} = [1\, a_{w_1}^{nn} \ \dots a_{w_q}^{nn}]$ is the LPC of the nn^{th} entry of the noise codebook with p and q being the respective linear prediction (LP) model orders. Note that i denotes the time-frame index, since the estimation is performed on a frame-by-frame basis. A schematic diagram of this method is shown in Fig. 2.4.

Under Gaussianity assumptions and using the equivalence between the log-likelihood and the Ikatura-Saito distortion (IS) [17] the estimation can be performed according to

$$\{ss_i^*, nn_i^*\} = \arg\min_{ss,nn}\{\min_{\sigma_{x,i}^2, \sigma_{w,i}^2} d_{IS}(P_{yy}(\omega_k, i)), \hat{P}_{yy}^{ss,nn}(\omega_k, i))\} \tag{2.20}$$

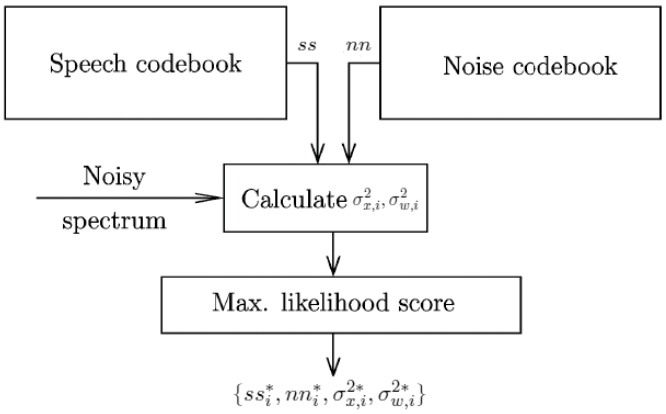

Figure 2.4: Block diagram of the codebook-based approach proposed in [39].

where d_{IS} is the Ikatura-Saito distortion measure given by

$$d_{IS}(P_{yy}(\omega_k,i),\hat{P}^{ss,nn}_{yy}(\omega_k,i)) = \frac{1}{2\pi}\int_0^{2\pi}(\frac{P_{yy}(\omega_k,i)}{\hat{P}^{ss,nn}_{yy}(\omega_k,i)} - \ln(\frac{P_{yy}(\omega_k,i)}{\hat{P}^{ss,nn}_{yy}(\omega_k,i)}) - 1)d\omega_k, \tag{2.21}$$

and

$$\hat{P}^{ss,nn}_{yy}(\omega_k,i) = \frac{\sigma^2_{x,i}}{|A^{ss}_x(\omega_k)|^2} + \frac{\sigma^2_{w,i}}{|A^{nn}_w(\omega_k)|^2} \tag{2.22}$$

the spectral envelope of the noisy speech based on the speech and the noise codebooks where

$$A^{ss}_x(\omega_k) = \sum_{l=0}^{p} a^{ss}_{x_l} e^{-j\omega_k l}, \quad A^{nn}_w(\omega_k) = \sum_{l=0}^{q} a^{nn}_{w_l} e^{-j\omega_k l} \tag{2.23}$$

are the complex spectrums of the ss^{th} vector from the speech codebook and the nn^{th} vector from the noise codebook respectively.

For given $A^{ss}_x(\omega_k)$ and $A^{nn}_w(\omega_k)$ and under the assumption of small modeling errors between $P^{ss,nn}_{yy}(\omega_k,i)$ and $\hat{P}^{ss,nn}_{yy}(\omega_k,i)$ [1], the excitation variances that minimize Eq. 2.21 are obtained according to

$$\mathbf{C}\begin{bmatrix}\sigma^2_{x,i}\\ \sigma^2_{w,i}\end{bmatrix} = \mathbf{D} \tag{2.24}$$

where

$$\mathbf{C} = \begin{bmatrix} \left|\left|\frac{1}{P^2_{yy}(\omega_k,i)|A^{ss}_x(\omega_k,i)|^4}\right|\right| & \left|\left|\frac{1}{P^2_{yy}(\omega_k,i)|A^{ss}_x(\omega_k,i)|^2|A^{nn}_w(\omega_k,i)|^2}\right|\right| \\ \left|\left|\frac{1}{P^2_{yy}(\omega_k,i)|A^{ss}_x(\omega_k,i)|^2|A^{nn}_w(\omega_k,i)|^2}\right|\right| & \left|\left|\frac{1}{P^2_{yy}(\omega_k,i)|A^{nn}_w(\omega_k,i)|^4}\right|\right| \end{bmatrix}$$

and

$$\mathbf{D} = \begin{bmatrix} \left|\left|\frac{1}{P_{yy}(\omega_k,i)|A^{ss}_x(\omega_k,i)|^2}\right|\right| \\ \left|\left|\frac{1}{P_{yy}(\omega_k,i)|A^{nn}_w(\omega_k,i)|^2}\right|\right| \end{bmatrix}$$

[1] using a series expansion for $ln(x)$ up to second order terms [39]

where $||f(\omega_k, i)|| = \int |f(\omega_k, i)|\, d\omega_k$

To get a better spectral fit between the modeled and the observed noisy spectra, such that the above mentioned approximations can be made valid, it is necessary to use the AR spectrum of the observed noisy speech rather than the DFT-based periodogram. Thus, P_{yy} in the preceding equations is obtained as

$$P_{yy}(\omega_k, i) = \frac{\sigma^2_{y,i}}{|A_y(\omega_k, i)|^2}, \quad A_y(\omega_k, i) = \sum_{l=0}^{p} a^i_{y_l} e^{-j\omega_k l} \tag{2.25}$$

The estimation process can be summarized as follows. For each pair of speech and noise spectral shapes from the representative codebooks, the excitation variances are calculated according to Eq. 2.24 and the distortion(Eq. 2.21) is evaluated. Codebook combinations that result in a negative value for either the speech or noise excitation variance are discarded since they are infeasible due to the non-negativity constraints on the variance. The speech and noise spectra globally minimizing the distortion are determined. These spectra together with the corresponding variances represent the ML estimate of the speech and noise segment i. Note that for each frame of noisy speech, the noise codebook is augmented with an estimate of the noise LP vector obtained from the noisy observation, using the minimum statistics approach ([31], Appendix B). Thus, one of the trained noise codebook entries is chosen only if it provides a better representation of the noise than the minimum statistics approach. Fig.2.5 and Fig 2.6 show the speech and noise models obtained from the above described approach for a voiced speech segment and an unvoiced speech segment, respectively, degraded by white Gaussian noise at 20 dB overall SNR. Note that in our implementation the noise envelope is obtained only from the minimum statistics approach.

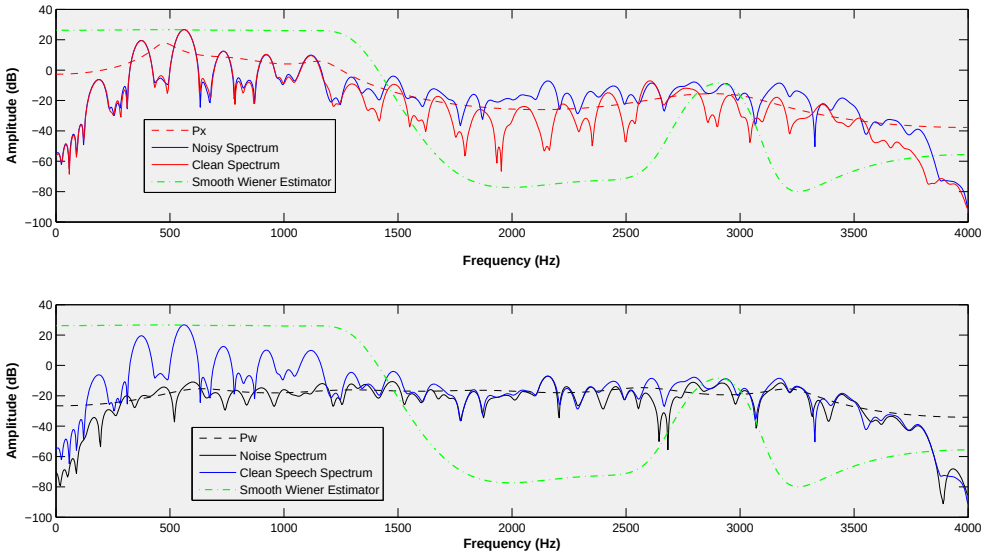

Figure 2.5: ML estimation voiced example.

The above estimation scheme can handle both quickly changing noise envelopes and quickly changing noise energy. The difference entries of the noise codebook deal with changing

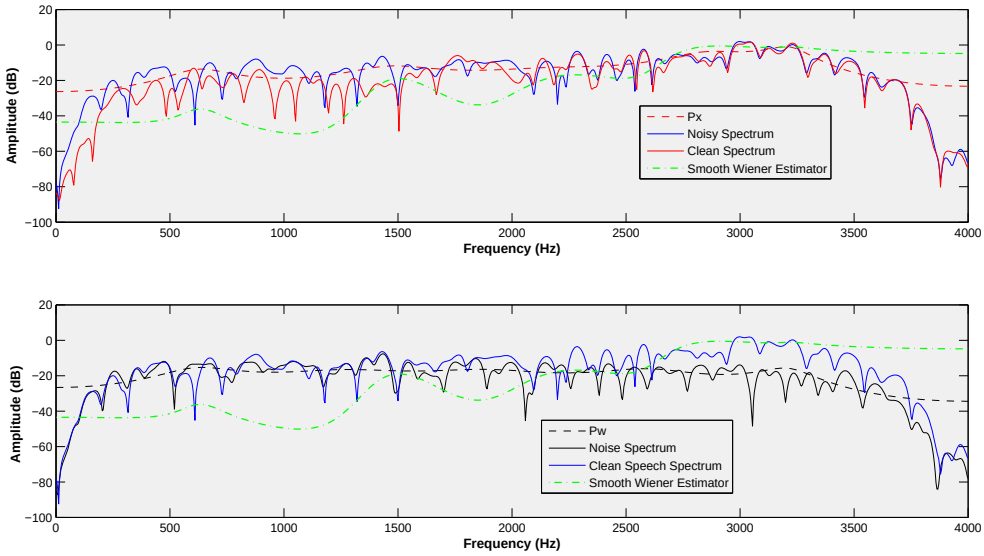

Figure 2.6: ML estimation unvoiced example.

noise envelopes while the instantaneous estimation of the excitation variances handles changing noise energy. The estimated parameters of speech and noise can be used to construct the following wiener filter (smoother across frequency bins than the one of Eq. 2.7):

$$H(\omega_k, i) = \frac{\frac{\sigma^{2*}_{x,i}}{|A^{ss*}_x(\omega_k, i)|}}{\frac{\sigma^{2*}_{x,i}}{|A^{ss*}_x(\omega_k, i)|} + \frac{\sigma^{2*}_{w,i}}{|A^{nn*}_w(\omega_k, i)|}} \tag{2.26}$$

which can be used to estimate the clean signal from the noisy observation. In Fig 2.5 and Fig 2.6 the respective wiener filter estimator of Eq. 2.26 is depicted (scaled).

Chapter 3

Prior Knowledge

Generally, most speech enhancement methods result in some loss of speech information, the severity of which depends on the SNR and the speech enhancement scheme. In this chapter we present two resent methods that try to enhance the performance of a conventional speech enhancement method by applying post-processing restoration modules. The first one, proposed in [44], uses a codebook mapping technique to restore the lost or suppressed speech information, while the second one, originally proposed in [33], uses nonlinearity to preserve speech harmonics that have been degraded by the conventional noise reduction schemes. The aforementioned methods will be used in chapter 5 in order to enhance the excitation signal obtained from a conventional speech enhancement method.

3.1 Codebook Based Post-Processing Method

Zavarehei et al. [44] proposed a post-processing algorithm for retrieving parts of the speech spectrum that may be lost to noise or suppressed by the conventional speech enhancement methods. Fig.3.1 shows the block diagram of the speech enhancement scheme that was proposed. They focused on reviving severely damaged subbands of speech using a harmonic plus noise model (HNM) [24]. Application of HNM enforces a harmonically structured reconstruction of speech spectral amplitude. Furthermore, HNM provides a good model for tracking natural energy contours of the signal across time and frequency. They proposed a weighted codebook mapping algorithm (WCBM) for estimation and restoration of the HNM parameters. Incorporation of a priori knowledge of speech is done through codebooks, trained on clean speech. The codebooks are accessed using a weighted distance measure of the codebook entries from the enhanced speech vector. A weight vector is adaptively obtained according to the subbands' SNRs. Higher weights are given to less-distorted subbands (high SNR values) and lower weights to those which suffer from distortion (low SNR values). The enhanced speech is a weighted interpolation of the output of the conventional speech enhancement scheme and the codebook output. WCBM exploits the high correlation between the succeeding harmonics of a speech frame i and aims at reviving those that are distorted.

3.1.1 Harmonic Noise Model of speech

The useful effects, of using HNM for modeling the harmonic structure of speech, on the intelligibility and quality of synthesized speech have made these models popular for various

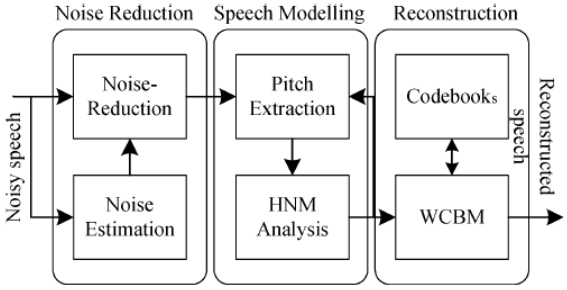

Figure 3.1: Block diagram of the speech enhancement system proposed by Zavarehei et al., source [44].

applications [34,40,43]. According to HNM the speech segments are analyzed and synthesized entirely in spectral amplitude domain ($|X(f_k, i)|$). The unprocessed speech phase is used for resynthesis.

From each harmonic subband three parameters are extracted:

- the harmonic amplitude $A(f_h, i)$, which is represented by the square-root energy of the subband

$$A(f_h, i) = \sqrt{\int_{f_h-\Delta f}^{f_h+\Delta f} X^2(f_k, i) df_k} \tag{3.1}$$

- the harmonicity $V(f_h, i)$, which represents the voicing degree of the subband

$$V(f_h, i) = 1 - \frac{\sqrt{\int_{f_h+\Delta f}^{f_h-\Delta f} (X(f_k, i) - A(f_h, i)G(f_k - f_h))^2 df_k}}{A(f_h, i)} \tag{3.2}$$

- the harmonic central frequency f_h, which is extracted locally to maximize the harmonicity of the signal around that harmonic

$$f_h = \arg\max_{f_h} V(f_k, i),\ hF_0 - \Delta F_H < f_k < hF_0 + \Delta F_H \tag{3.3}$$

where i is the time-frame index, h is the harmonic index, F_0 is the fundamental frequency, $2\Delta f$ is the harmonic bandwidth, ΔF_H is an empirically search range and $G(f_k) = \alpha exp(-(\frac{\beta f_k}{\Delta f})^2)$ a Gaussian-shaped function used to model the harmonics. The parameters α and β are calculated in such a way to preserve unity power of the Gaussian shape spectrum, i.e. $\int_{-\Delta f}^{\Delta f} G^2(f_k) df_k = 1$.

Given the set of HNM parameters for a speech segment i, the spectral amplitudes are regenerated according to:

$$|X_{HNM}(f_k, i)| = \sum_{h=1}^{N_H} A(f_h, i)(V(f_h, i)G(f_k - f_h) + (1 - V(f_h, i))R(f_k - f_h)) \tag{3.4}$$

where $X_{HNM}(f_k, i)$ is the HNM-synthesized amplitude spectrum, $R(f_k)$ is the noise component of the harmonic subband (Rayleigh distributed random variable satisfying $\int_{-\Delta f}^{\Delta f} R^2(f_k)df_k = 1$ [10]) and N_H the number of harmonics of the segment i depending on the fundamental frequency F_0 and the sampling frequency F_S. Note that each harmonic subband is reconstructed as a combination of voiced and unvoiced parts. The harmonicity specifies the proportion of voiced and unvoiced energy in each subband. In other words, harmonicity helps to avoid the hard decision of voiced or unvoiced subbands.

3.2 Harmonic Regeneration

As it has been mentioned before, one major limitation that exists in the classical speech enhancement schemes is that some harmonics are considered as noise only components (especially in low SNR values) and consequently are suppressed by the speech enhancement system. To overcome this problem, Plapous et al. [33,34] proposed a method called harmonic regeneration noise reduction (HRNR) that takes into account the harmonic structure of the speech. In this approach, the output signal of a common noise reduction algorithm is further processed to create an artificial signal where the missing harmonics have been automatically regenerated. Then this artificial signal is used to compute a suppression gain that tries to preserve all the harmonics of the clean speech signal.

3.2.1 Principle of harmonic regeneration

A simple and efficient way to restore speech harmonics consists of applying a nonlinear function NL (absolute value, minimum or maximum relative to a threshold, etc.) to the time speech segment obtained from a common noise reduction method. Then the artificially restored segment is obtained by

$$x_{harmo,i}(n) = NL(\hat{x}_i(n)) \tag{3.5}$$

Note that the restored harmonics of $x_{harmo,i}(n)$ are created at the same positions as the clean speech ones. This very interesting and important characteristic of the method is ensured because of nonlinearity in the time domain (see next section for more details). For illustration, Fig. 3.2 shows the typical behavior of the nonlinearity. Fig. 3.2(a) represents a clean speech segment and Fig. 3.2(b) the enhanced speech segment obtained with wiener filter (Eq. 2.10). It appears that some harmonics have been completely suppressed or severely degraded. Fig 3.2(c) shows the artificially restored segment obtained using Eq. 3.5 where the nonlinearity (half wave rectification, i.e., the maximum relative to 0) has restored the harmonics. However, the harmonic amplitudes of the artificial signal are biased compared to clean speech. As a consequence this signal cannot be used directly as clean speech estimation. Nevertheless, it possesses very useful information that can be exploited to refine the *a priori* SNR. This is

$$\xi(k,i)^{HRNR} = \frac{\rho(k,i)\,|\hat{x}(k,i)|^2 + (1-\rho(k,i))\,|x_{harmo}(k,i)|^2}{\hat{P}_{WW}(k,i)} \tag{3.6}$$

The $\rho(k,i)$ parameter controls the mixing level of $|\hat{x}(k,i)|^2$ and $|x_{harmo}(k,i)|^2$ ($0 \leq \rho(k,i) \leq 1$). The value of which should be close to 1 when the estimation provided by the speech

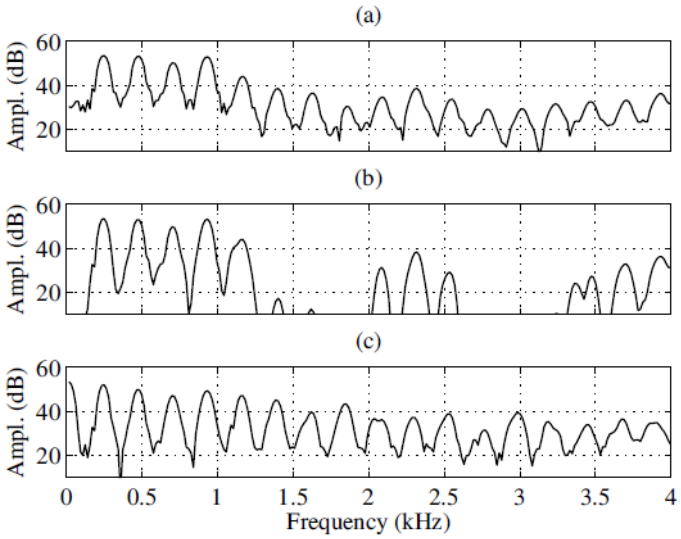

Figure 3.2: Effect of the nonlinearity, source [33].

enhancement system is reliable and close to 0 when it is unreliable. In [34], in order to match this behavior, they proposed $\rho(k,i)$ to be equal to the Wiener filter gain function while in [33] the parameter ρ was chosen to be a constant value.

The refined *a priori* SNR $\xi(k,i)^{HRNR}$ is then used to compute an estimate of the clean speech, in which all the harmonics will be presence. Following Eq. 1.6, this is

$$\hat{x}(k,i) = g(\xi^{HRNR}(k,i), \gamma(k,i), y(k,i)) \tag{3.7}$$

3.2.2 Theoretical Analysis of Harmonic Regeneration

To analyze the harmonic regeneration step we focus on a particular nonlinear function, the half wave rectification

$$x_{harmo,i}(n) = max(\hat{x}_i(n), 0) = \hat{x}_i(n)u(\hat{x}_i(n)) \tag{3.8}$$

where u is the step function. Fig. 3.3 shows a segment of an enhanced voiced speech signal $\hat{x}_i(n)$ (dotted line) and the corresponding $u(\hat{x}_i(n))$ (dashed line). It can be observed that the signal $u(\hat{x}_i(n))$ amounts to a repetition of an elementary waveform (solid line) with periodicity T_0, corresponding to the segments fundamental frequency. By definition, the Fourier Transform (FT) of $u(\hat{x}_i(n))$ comes down to a sampled version of the elementary waveform's FT

$$FT\{u(\hat{x}_i(n))\} = \frac{1}{T}\sum_{m=-\infty}^{+\infty} R_i(\frac{m}{T_0})\delta(f - \frac{m}{T_0}) \tag{3.9}$$

where δ denotes the Dirac function, f the continuous frequency and $R_i(\frac{m}{T_0})$ is the FT of the elementary waveform taken at discrete frequencies $\frac{m}{T_0}$. Note that the sampling frequency coincides with the harmonic positions of the elementary waveform $(\frac{1}{T_0}, \frac{2}{T_0}, \ldots)$.

Thus the spectrum of the $x_{harmo,i}(n)$ is the convolution between the spectrum of $\hat{x}_i(n)$ (Fig. 3.2(b)) and the harmonic comb $FT\{u(\hat{x}_i(n))\}$. The fact that these two spectrums have

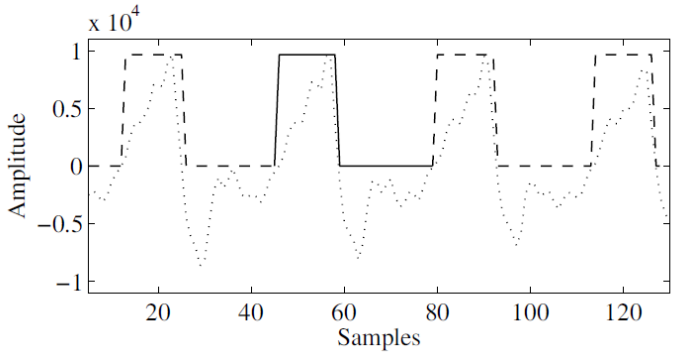

Figure 3.3: Voiced speech segment, source [33].

the same fundamental frequency $F_0 = \frac{1}{T_0}$ explains the harmonic regeneration phenomenon. The harmonics are created at the same position as the clean speech ones. Furthermore, the envelope of $FT\{u(\hat{x}_i(n))\}$, rapidly decreases when $|m|$ increases, thus a missing harmonic is generated only using the information of the few neighboring harmonics.

It is also important to investigate the behavior of the harmonic regeneration process for unvoiced speech. Let us consider an hybrid signal where the lower part of the spectrum is voiced and the upper unvoiced. The FT of $u(\hat{x}_i(n))$ will still be a harmonic comb, with fundamental frequency imposed by the voiced part. Then the voiced part of the spectrum of the $x_{harmo,i}(n)$ will be exactly the same as in voiced only case. However since each frequency bin is obtained using only its corresponding neighboring area in the spectrum of $\hat{x}_i(n)$, the unvoiced spectrum part will lead to an unvoiced restored spectrum. In the case of an unvoiced speech segment, the $FT\{u(\hat{x}_i(n))\}$ will be an undetermined spectrum and the convolution will lead to an unvoiced spectrum. Thus, the unvoiced segment will not be degraded.

Chapter 4

Estimation of Speech Spectral Envelope

In this chapter we present a post-processing speech spectral envelope restoration approach for enhancing the speech spectral envelope obtained after applying LPA to the output signal of a conventional speech enhancement method ("enhanced envelope"). A schematic diagram of this approach is shown in Fig. 4.1. The restoration is achieved through incorporation of a priori knowledge of speech spectral envelopes stored in an LSF codebook, pretrained on clean speech. The restoration is accomplished using a weighted codebook mapping algorithm.

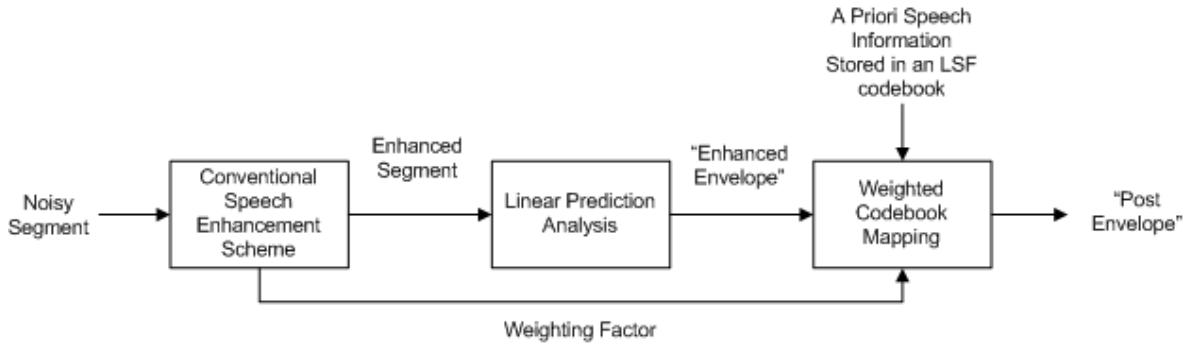

Figure 4.1: Block diagram of the post-processing speech spectral envelope restoration approach.

4.1 LSF Codebook

Codebook mapping is a simple technique of incorporating prior knowledge in the signal estimation and extrapolation process. This technique has been widely used in different aspects of speech processing, including speech enhancement [37, 39, 44], bandwidth extension [25], etc. In our approach a 10 bit (1024 codewords) speech codebook of LSFs was constructed. This size was chosen since on average it provides a good trade off between quantization errors and memory usage. The dimension of the codebook, i.e. the order p of the LPA, was 10 [23]. For training the codebook we used the entire TIMIT-TRAIN database [16], which consist of 4620 clean speech sentences. The sampling frequency was $F_s = 8000\,Hz$ and the speech signals

were limited to telephone bandwidth ($300 - 3400\,Hz$). We used $50\,\%$ overlapping frames of $32\,ms$ and a cosine-squared window, which has the perfect recunstruction property. Beginning and trailing silences and frames of a sentence with energy lower than $40\,dB$ of the maximum frame energy of the sentence were excluded from the training phase. This resulted in around 750000 training vectors. The Linde, Buzo, Gray (LBG) vector quantization algorithm was used for training the codebook [5]. The basic principles of the LBG algorithm are presented in Appendix C. Fig 4.2 and Fig 4.3 show the original and the quantized envelopes of a voiced clean segment and an unvoiced clean segment, respectively.

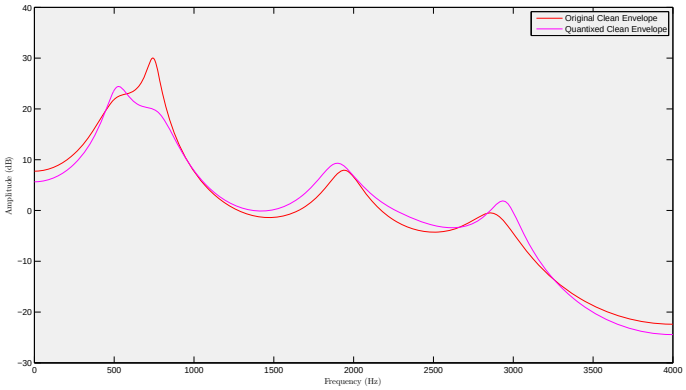

Figure 4.2: Quantization effect on the envelope of a voiced segment.

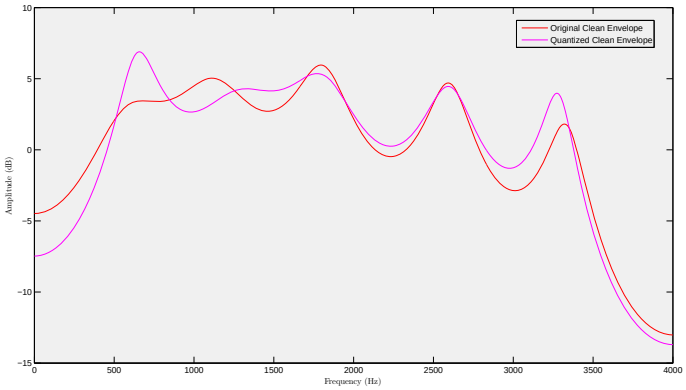

Figure 4.3: Quantization effect on the envelope of a unvoiced segment.

4.2 Codebook Mapping

In order to apply our algorithm, first the LPC of the output of the speech enhancement algorithm are extracted. The order of LPA of the enhanced frame is chosen to be the same with the dimension of the codebook, i.e. $p = 10$. The "distance" between the "enhanced envelope" and the envelopes represented by the codewords ("codebook envelopes") are estimated in

terms of log-spectral distortion (SD)[1]

$$SD_m^i = \sqrt{\frac{1}{2\pi}\int_0^{2\pi}(10\log_{10}(\frac{1}{|A_{\hat{x}}(\omega_k,i)|^2}) - 10\log_{10}(\frac{1}{|A_m(\omega_k)|^2}))^2 d\omega_k} \tag{4.1}$$

where i the time-frame index, since our method is applied on a frame-by-frame basis, and

$$A_{\hat{x}}(\omega_k,i) = \sum_{l=0}^{p} a_{\hat{x}_l}^i e^{-j\omega_k l}, \quad A_m(\omega_k) = \sum_{l=0}^{p} a_{m_l} e^{-j\omega_k l} \tag{4.2}$$

are the complex spectrums of the "enhanced envelope" and the m^{th} vector from the speech codebook, respectively, with $\mathbf{a_{\hat{x}}}^i = [1\, a_{\hat{x}_1}^i \ldots, a_{\hat{x}_p}^i]$ and $\mathbf{a_m} = [1\, a_{m_1} \ldots, a_{m_p}]$ being the respective LPC. An estimate of the codebook mapped LSF vector is obtained from the L entries of the codebook with the smallest SD, as

$$\boldsymbol{\lambda_{CB}}^i = \sum_{j=1}^{L} q_j^i \lambda_j \tag{4.3}$$

where q_j is the weight of the codeword λ_j (LSF representation of $\mathbf{a_j}$) and is proportional to the reciprocal of the SD_j^i between $\mathbf{a}_j$ and $\mathbf{a_{\hat{x}}}^i$:

$$q_j^i = \frac{1}{SD_j^i \sum\limits_{L} \frac{1}{SD_j^i}} \tag{4.4}$$

Note that the interpolation is performed in the LSF domain to assure stable spectral envelope (see section A.5).

In this work and for a codebook size of 10 bits, L was experimentally set equal to 6. For optimizing the parameter L, we quantized all the 192 clean speech utterances of the TIMIT core test set [16]. Table 4.2 shows the average performance, in terms of mean SD (Eq. 4.1 - computed explicitly from the telephone bandwidth frequency bins) of the 10 bits codebook for different values of the parameter L. Note that the SD for an utterance is computed as the average of the instantaneous SD for the individual frames.

L	1	2	3	4	5	6	7	8	9	10
SD(dB)	2.56	2.14	2.02	1.97	1.96	1.96	1.97	1.98	1.99	2.01

Table 4.1: Optimizing parameter L

An alternative approach would be to introduce in Eq. 4.1 a frequency dependent weighting factor which will be used as a measure of the reliability of the speech enhancement system. In other words, for frequency bins with high SNR value the weight has to be close to 1 while for frequency bins where the noise is dominant the weight has to be close to 0. For obtaining that weight two different factors, representing the reliability of the speech enhancement system, were considered

[1]SD is an objective quality measure of speech quality that has been reported to have a high correlation with subjective quality

- the a priori SNR $\xi(\omega_k, i)$
- the speech enhancement system's gain function $H(\omega_k, i)$

In practice, the modified decision directed *a priori* SNR estimator (Eq. 2.18) was used for computing the *a priori* SNR and the Wiener filter (Eq. 2.10) was used as the gain function. Fig. 4.4 shows a voiced speech segment and the aforementioned frequency dependent weighting factors, where *a priori* SNR is normalized according to Eq 4.6. Notice that those factors are scaled to make the figure clearer.

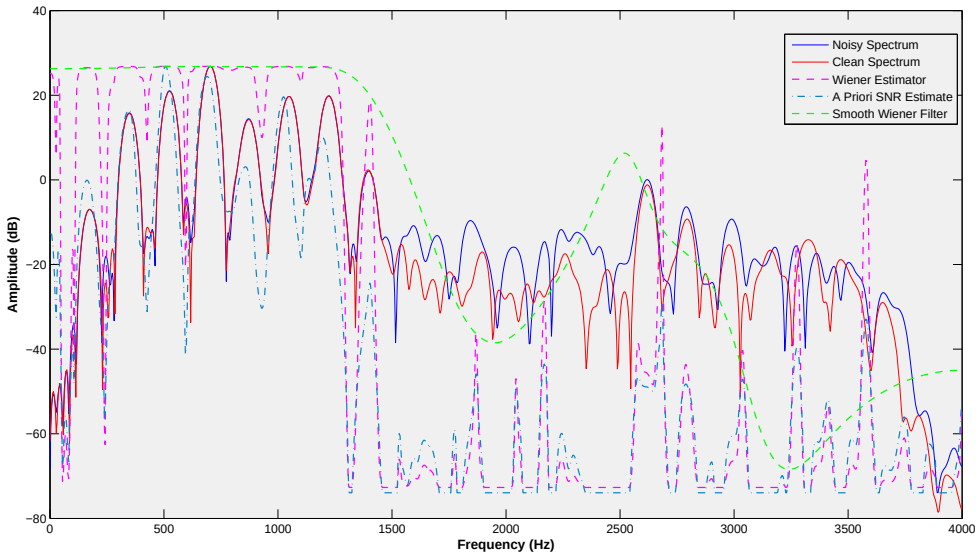

Figure 4.4: Possible weights for the weighted codebook mapping envelope estimation approach.

Since *a priori* SNR and Wiener filter are evolved sharply across frequency a third weighting factor was considered. This is the smooth Wiener filter (Eq. 2.26), previously discussed in section 2.3.3. To derive the smooth Wiener filter gain function we use the same speech codebook as the one described in section 4.1 and a single entry noise codebook consisted only of the LSF obtained from the minimum statistics noise tracking approach (Appendix B). In Fig.4.4 the smooth version of the Wiener filter (ML Wiener) is depicted, as well. Notice that its peaks occur close to the formants regions.

In the weighted codebook mapping approach it might happen that for the reliable regions, i.e. frequency bins where the weight is close to 1, the "enhanced envelope" and the "codebook envelope" have quite the same shape but different power levels. Therefore, a compensation factor c_m^i is introduced in the SD measure

$$SD_m^i = \sqrt{\frac{1}{2\pi}\int_0^{2\pi} w(\omega_k, i)(10\log_{10}(\frac{1}{|A_{\hat{x}}(\omega_k, i)|^2}) - 10\log_{10}(\frac{c_m^i}{|A_m(\omega_k)|^2}))^2 d\omega_k} \tag{4.5}$$

where $w(\omega_k, i)$ is the weighting factor given by

$$w(\omega_k,i) = \begin{cases} \frac{log_{10}(\hat{\xi}(\omega_k,i)) - min(log_{10}(\hat{\xi}(\omega_k,i)))|_{i=const}}{max(log_{10}(\hat{\xi}(\omega_k,i)))|_{i=const} - min(log_{10}(\hat{\xi}(\omega_k,i)))|_{i=const}} \\ \frac{\hat{\xi}(\omega_k,i)}{\hat{\xi}(\omega_k,i)+1} \\ \frac{\frac{\sigma^{2*}_{x,i}}{|A^{ss*}_x(\omega_k,i)|}}{\frac{\sigma^{2*}_{x,i}}{|A^{ss*}_x(\omega_k,i)|} + \frac{\sigma^{2*}_{w,i}}{|A_w(\omega_k,i)|}} \end{cases} \tag{4.6}$$

Note that Eq. 4.6 rescales the *a priori* SNR (in dB scale) to values between 0 and 1. Furthermore, since in this thesis we are using the modified decision directed approach (Eq. 2.18) to estimate the *a priori* SNR, the Wiener gain function is bounded to values between 0.01 and 1.

Given $A_{\hat{x}}(\omega_k, i)$ and $A_m(\omega_k)$ the optimal compensation factor c^i_m can be determined by differentiating Eq. 4.5 with respect to c^i_m and setting the result equal to O. By leaving out the square root for simplicity we get

$$c^i_m = 10^{\frac{\int_0^{2\pi} w(\omega_k,i) \log_{10}(\frac{|A_m(\omega_k)|}{|A_{\hat{x}}(\omega_k,i)|}) d\omega_k}{20 \int_0^{2\pi} w(\omega_k,i) d\omega_k}} \tag{4.7}$$

As before, an estimate of the weighted codebook mapped LSF vector is obtained from the L entries of the codebook with the smallest SD. Fig. 4.5 and Fig. 4.6 show the typical behavior of the codebook mapping method for a voiced and an unvoiced envelope, respectively. It appears that many formants are regenerated using the codebook mapping technique, whereas they are suppressed in the conventional speech enhancement scheme. Note that we use the Wiener filter as a weighting factor and highpass filtered stationary white Gaussian noise at 5dB global SNR to make the restoration effect clearer.

In order to evaluate the performance of the four weighting factors (no weight, Wiener estimator, normalized a priori SNR, ML Wiener) we apply the above described codebook mapping algorithm to all the 192 speech utterances of the TIMIT core test set [16], to which computer generated stationary white Gaussian noise, limited to telephone bandwidth, has been added at four different input SNR levels, i.e. 0dB, 5dB, 10dB and 15dB. Table 4.2 shows the mean SD scores (computed explicitly from the telephone bandwidth frequency bins) for the "noisy envelope"[2], the "enhanced envelope" and the "post envelope"[3]. It can been seen that at most SNR values (except at 15 dB) the gain function weighted post-processed method results in slightly better performance (in terms of SD) than all the other methods. The normalized a priori SNR performs worst of all the other methods suggesting that weights with many zero values across frequency do not comprise a good choice. For further evaluation of the post-processing speech spectral envelope restoration approach see section 6.2.

[2]We use the term "noisy envelope" to denote the envelope obtained from the noisy signal after LPA

[3]We use the term "post envelope" to denote the envelope obtained from the post-processed speech spectral envelope restoration method presented in this chapter

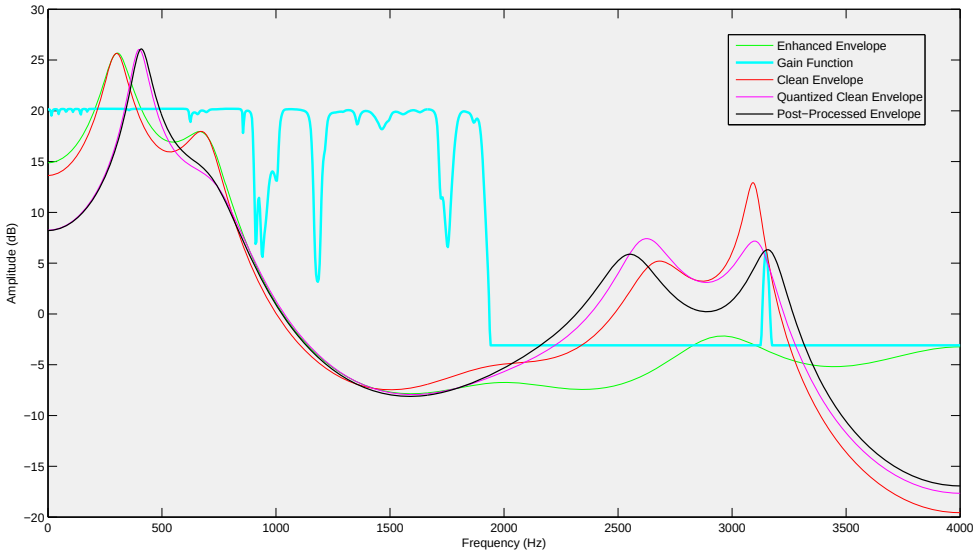

Figure 4.5: Codebook mapping for a voiced segment

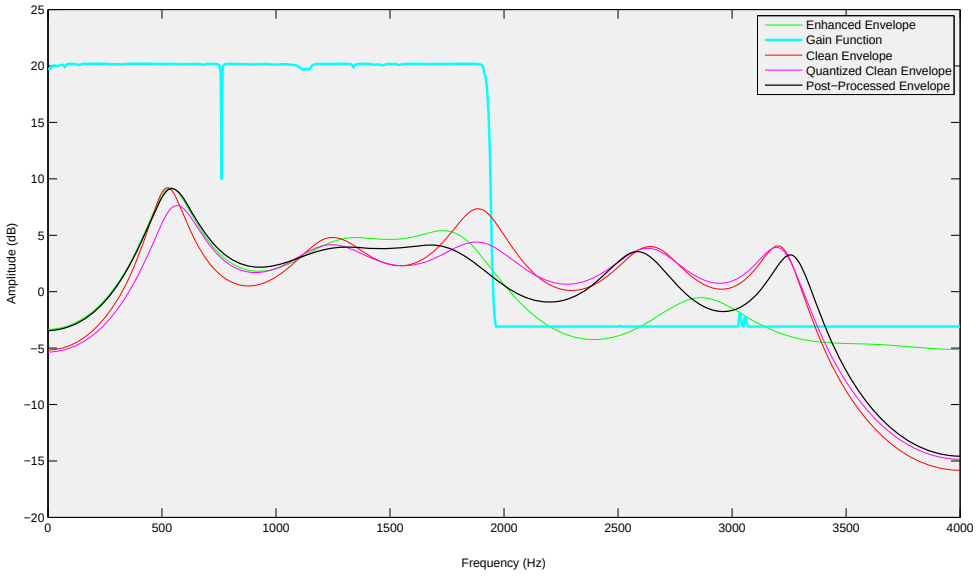

Figure 4.6: Codebook mapping for an unvoiced segment

	SD (dB)					
	"Noisy	"Enhanced	"Post Envelopes"			
SNR	Envelope"	Envelope"	No Weight	Wiener	ML Wiener	A Priori SNR
0	8.10	7.04	6.97	6.87	6.96	7.55
5	7.24	6.03	5.93	5.88	5.96	6.55
10	6.18	5.21	5.05	5.04	5.09	5.62
15	5	4.57	4.76	4.81	4.82	5.43

Table 4.2: Evaluating the weighting factors

Chapter 5

Estimation of Speech Excitation

In this chapter we present two different approaches for enhancing the excitation signal obtained after conducting LPA to the output signal of a conventional speech enhancement method ("enhanced excitation"). The first one, based on [33] (section 3.2), uses nonlinearity to preserve speech excitation harmonics that have been degraded by the conventional noise reduction schemes, while the second one, inspired by [44] (section 3.1), uses a codebook mapping technique to restore the lost or suppressed excitation harmonics.

5.1 Harmonic Regeneration of the Excitation

As it has been discussed in section 3.2 a simple and efficient way to restore speech harmonics consists of applying a nonlinear function NL (absolute value, minimum or maximum relative to a threshold, etc.) to the time speech segment obtained from a common noise reduction method.

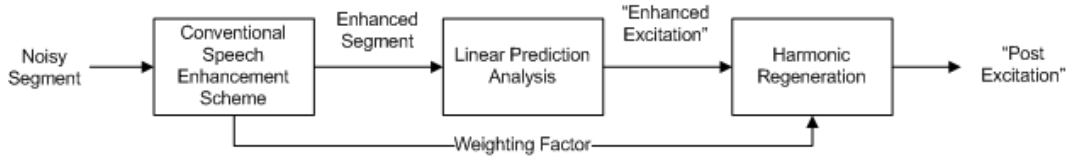

Figure 5.1: Block diagram of the HR excitation approach.

We apply that method directly to the "enhanced excitation" (Fig 5.1). Then the artificially restored excitation is obtained by

$$e_{harmo,i}(n) = NL(\hat{e}_i(n)) \tag{5.1}$$

where $\hat{e}_i(n)$ the "enhanced excitation", i the time frame index since the algorithm is applied on a frame-by-frame basis and n the discrete-time index. However, the amplitudes of the artificially restored excitation are biased compared to the "enhanced excitation". In order to compensate for this bias we propose to multiply the complex spectrum of the artificially

restored excitation $E_{harmo}(\omega_k, i)$ with a factor $c(i)$, given by

$$c(i) = \sqrt{\frac{\int_0^{2\pi} \left|\hat{E}(\omega_k, i)\right|^2 d\omega_k}{\int_0^{2\pi} |E_{harmo}(\omega_k, i)|^2 d\omega_k}} \tag{5.2}$$

where $\hat{E}(\omega_k, i)$ is the complex spectrum of the "enhanced excitation". In this way we conserve the energy on a frame-by-frame basis.

The above artificially restored excitation $e_{harmo,i}(n)$ contains useful information that can be exploited to refine the estimated speech excitation's amplitude

$$|\tilde{E}(\omega_k, i)| = \rho(\omega_k, i)|\hat{E}(\omega_k, i)| + (1 - \rho(\omega_k, i))|c(i)E_{harmo}(\omega_k, i)| \tag{5.3}$$

The unprocessed phase of the "enhanced excitation" is used for estimating the "HR excitation's"[1] complex spectrum. The parameter $\rho(\omega_k, i)$ is used to control the mixing level of $|\hat{E}(\omega_k, i)|$ and $|E_{harmo}(\omega_k, i)|$, with $0 \leq \rho(\omega_k, i) \leq 1$. The value of which should be close to 1 when the estimation provided by the speech enhancement system is reliable and close to 0 when it is unreliable. To match this behavior any of the factors of Eq. 4.6 can be used. Fig 5.1 shows the typical behavior of the harmonic regeneration method. Notice that the stationary white Gaussian noise is bandlimited to make the regeneration effect clearer. It is clear that the most of the missing harmonics are regenerated. For the evaluation of the harmonic regeneration model of the excitation see section 6.3.

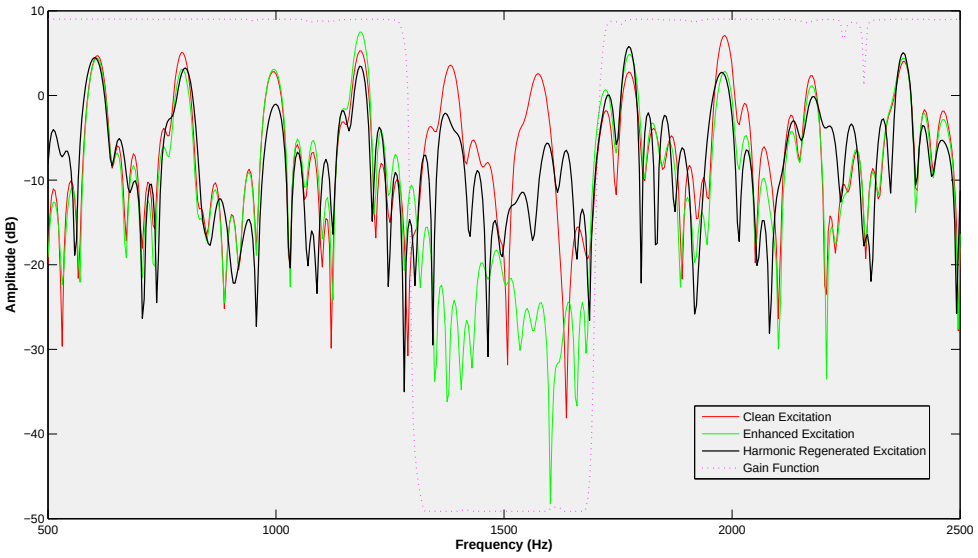

Figure 5.2: Effect of the harmonic regeneration method on the excitation of a voiced frame.

[1]We use the term "HR excitation" to denote the excitation obtained from the post-processing excitation restoration method presented in this section

5.2 Codebook Based Estimation of the Excitation

Zavarehei et al. [44] proposed a post-processing algorithm for retrieving parts of the speech spectrum that may be lost to noise or suppressed by the conventional speech enhancement methods. They focused on reviving severely damaged subbands of speech using a HNM. For a detailed presentation of Zavarehei's post-processing speech enhancement method see section 3.1. We apply that method for enhancing the excitation signal obtained after conducting LPA to the signal on the output of a conventional speech enhancement method.

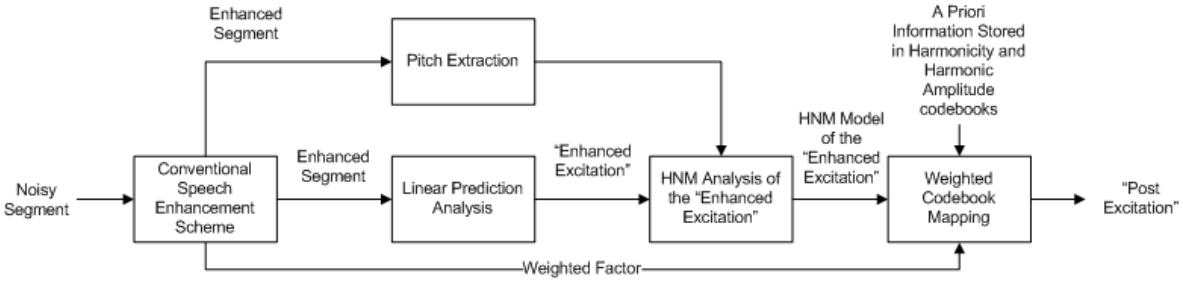

Figure 5.3: Block diagram of the HNM excitation estimation approach.

5.2.1 Harmonic Noise Model of Excitation

A variant of the HNM is used to model the excitation amplitude $|E(f_k, i)|$. Three parameters are extracted from each subband:

- the harmonic amplitude $A(f_h, i)$,
- the harmonicity $V(f_h, i)$ and
- the harmonic frequency f_h^i

The harmonic amplitudes represent the square root energy of the corresponding subband. The harmonicity of each harmonic is a real-valued measure between 0 and 1 that represents the voicing degree of each subband. Given a set of HNM parameters for a segment i the excitation amplitude can be modeled as a mixture of N_H harmonics and noise as

$$|E_{HNM}(f_k, i)| = \sum_{h=1}^{N_H} A(f_h^i, i)(V(f_h^i, i)G(f_k - f_h^i) + (1 - V(f_h^i, i))R(f_k - f_h^i)) \tag{5.4}$$

where $G(f_k)$ is a Gaussian-shaped function which models the shape of each harmonic and $R(f_k)$ is the Rayleigh distributed noise component of the excitation. The number of harmonics N_H of the segment i depends on the fundamental frequency F_0 and the sampling frequency F_S, i.e $N_H = \left\lfloor \frac{\frac{F_S}{2} - \frac{F_0}{2}}{F_0} \right\rfloor$. Note that the unprocessed excitation phase is used for resynthesis of the excitation's complex spectrum.

For each segment i the fundamental frequency (pitch frequency) F_0 is extracted using an autocorrelation based method [7]. Due to possible inaccuracies of pitch frequency estimate, multiples of pitch frequency estimate might not exactly coincide with the "true" harmonic

frequencies. Thus the exact frequency of each harmonic is obtained locally to maximize the harmonicity around that harmonic

$$f_h^i = \arg\max_{f_k} V(f_k, i),\ |f_k - hF_0| < 30 \tag{5.5}$$

where the search range is limited to $30\,Hz$ around the nominal value of each harmonic, i.e. hF_0 with $h = 1, \ldots, N_H$. $V(f_k, i)$ is the harmonicity at frequency f_k defined as

$$V(f_k, i) = 1 - \frac{\sqrt{\int_{f_k - \frac{F_0}{2}}^{f_k + \frac{F_0}{2}} |E(\nu, i) - A(f_k, i)G(\nu - f_k)|^2 d\nu}}{A(f_k, i)} \tag{5.6}$$

and $A(f_k, i)$ is the harmonic amplitude at frequency f_k

$$A(f_k, i) = \sqrt{\int_{f_k - \frac{F_0}{2}}^{f_k + \frac{F_0}{2}} |E(\nu, i)|^2 d\nu} \tag{5.7}$$

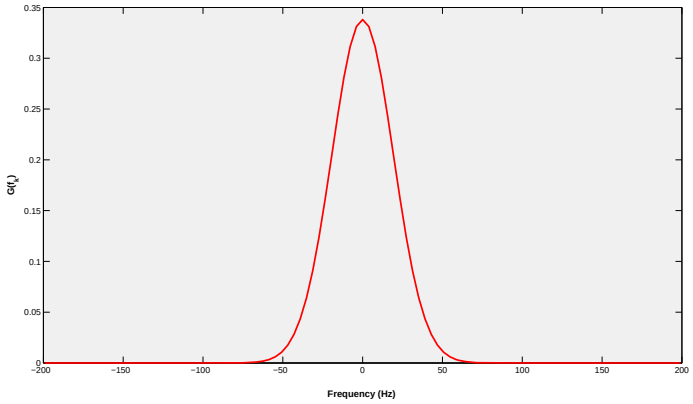

Figure 5.4: Gaussian window used to model excitation harmonics for $F_0 = 200Hz$.

The Gaussian-shaped function (Fig .5.4) is given by

$$G(f_k, i) = \begin{cases} \alpha e^{-\beta f_k^2}, & |f_k| \leq \frac{F_0}{2} \\ 0, & \text{elsewhere} \end{cases} \tag{5.8}$$

where $\beta = 0.0013\bar{4}$ and $\alpha = \frac{1}{\sqrt{\int_{-F_0/2}^{F_0/2} e^{-2\beta f_k^2} df_k}}$ which results in a unity power Gaussian shaped spectrum, i.e. $\int_{-F_0/2}^{F_0/2} |G(f_k)|^2 df_k = 1$. Furthermore the Rayleigh distributed spectral magnitude of the noise component of the excitation is zero outside $[-\frac{F_0}{2}, \frac{F_0}{2}]$ and has unity power,

i.e. $\int_{-F_0/2}^{F_0/2} |R(f_k)|^2 df_k = 1$. Fig. 5.5 shows the excitation amplitude of a clean voiced speech segment, the reconstructed excitation amplitude using HNM and the extracted harmonicity values. The synthesized amplitude spectrum of the clean segment using the HNM reconstructed excitation is shown in Fig. 5.6.

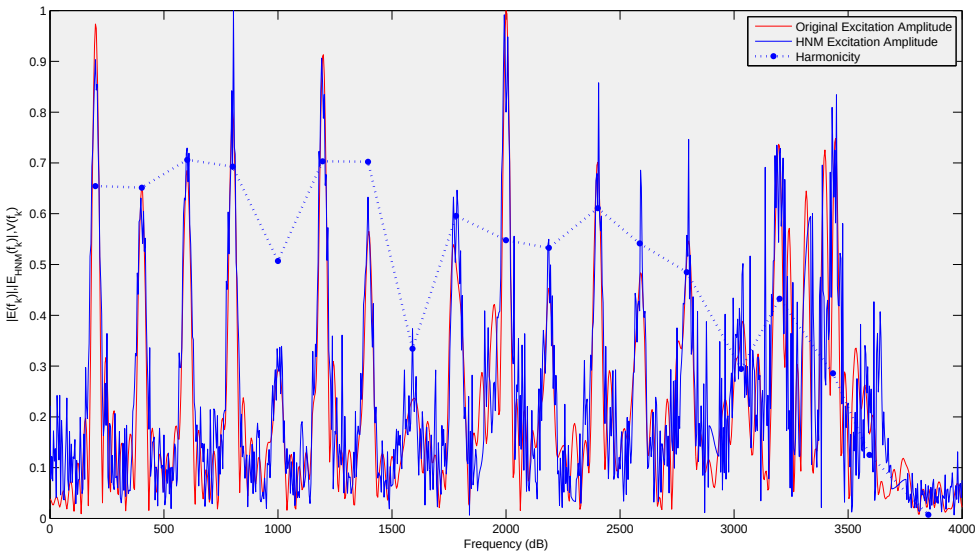

Figure 5.5: Excitation amplitude of a voiced speech frame, the reconstructed excitation amplitude using HNM and the extracted harmonicity values.

5.2.2 Codebook Mapping Estimation

In order to reconstruct the excitation spectrum, a weighted codebook mapping method is implemented similar with the one proposed in [44].

Training of Codebooks

Two different codebooks were trained. One on energy-normalized harmonic amplitude values of excitation of clean speech and the other one on the harmonicity degree vectors of the harmonic subbands. The harmonic amplitudes are normalized as

$$\mathbf{B} = \frac{1}{\sqrt{\sum_{h=1}^{N} A(f_h)^2}} \mathbf{A} \tag{5.9}$$

where $\mathbf{A} = [A(f_1), \ldots, A(f_N)]$ and $\mathbf{B} = [B(f_1), \ldots, B(f_N)]$ are the harmonic amplitudes and the normalized harmonic amplitudes, respectively of an excitation of a speech segment. The rationale behind the normalization of the harmonic amplitudes is that the codebook becomes energy independent while preserving the spectrum's shape. The number of harmonics in each

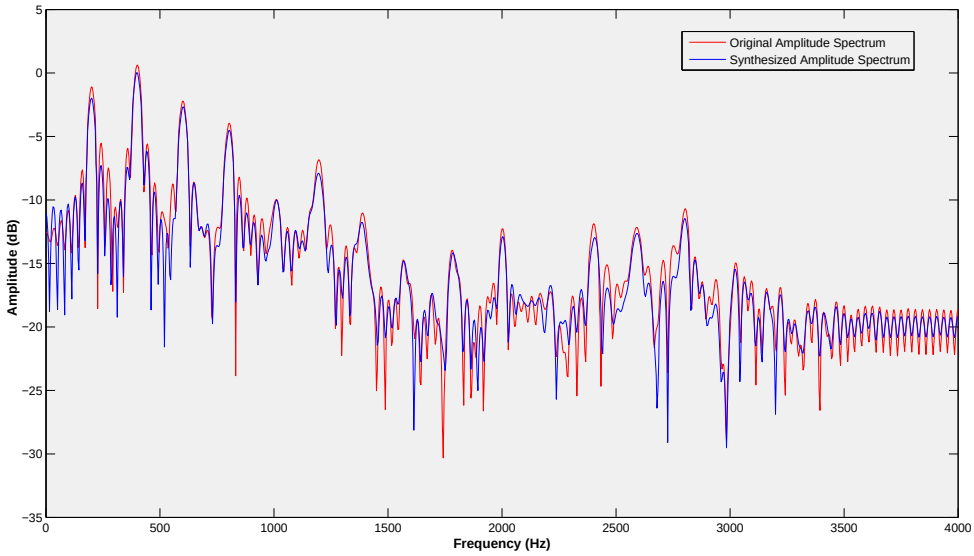

Figure 5.6: The synthesized amplitude spectrum using for synthesis the reconstructed HNM excitation amplitude of Fig. 5.5.

frame depends on the fundamental frequency F_0 and on the sampling frequency F_S and hence is variable. A fixed number of N harmonics were used for training the codebooks. If a segment has more than N harmonics, i.e. $\left\lfloor \frac{\frac{F_S}{2}-\frac{F_0}{2}}{F_0} \right\rfloor > N$, the harmonic amplitude and harmonicity training vectors are truncated to N. On the other hand, if a frame has less than N harmonics the respective training vectors are padded with zeros. This is justified by the fact that during the codebook mapping process, only the existing harmonics are taken into account.

In our approach two 10 bits (1024 codewords) were constructed. One on energy-normalized harmonic amplitude values of excitation of clean speech segments and the other one on the harmonicity degree vectors of the harmonic subbands. This size was chosen since on average it provides a good trade off between quantization errors and memory usage. The dimension of the codebook, i.e. the maximum number of harmonic frequencies, was $N = 40$ [44]. For training the codebooks we used the entire TIMIT-TRAIN database [16], which consist of 4620 clean speech sentences. The sampling frequency was $F_s = 8000\,Hz$ and the speech signals was limited to telephone bandwidth $(300 - 3400\,Hz)$. We used rectangular windowed overlapping frames of $32\,ms$. Beginning and trailing silences and frames of a sentence with energy lower than $40\,dB$ of the maximum frame energy of the sentence were excluded from the training phase. These resulted in around 750000 training vectors. The Linde, Buzo, Gray (LBG) vector quantization algorithm was used for training the codebook [5].

Codebook Mapping Algorithm

In applying the codebook mapping algorithm, first the HNM parameters of the " enhanced excitation" of enhanced frame i are extracted. The weighted distance between the harmonic amplitudes of the "enhanced excitation" and the codewords from the harmonic amplitude

codebook are estimated as

$$D_m^i = \sum_{h=1}^{N_i} (W_h^i(B_{NR}(f_h, i) - C(f_h, m)))^2 \tag{5.10}$$

where $\mathbf{B_{NR}}^i = [B_{NR}(f_1, i), \ldots, B_{NR}(f_{N_i}, i)]$ is the energy normalized amplitude vector obtained from the "enhanced excitation", $\mathbf{C}_m = [C(f_1, m), \ldots, C(f_N, m)]$ is the m^{th} entry of the amplitude codebook and W_h^i is the weight of the h^{th} harmonic. The number of harmonic frequencies N_i that are taken into account at this stage is the minimum of the dimension of the codebook N and the number of existing harmonics in the frame N_H, i.e. $N_i = min(N_H, N)$. Note that if $N_H > N$ the harmonics above the limit will not be processed by the codebook mapping algorithm and will be replaced only by the ones of the "enhanced excitation".

The value W_h^i is between 0 and 1 and is used as a measure of the reliability of the "enhanced excitation". Any of the factors of Eq.4.6 can be used for obtaining these weights, according to the following formula

$$W_h^i = \frac{w_h(i) - \min\limits_h\{w_h(i)\}|_{i=const}}{\max\limits_h\{w_h(i)\}|_{i=const} - \min\limits_h\{w_h(i)\}|_{i=const}} \tag{5.11}$$

where $w_h(i) = \sum\limits_{f_k=f_h-F_0/2}^{f_h+F_0/2} w(f_k, i)$ is the weighting factor of h^{th} subband and $w(f_k, i)$ is the weighting factor given by Eq. 4.6. Note that Eq. 5.11 rescales the weights to range $[0, 1]$ on a frame-by-frame basis.

An estimate of the codebook amplitude vector is obtained from the $L = 3$ (according to [44]) entries of the codebook with the lowest distances from $\mathbf{B_{NR}}^i$, as

$$\mathbf{B_{CB}}^i = \sum_{j=1}^{L} q_j \mathbf{C}_j \tag{5.12}$$

where q_j is the weight of the codeword $\mathbf{C}_j$ and is proportional to the reciprocal of the distance D_j^i of the codeword $\mathbf{C}_j$ from the "enhanced excitation" amplitude vector $\mathbf{B_{NR}}^i$:

$$q_j^i = \frac{1}{SD_j^i \sum\limits_L \frac{1}{SD_j^i}} \tag{5.13}$$

The resulting energy-normalized amplitude codebook vector $\mathbf{B_{CB}}^i$ needs to be denormalized before combining it with the noise reduced vector. This is

$$\mathbf{A_{CB}}^i = \alpha_{dn} \mathbf{B_{CB}}^i \tag{5.14}$$

where

$$\begin{aligned} \alpha_{dn} &= \arg\min_{\alpha}\{\sum_{h=1}^{N_i} (W_h^i(A_{NR}(f_h, i) - \alpha B_{CB}(f_h, i)))^2\} \\ &= \frac{\sum\limits_{h=1}^{N_i} (W_h^i)^2 A_{NR}(f_h, i) B_{CB}(f_h, i)}{\sum\limits_{h=1}^{N_i} (W_h^i B_{CB}(f_h, i))^2} \end{aligned} \tag{5.15}$$

The denormalized $\mathbf{A_{CB}}^i$ vector is then weighted-averaged with the noise reduced vector $\mathbf{A_{NR}}^i$ in a way that elements with higher weights W_h^i are less affected than those with lower weights:

$$\tilde{A}(f_h, i) = W_h^i A_{NR}(f_h, i) + (1 - W_h^i) A_{CB}(f_h, i) \tag{5.16}$$

The harmonicities of the harmonics $\tilde{V}(f_h, i)$ are obtained from the harmonicity codebook in the same fashion but without the normalization step. The reconstructed HNM excitation amplitude $\left|\tilde{E}_{HNM}(f_k, i)\right|$ is computed by substituting $\tilde{A}(f_h, i)$ and $\tilde{V}(f_h, i)$ to Eq. 5.4. Note that the unprocessed "enhanced" excitation phase is used for synthesizing the "HNM excitation's"[2] complex spectrum. Fig. 5.7 illustrates the amplitude excitation of a voiced speech segment (red solid line). The clean speech was contaminated by computer generated bandlimited (1000 Hz - 2000 Hz) stationary white Gaussian noise to make the restoration process clearer. From the "enhanced excitation" (green solid line) it is evident that the harmonic structure of the spectrum is distorted and suppressed in the frequency bins where the noise is dominant. The effect of applying the above described codebook mapping restoration method is shown with the blue solid line. The harmonic structure is restored in terms of amplitude and shape at most of the harmonics. For the evaluation of the HNM restoration scheme of the excitation see section 6.3.

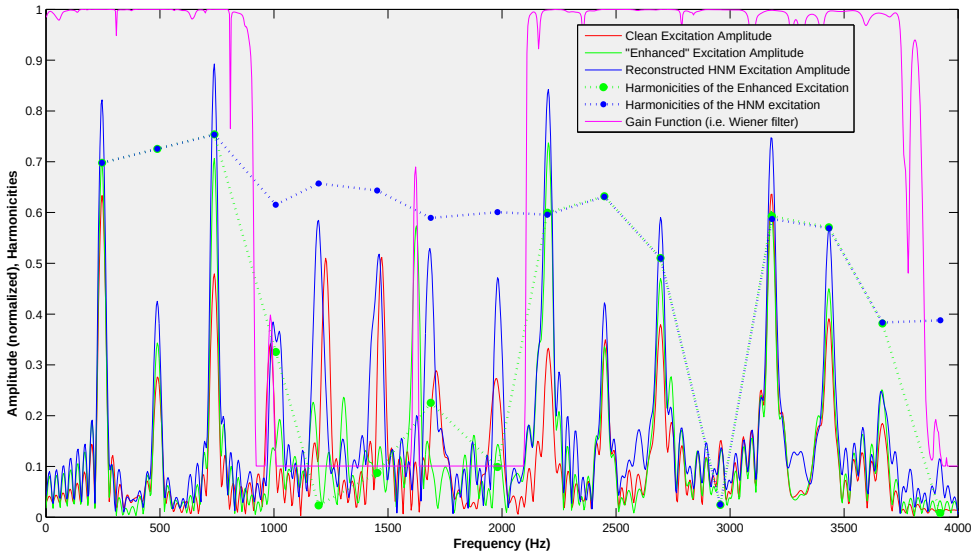

Figure 5.7: Illustration of the performance of the HNM codebook mapping restoration approach for the excitation.

[2]We use the term "HNM excitation" to denote the excitation obtained from the post-processing excitation restoration method presented in this section

Chapter 6

Results and Discussion

In this chapter we evaluate the performance of the post-processing speech enhancement method described in the previous chapters. First we describe the objective and subjective measures of speech quality that we use and the experimental setup. This is followed by a description of a number of experiments to evaluate the performance of different aspects of the post-processing system. Experiments were conducted to evaluate the performance of the spectral envelope estimation approach, the excitation estimation approaches and the overall post-processing speech enhancement system.

6.1 Quality Measures & Experimental Setup

The objective measures of speech quality used in this thesis were the segmental signal-to-noise ratio (SSNR), computed as the average of the SNR for each segment in the utterance:

$$SNR_i = 10\log_{10}\left(\frac{\sum_{n=0}^{K-1} x_i(n)^2}{\sum_{n=0}^{K-1} (x_i(n) - \hat{x}_i(n))^2}\right) \tag{6.1}$$

and the Perceptual Evaluation of Speech Quality (PESQ) [35], an ITU recommendation that has been reported to have a high correlation to subjective quality. The test set for computing the objectives measures consisted of ten speech utterances (5 male and 5 female) randomly chosen from the TIMIT core test set. Experiments were conducted for noisy speech at 0, 5, 10 and 15 dB global SNR for car noise and white Gaussian noise.

Since subjective listening tests are recognized as being the most reliable way of measuring the quality of speech, we evaluate the performance of the post-processing speech enhancement method by contacting a series of subjective listening tests using the method called "multi stimulus test with hidden reference and anchor" (MUSHRA) [3]. The subjects were asked to score the stimuli according to the scale shown in Fig 6.1. We used the clean speech signal as a reference signal (which was also used as a hidden stimuli) and a noisy speech as a hidden anchor [1]. The subject could switch between the reference signal and any of the stimuli. A total of 18 listeners (10 males and 8 females) participated in the test. The subjects were between the age of 20 and 50 years old. The listening test material was rendered to them through

[1] The anchor is specified analogously for each of the experiment

high-quality headphones (Sennheizer HD 600). The test set for computing the subjective measures consisted of four speech utterances (2 male and 2 female) randomly chosen from the TIMIT core test set. Experiments were conducted for noisy speech at 0 and 10 dB global SNR for car noise and white Gaussian noise.

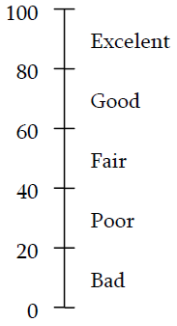

Figure 6.1: Mean opinion scores (MOS) scale used in the MUSHRA test

For the experiments we used a frame length of 256 samples with an overlap of 50%. The size of the fast fourier transform (FFT) was 2048. The modified decision directed approach (Eq. 2.18) with a smoothing factor $\alpha = 0.98$ was used to estimate the a priori SNR. The Wiener estimator (Eq. 2.10) was used as the conventional speech enhancement method. The gain function, i.e. Wiener estimator, was chosen as the weighting factor (Eq. 4.6).

6.2 Evaluation of the spectral envelope estimation approach

In this section we compare the envelope obtained from the post-processing speech spectral envelope restoration method presented in chapter 4 ("post envelope") with the "enhanced envelope". In other words, we investigate if the codebook mapping approach enhances the spectral envelope obtained from a conventional speech enhancement method (see the problem definition on page 4). To do so the "clean excitation"[2] is used for resynthesis (see Appendix A). The noisy signal is synthesized by passing the "clean excitation" through the "noisy envelope", the enhanced signal is synthesized by passing the "clean excitation" through the "enhanced envelope" and the post-processed signal is synthesized by passing the "clean excitation" through the "post envelope". This allows us to leave out the influence of the excitation on the quality of the signals and to focus only on the performance of the post-processed spectral envelope restoration scheme.

Table 6.1 shows the objective quality measures for the "noisy envelope", the "enhanced envelope" and the "post envelope". It can be seen that for all SNR values and for both noise types the post-processing speech spectral envelope restoration approach only slightly improves the performance of the conventional speech enhancement scheme. In terms of SSNR the improvement seems to depend on the SNR (the higher the SNR the "greater" the improvement). Note that for higher than 10dB SNR values the performance of both the "enhanced envelope" and the "post envelope" degrades (in terms of SSNR) comparing to the "noisy envelope". Since the respective PESQ scores are counter to that, we interpret the higher SSNR

[2]We use the term "clean excitation" to denote the excitation obtained from the clean speech signal after LPA

values for relatively high SNRs as an indication that the quality of the "noisy envelope" is very close to the quality of the "enhanced envelope".

		SSNR			PESQ		
SNR	Noise	N.Env.	E.Env.	P.Env.	N.Env.	E.Env.	P.Env.
	White	1.06	2.27	2.36	1.50	1.96	1.96
0	Car	0.71	1.87	1.93	2.13	2.54	2.59
	White	2.5	3.07	3.19	1.77	2.68	2.75
5	Car	2.93	3.09	3.21	2.53	2.9	2.96
	White	4.3	3.83	3.98	2.19	3.14	3.24
10	Car	5.48	4.33	4.63	2.96	3.35	3.38
	White	6.62	4.7	4.89	2.72	3.55	3.62
15	Car	9.27	6.49	6.76	3.33	3.72	3.75

Table 6.1: SSNR and PESQ values averaged over 10 utterances at 0, 5, 10 and 15dB global SNR for the "noisy envelope" (N.Env), "enhanced envelope" (E.Env.) and "post envelope" (P.Env).

The aforementioned claims are also confirmed from the results obtained from the subjective listening tests shown in Fig 6.2 and 6.3. Note the relatively high MOS (Fig 6.2) that the "noisy envelope" gets at 10dB SNR for both noise types (especially for the low pass car noise). Furthermore, in almost all cases (except from the white Gaussian noise at 0dB SNR - extremely noisy condition extended across the whole frequency band) 75% of the "post envelope" samples overperform the respective "enhanced envelope" samples. Despite of that percentage, the box plots in Fig 6.2 show no significant improvement, since the notches always overlap. We deduce the same in case we substract the MOS of the "noisy envelope" from the MOS of the "enhanced envelope" and from the MOS of the "post envelope'.'

By comparing the quality (subjective & objective) measures with the Fig 4.5 and 4.6 it can be seen that the post-processing method for estimating the spectral envelope has a "significant" advantage only for noise types that are bandlimited (e.g. filtered Gaussian, siren noise, etc.), since then there is enough information in the weighting factor (Eq. 4.6) to get a good match (in SD terms - Eq. 4.5) from the LSF codebook. Furthermore, it worth to mention that we conducted informal listening tests with bigger codebook sizes (11 and 15 bits), different LPA orders ($p = 6, 8$ *and* 12), smaller segment size ($20ms$) and 75% overlap between succeding segments, but still no significant improvement in perception was noticed.

6.3 Evaluation of the excitation estimation approaches

In this section we compare the excitation signals obtained from the two post-processing excitation restoration methods (chapter 5) with the "enhanced excitation". In other words, we investigate if those two approaches enhance the excitation obtained from a conventional speech enhancement method (see the problem definition on page 4). To do so the "clean envelope"[3] is used for resynthesis (see Appendix A). The noisy signal is synthesized by passing the "noisy excitation"[4] through the "clean envelope", the enhanced signal is synthesized by

[3]We use the term "clean envelope" to denote the envelope obtained from the clean speech signal after LPA
[4]We use the term "noisy excitation" to denote the excitation obtained from the noisy signal after LPA

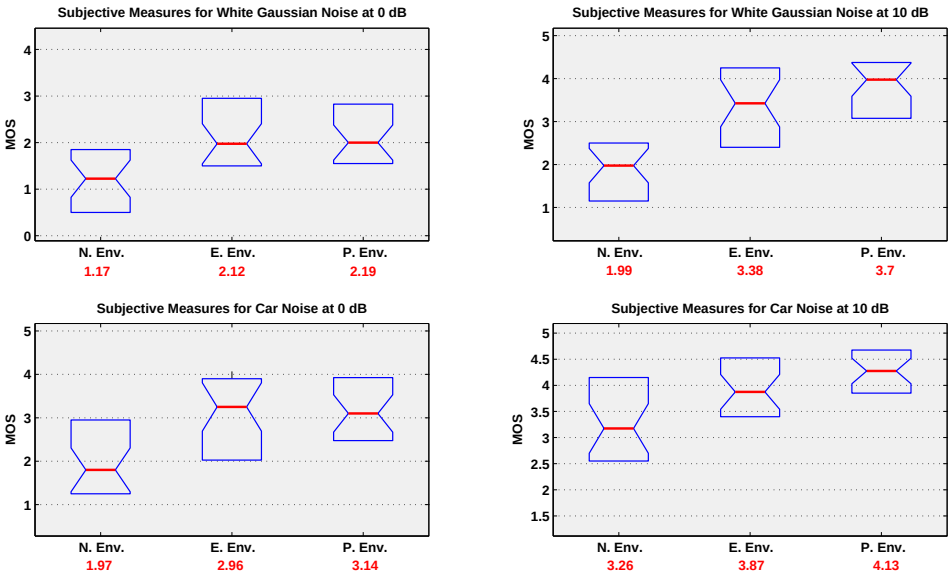

Figure 6.2: Subjectives measures for evaluating the performance of the spectral envelope estimation approach. Each box has horizontal lines at the lower quartile (25% of the samples), median (50% of the samples) and upper (75% of the samples) quartile values. Notches display the variability of the median between samples. The width of the notch is computed so that box plots whose notches do not overlap have different medians at the 5% significance level. The red typed values denote the respective mean values.

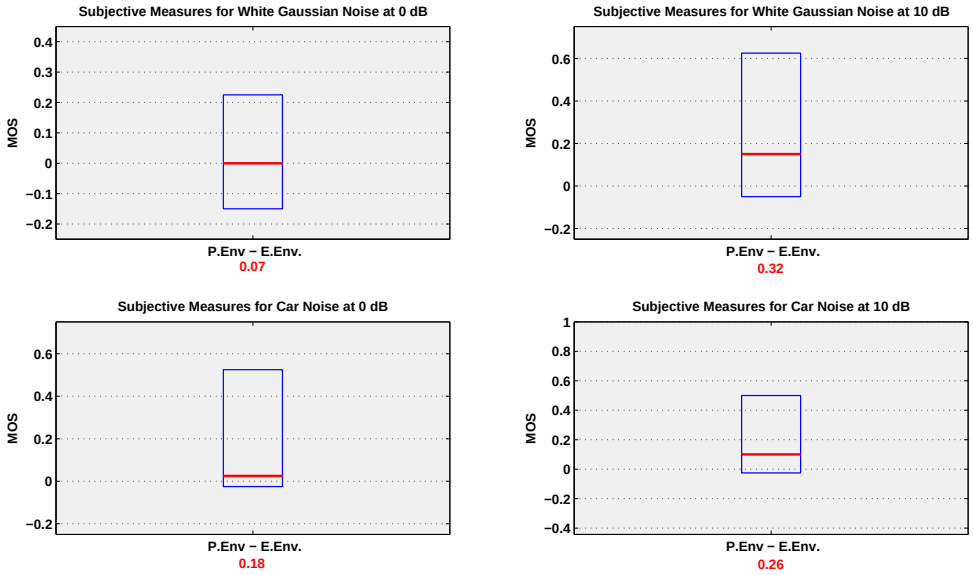

Figure 6.3: Differences between the MOS of the "post envelope" and the "enhanced envelope". Each box has horizontal lines at the lower quartile (25% of the samples), median (50% of the samples) and upper (75% of the samples) quartile values. The red typed values denote the respective mean values.

passing the "enhanced excitation" through the "clean envelope", the HR signal is synthesized by passing the "HR excitation" (Eq. 5.3) through the "clean envelope" and the HNM signal is synthesized by passing the "HNM excitation" (Eq. 5.4) through the "clean envelope". This allows us to focus only on the performance of the excitation estimation approaches.

Table 6.2 shows the objective quality measures for the "noisy excitation", the "enhanced excitation", the "HNM excitation" and the "HR excitation". It can be seen that in general both the post-processing methods perform as well as the conventional speech enhancement method.

From the results obtained from the subjective listening tests (Fig 6.4 and 6.5) several conclusions can be drawn:

- The excitation signal influences the quality of a signal much more than the spectral envelope does. This is deduced from the MOS that the "noisy envelope" (Fig 6.2) and the "noisy excitation" (Fig 6.4) get.

- In extremely noisy conditions (white Gaussian noise at 0dB SNR) both the post-processing methods perform as well as the conventional speech enhancement scheme (Fig. 6.5).

- In medium noise conditions (10dB SNR) both the post-processing methods improve the quality of the speech signals (Fig. 6.5). It can be seen from the box plots in Fig 6.4 that this improvement is not statistically significant since the notches always overlap.

- For lowpass noise types (car noise) and at low SNR values (0 dB) both the post-processing methods have a significant advantage over the conventional speech enhance-

SNR	Noise	SSNR N. Exc.	SSNR E. Exc.	SSNR HNM Exc.	SSNR HR Exc.	PESQ N. Exc.	PESQ E. Exc.	PESQ HNM Exc.	PESQ HR Exc.
0	White	-4.02	-2.4	-1.88	-2.38	1.07	1.96	2	1.95
	Car	-4.05	-0.38	-0.65	-0.31	1.14	1.93	1.97	1.96
5	White	-1.74	-0.88	-0.55	-0.83	1.2	2.2	2.24	2.2
	Car	-1.61	1.41	1.11	1.44	1.4	2.35	2.37	2.4
10	White	0.90	0.53	0.65	0.53	1.47	2.42	2.45	2.4
	Car	1.01	3.28	3.01	3.26	1.74	2.75	2.8	2.77
15	White	3.47	2.79	2.71	2.76	1.8	2.72	2.76	2.72
	Car	3.52	5.64	5.27	5.55	2.13	3.18	3.18	3.16

Table 6.2: SSNR and PESQ values averaged over 10 utterances at 0, 5, 10 and 15dB global SNR for the "noisy excitation" (N. Exc.), "enhanced excitation" (E. Exc.), "HNM excitation" (HNM Exc.) and "HR excitation" (HR Exc.).

ment scheme.

- In all cases the HNM method performs better than the HR.
- There is no objective measure that can predict 100% accurate the subjective listening tests.

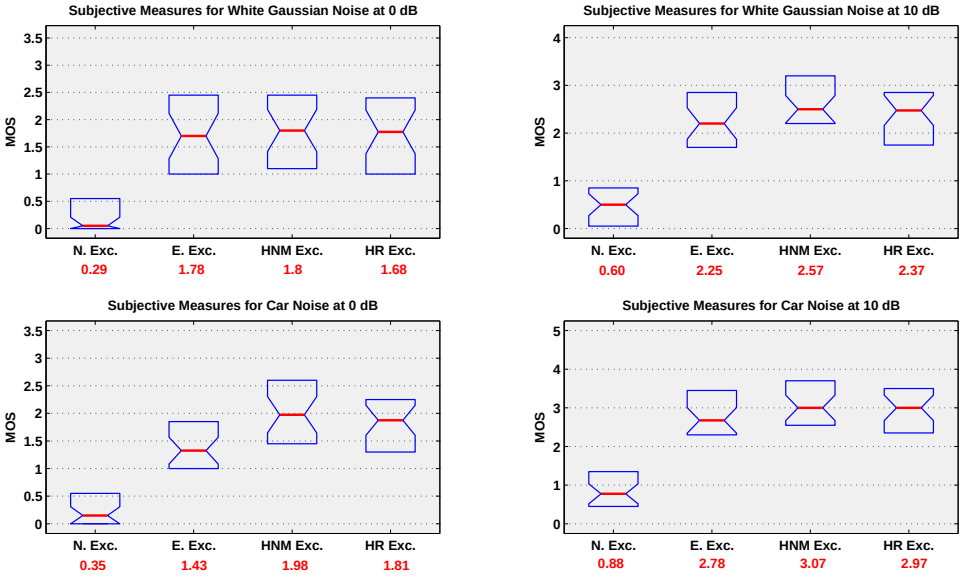

Figure 6.4: Subjectives measures for evaluating the performance of the excitation estimation approaches. Each box has horizontal lines at the lower quartile (25% of the samples), median (50% of the samples) and upper (75% of the samples) quartile values. The width of the notch is computed so that box plots whose notches do not overlap have different medians at the 5% significance level. The red typed values denote the respective mean values.

HR method is a simple way to restore speech excitation harmonics using nonlinearity (see section 3.2.1). It is easier to restore harmonics when only a few are degraded or missing which explains the better behavior for high SNRs, lowpass noise types (Fig 6.5) and bandlimited noise cases (Fig 5.1). HNM method is a quite complicated method to revive severely damaged subbands of speech excitation using a priori knowledge stored in codebooks and a weighted codebook mapping algorithm. Thereafter is natural to perform better than the HR method. Note that white Gaussian noise at 0dB global SNR results in such loss of speech information that even the HNM fails to reveal the degraded harmonics. Errors in extracting the fundamental frequency play also a roll.

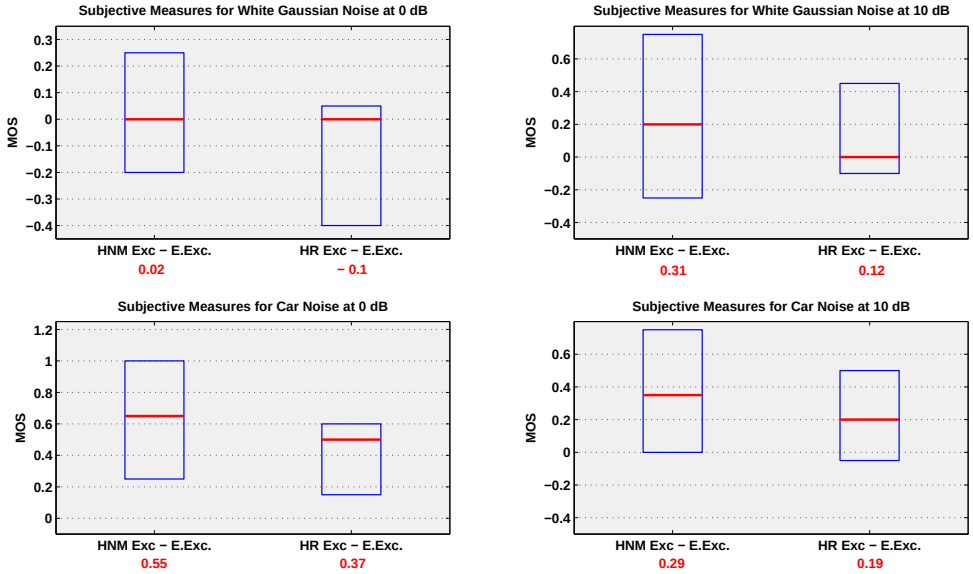

Figure 6.5: Differences between the MOS of the "HNM excitation" and the "enhanced excitation" and between the "HR excitation" and the "enhanced excitation". Each box has horizontal lines at the lower quartile (25% of the samples), median (50% of the samples) and upper(75% of the samples) quartile values. The red typed values denote the respective mean values.

6.4 Evaluation of the overall post processing speech enhancement scheme

In this section we compare the overall performance of the post-processing speech enhancement method presented in this thesis with a conventional speech enhancement method. The post-processing speech spectral envelope restoration method (chapter 4) and the two post-processing excitation restoration methods (chapter 5) yield to two add-on post-processing systems. The HR signal, synthesized by passing the "HR excitation" (Eq. 5.3) through the "post envelope" and the HNM signal, synthesized by passing the "HNM excitation" (Eq. 5.4) through the "post envelope". The enhanced signal is the output of a conventional speech enhancement method (Wiener filter).

Table 6.3 shows the objective quality measures for the noisy, the enhanced, the HNM and the HR signals. It can be seen that in general the post-processing methods perform as well as the conventional speech enhancement method. This is also expected based on the objective measures of the individual parts of the post-processing speech enhancement system (Table 6.1 and 6.2).

		SSNR				PESQ			
SNR	Noise	Noisy	Wiener	HNM	HR	Noisy	Wiener	HNM	HR
0	White	-4.26	0.61	0.27	0.45	1.07	1.39	1.4	1.38
	Car	-4.42	0.16	-0.33	0.1	1.15	1.54	1.59	1.6
5	White	-1.63	3.16	2.73	2.93	1.24	1.83	1.88	1.85
	Car	-1.52	2.68	1.95	2.51	1.44	2.02	2.02	2.06
10	White	1.91	5.76	5.2	5.44	1.54	2.2	2.23	2.26
	Car	1.91	5.49	4.72	5.25	1.82	2.45	2.49	2.5
15	White	5.80	8.38	7.61	7.97	1.91	2.61	2.65	2.64
	Car	5.85	8.57	7.55	8.24	2.26	2.92	2.93	2.94

Table 6.3: SSNR and PESQ values averaged over 10 utterances at 0, 5, 10 and 15dB global SNR for the noisy, the enhanced, the HNM and the HR signals

From the results obtained from the subjective listening tests (Fig 6.7 and 6.5) several conclusions can be drawn:

- In white noise conditions both the post-processing methods perform as well as the wiener filter (Fig. 6.7).
- For lowpass noise types (car noise) the HR method improves the quality of the speech signals (Fig. 6.7). It can be seen from the box plots in Fig 6.6 that this improvement is not statistically significant.
- For lowpass noise types (car noise) the HNM method improves the quality of the speech signals. It can be seen from the box plots in Fig 6.6 that this improvement is statistically significant only for 0db SNR.
- In all cases the HNM method performs slightly better than the HR.
- There are some inconsistencies between the subjective measures of the overall system and those of the individual parts, e.g. HR method at 0dB. Further evaluation of the processed speech signals is needed with a method specifically designed for speech enhancement algorithms evaluation (ITU-T standard (P.835) [1]).

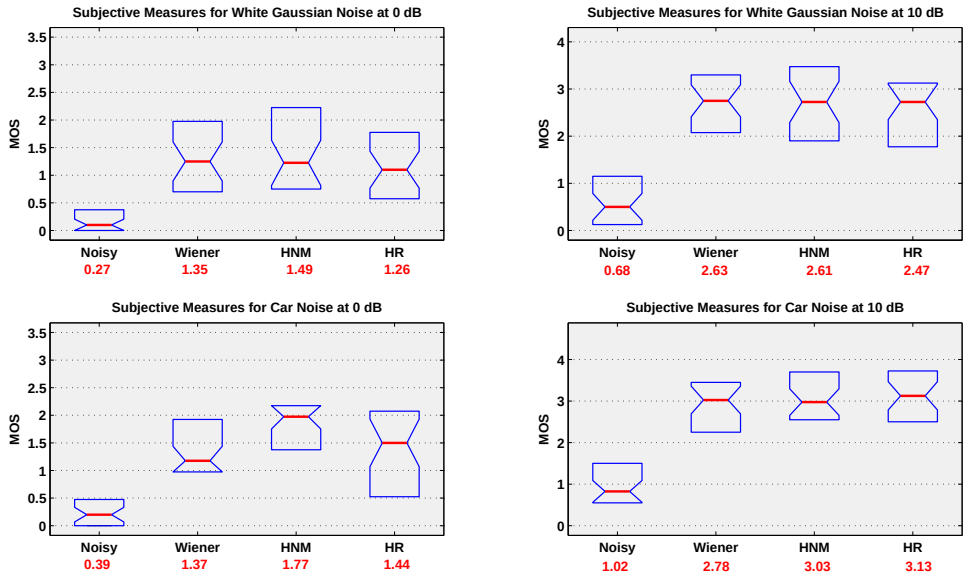

Figure 6.6: Subjectives measures for evaluating the overall performance of the speech post-processing speech enhancement system. Each box has horizontal lines at the lower quartile (25% of the samples), median (50% of the samples) and upper (75% of the samples) quartile values. The width of the notch is computed so that box plots whose notches do not overlap have different medians at the 5% significance level. The red typed values denote the respective mean values.

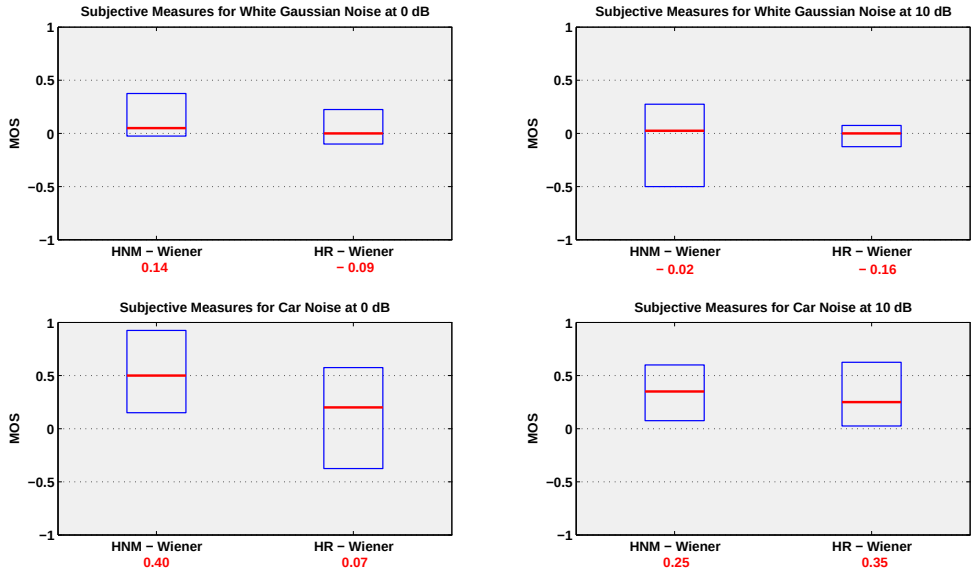

Figure 6.7: Differences between the MOS of the HNM signal and the wiener filter and between the HR signal and the wiener filter. Each box has horizontal lines at the lower quartile (25% of the samples), median (50% of the samples) and upper (75% of the samples) quartile values. The red typed values denote the respective mean values.

Chapter 7

Conclusions and Future Work

The goal of this thesis was to investigate the possibility of enhancing the performance of a conventional speech enhancement system by applying post-processing restoration modules. Conventional speech enhancement systems result in some loss of speech information, the severity of which depends on the SNR and on the system. We addressed the important issue of restoration of those speech parts that are lost due to noise or suppressed by the speech enhancement scheme. In order to do so we model the speech production process with LPA. This yielded to a two step problem: estimation of the spectral envelope and estimation of the excitation signal.

A scheme based on a weighted codebook mapping technique was proposed for enhancing the speech spectral envelope. While codebook mapping works well in relevant to speech enhancement applications, such as bandwidth extension, it fails to improve significantly the performance of a conventional speech enhancement scheme for real life noise types.

Two different methods were considered for enhancing the excitation signal. The first one (HR) uses nonlinearity to preserve speech excitation harmonics that have been degraded by the conventional noise reduction scheme, while the second one (HNM) uses a weighted codebook mapping technique to restore the lost or suppressed excitation harmonics. For lowpass noise types, such as car noise, at low SNR values both the post-processing methods have a significant advantage over the conventional speech enhancement scheme. In white noise conditions, though, both the post-processing methods fail to improve significantly the performance of a conventional speech enhancement scheme.

The proposed system when used as an add-on post-processing module with an existing noise reduction scheme improves the quality of the speech signal only for lowpass noise types at low SNR values.

Since the excitation signal influences the quality of a processed signal much more than the spectral envelope does, future work should focus on deriving accurate methods for enhancing the excitation signal. The recently proposed harmonicity measurements [43] for assessing the speech quality could be incorporated into the HR post-processing system. In case of the HNM model, accurate pitch tracking should be guaranteed. Optimization of the parameter L could probably results in some improvement in terms of speech quality. Different sizes and dimensions of the harmonic amplitude and the harmonicity codebooks should also be tried.

Appendix A

Linear Prediction Analysis of Speech

Linear prediction Analysis (LPA) is widely used in many speech applications (noise reduction, speech compression, speech coding, speech recognition etc.). This Appendix provides the most relevant aspects of LPA for the purposes of this thesis.

A.1 Autoregressive (AR) model of speech

Speech production process is well modeled with LPA. Indeed, it is well recognized that a speech signal can be written in the following form [30]

$$x(n) = \sum_{l=1}^{p} a_l x(n-l) + Ge(n) \tag{A.1}$$

where n is the discrete-time index, p represents the number of coefficients in the model, $a_l,\ l = 1, \ldots, p$, are defined as the linear prediction coefficients (LPC), G is the gain (also referred to as variance) of the *excitation*, and $e(n)$ is the *excitation* signal, which can be either a quasi-periodic train of impulses or a random noise source (also a combination of both signals for voiced fricatives such as /z/). The periodic source produces voiced sounds such as vowels and nasals (e.g. /I/,/m/), and the noise source produces unvoiced sounds such as the unvoiced fricatives(e.g /s/).

Due to non-stationarity of the vocal track and the excitation [8] the LPA is performed on a frame-by-frame basis, i.e. model parameters are estimated based on a short segment of speech $\mathbf{x}_i$. Note that we dropped the frame index i for notational convenience. Eq. A.1 can be rewritten in the frequency domain, by applying the z-transform. If $H(z)$ is the transfer function of the model of human vocal track (spectral envelope), we have:

$$H(z) = \frac{X(z)}{E(z)} = \frac{G}{1 - \sum_{l=1}^{p} a_l z^{-l}} = \frac{G}{A(z)} \tag{A.2}$$

which is an all-pole transfer function. For each frame the model parameters are the vector of LPC $\mathbf{a} = [1\, a_1\, \ldots a_p]$ and the variance of the *excitation* signal G. In Fig. A.1 the block diagram of the model is depicted.

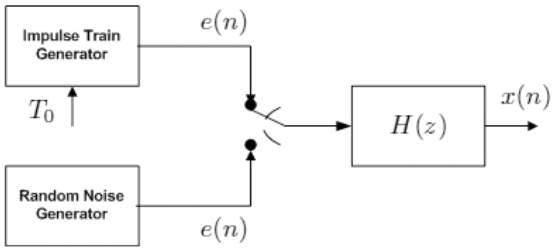

Figure A.1: Simplified source-filter model of speech.

A.2 Forward Linear Prediction

The basic idea behind LPA is that each speech sample $x(n)$ is approximated as a linear combination of past samples. A linear predictor of order p is defined by

$$\hat{x}(n) = \sum_{l=1}^{p} \alpha_l x(n-l) \tag{A.3}$$

The sequence of $\hat{x}(n)$ is the prediction of $x(n)$ by the sum of p past samples. The system function associated with p_{th} order predictor is a finite-length impulse response (FIR) filter of length p, given by

$$P(z) = \sum_{l=1}^{p} \alpha_l z^{-l} \tag{A.4}$$

The *prediction error* sequence is given by

$$\begin{aligned} \epsilon(n) &= x(n) - \hat{x}(n) \\ &= x(n) - \sum_{l=1}^{p} \alpha_l x(n-l) \end{aligned} \tag{A.5}$$

and the associated *prediction error* filter is defined as:

$$A(z) = 1 - \sum_{l=1}^{p} \alpha_l z^{-l} = 1 - P(z) \tag{A.6}$$

which is illustrated in Fig. A.2(a)

Suppose that $x(n)$ is the output of an all-pole system as in Eq. A.1. Then if the predictor coefficients equal the model coefficients, i.e. $\alpha_k = a_k$ we write the *prediction error* as

$$\begin{aligned} \epsilon(n) &= x(n) - \sum_{l=1}^{p} \alpha_l x(n-l) \\ &= \sum_{l=1}^{p} a_l x(n-l) + Ge(n) - \sum_{l=1}^{p} \alpha_l x(n-l) \\ &= Ge(n) \end{aligned} \tag{A.7}$$

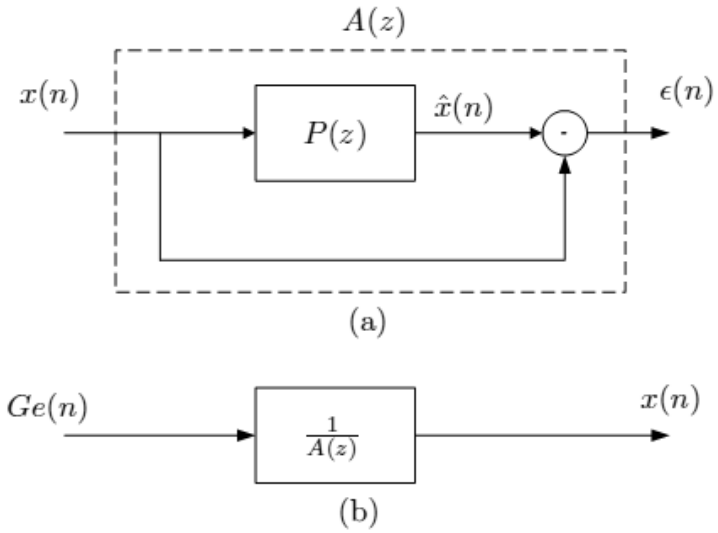

Figure A.2: Filtering view of linear prediction: (a) prediction error filter $A(z)$; (b) AR model of speech.

and thus the $Ge(n)$ can be recovered by passing $x(n)$ through $A(z)$. For this reason the $A(z)$ is sometimes called *inverse filter*. Correspondingly, when we pass $Ge(n)$ through $\frac{1}{A(z)}$, we obtain $x(n)$ as shown in Fig. A.2(b)

A.3 Error minimization

We would like to find the optimal linear predictor in terms of MSE. For that we define the error criterion:

$$J = E\{\epsilon(n)^2\} = E\{(x(n) - \hat{x}(n))^2\} \tag{A.8}$$

where $E\{.\}$ is the statistical expectation operator. The goal is to minimize the above criterion

$$\min_{\alpha_l} J(\alpha_l) \tag{A.9}$$

This is done by calculating the derivatives of $J(\alpha_l)$ with respect to α_l

$$\frac{\partial J(\alpha_l)}{\partial \alpha_l}, \quad l = 1, \ldots, p \tag{A.10}$$

From Eq.A.8 and Eq.A.10 we obtain the following set of normal equations

$$\begin{bmatrix} r(0) & r(1) & \ldots & r(p-1) \\ r(1) & r(0) & \ldots & r(p-2) \\ \vdots & \vdots & \vdots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix} \tag{A.11}$$

where $r(\tau) = E\{x(n)x(n+\tau)\}$.

The above normal equations involve expected values of random variables $r(\tau)$. In practice, we have a short segment of speech $\mathbf{x}_i$ from which $r(\tau)$ must be estimated. Depending on how the estimator of $r(\tau)$ is formed, we end up with either the covariance or autocorrelation method [30]. The autocorrelation method is preferred because it guarantees stable synthesis filters $H(z)$ and there is a computational efficient way to calculate the optimal α_l.

We still need to compute gain G in the model (Eq. A.2) It can be shown that if the gain G is chosen as the square root of the minimum square error, i.e. $G = \sqrt{\epsilon_{min}}$, then the signal energy is not altered on the average (energy conservation) [30]. Finally, the order p of the LPA is estimated from knowledge of speech production. In this thesis and for sampling frequency $F_s = 8\ kHz$ we used $p = 10$ [23]. Fig A.3 and Fig A.4 show a voiced and an unvoiced example of LPA, respectively.

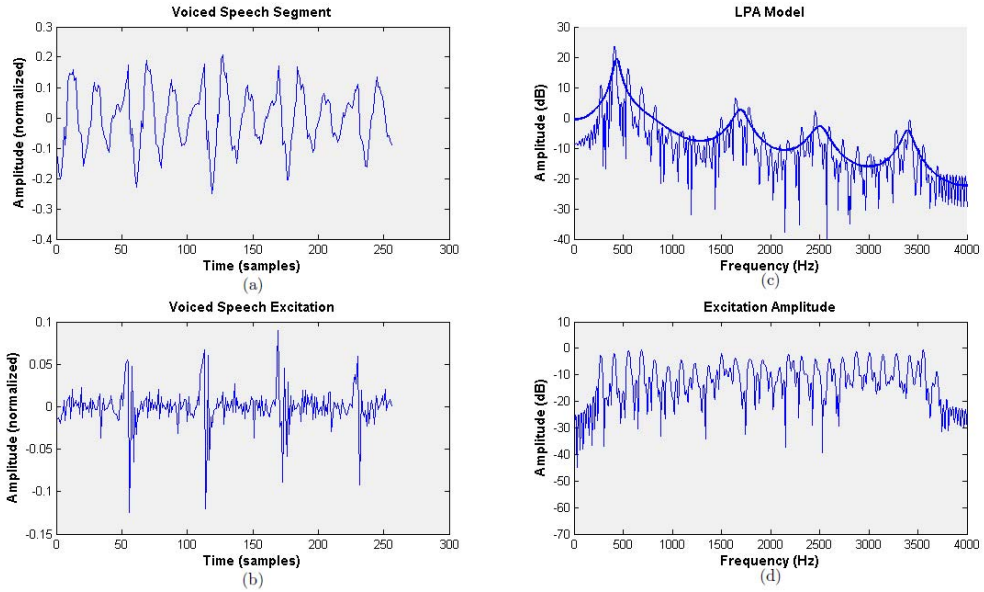

Figure A.3: LPA voiced segment example, where (a) shows a voiced speech segment, (b) shows the excitation of (a), (c) shows the amplitude spectrum and the envelope of (a) and (d) shows the frequency representation of (b).

A.4 Bandwidth widening

LPA has problems in estimating accurately the spectral envelope $H(z)$ for high pitch voiced sounds. The spectral information about a periodic signal is contained only at harmonics. For high pitched voices (i.e. female speaker), the harmonic spacing is too wide to provide an adequate sampling of the spectral envelope. For this reason, LPA does not provide accurate estimation of spectral envelope for female speakers. Such inaccurate estimation occurs mainly in formant bandwidth. This might result in unnatural (metallic) synthesized speech.

The most common procedure to overcome this problem is bandwidth widening [23]. In this procedure, each LPC coefficient α_l is multiplied by a factor γ^l, i.e. all α_l are replaced by $\gamma^l \alpha_l$. Such multiplication moves all the poles of $H(z)$ inwards by a factor γ and cause

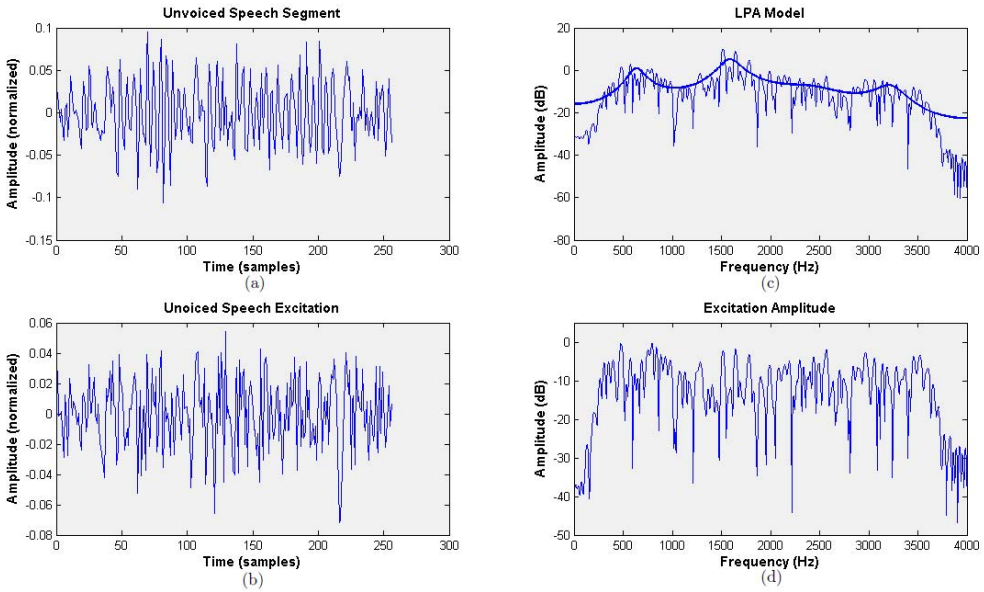

Figure A.4: LPA unvoiced example, where (a) shows an unvoiced speech segment, (b) shows the excitation of (a), (c) shows the amplitude spectrum and the envelope of (a) and (d) shows the frequency representation of (b).

bandwidth expansion for all the poles. In this thesis we used $\gamma = 0.996$ which corresponds to 30 Hz widening.

A.5 Line Spectrum Frequency (LSF)

The LSF representation is an alternative representation of the LPC coefficients with one-to-one correspondence. The LSF representation transforms $A(z)$ (Eq. A.6) to:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)}A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)}A(z^{-1}) \end{aligned} \tag{A.12}$$

The inverse mapping from LSF to LPC is given by

$$A(z) = \frac{1}{2}(P(z) + Q(z)) \tag{A.13}$$

LSFs are defined as the angles ω_i of the roots of $P(z)$ and $Q(z)$ in the upper half of the complex plane. While $A(z)$ has complex roots anywhere within the unit circle, $P(z)$ and $Q(z)$ have the very useful property of only having roots that lie on the unit circle [36]. Further, it can be shown [36] that the roots of $P(z)$ and $Q(z)$ occur in complex conjugated pairs (except for roots at $z = -1$ and $z = 1$) and that the LSFs are interlaced if and only if zeros of $A(z)$ are inside unit-circle. This is called the interlacing or ordering property, which means that if LSFs are ordered then $H(z)$ is stable. For all the above mentioned reasons, in this thesis the LSFs were chosen as the features that will represent the spectral envelope $H(z)$.

Appendix B

Minimum Statistics Noise Tracking

Estimation of the statistics of the background noise is an essential feature of all the single channel speech enhancement methods. If the noise estimate is too low, unnatural residual noise will be perceived in the enhanced signal. On the other hand, if the estimate is too high, the enhanced signal will be muffled. In this appendix we briefly present one of the most frequently used noise estimation schemes based on minimum statistics [31]. Because of its popularity and it's good performance, we chose to use this method for the estimation of the noise PSD.

B.1 Minimum Statistics' Principles

The minimum statistics noise tracking method is based on the observation that even during speech activity the PSD estimate of the noisy speech often decays to that of the noise process. The method rests on the fundamental assumption that during speech pauses or within brief periods in between words the speech energy is close to zero. Thus, by tracking the minimum power per frequency bin k within a finite window, large enough to bridge high power speech segments, the noise floor can be adequate estimated.

In practice we consider a recursively smoothed periodogram to estimate the noisy PSD. This is

$$\hat{P}_{YY}(k,i) = \alpha \hat{P}_{YY}(k,i-1) + (1-\alpha)|Y(k,i)|^2 \tag{B.1}$$

Parameter α is set to 0.85 which value represents a good compromise between smoothing the noise and tracking the speech signal. The noise PSD estimate $\hat{P}_{WW}(k,i)$ is obtained by picking the minimum value per frequency bin within a sliding window of D (empirically often set to 96) consecutive values of $\hat{P}_{YY}(k,i)$.

The above mentioned simplified minimum tracking algorithm provides a rough estimate of the noise power. In order to improve the method the following issues have to be addressed:

- The smoothing with the fixed smoothing parameter widens the peaks of the speech activity of the $\hat{P}_{YY}(k,i)$. This will lead to overestimation of the noise level as the sliding window for the minimum search might slip into broad peaks.
- Since the minimum of a set of noisy PSD values is generally smaller than the mean, the noise estimate is biased towards lower values.
- In case of rapidly increasing noise PSD the, minimum tracking lags behind.

In the next sections the above mentioned shortcomings of the simplified minimum tracking algorithm are discussed.

B.2 Optimal Time Varying Smoothing

The smoothed noisy PSD estimate has to satisfy conflicting requirements. On one hand the variance should be as small as possible requiring α to be closed to 1. On the other hand, $\hat{P}_{YY}(k,i)$ has to track possibly nonstationary noise and to follow the highly nonstationary peaks of the speech. These problems can be addressed only with a time-varying and frequency-depended smoothing parameter $\alpha(k,i)$. By minimizing an appropriate error criterion, we get

$$\alpha_{opt}(k,i) = \frac{1}{1+(\frac{\hat{P}_{YY}(k,i-1)}{P_{WW}(k,i)}-1)^2} \tag{B.2}$$

For analytical derivation of Eq. B.2 see [31]. In practice, we replace the true noise PSD $P_{WW}(k,i)$ by its latest estimate $\hat{P}_{WW}(k,i-1)$ and we limit the smoothing parameter to a maximum value $\alpha_{max} = 0.96$.

In general, the estimated noise PSD $\hat{P}_{WW}(k,i)$ lags behind the true one. As a consequence, the estimated smoothing parameter $\hat{\alpha}(k,i)$ (Eq. B.2) might be too small or too large. Given this uncertainty in the noise PSD, the tracking errors in the $\hat{P}_{YY}(k,i)$ must be monitored. In [31] they propose to monitor these errors by comparing the average, across frequency, noisy PSD estimate of the previous frame $\frac{1}{K}\sum_{k=0}^{K-1}\hat{P}_{YY}(k,i-1)$ to the average periodogram $\frac{1}{K}\sum_{k=0}^{K-1}|Y(k,i)|^2$. The comparison is implemented according to

$$\tilde{\alpha}_c(i) = \frac{1}{1+(\frac{\frac{1}{K}\sum_{k=0}^{K-1}\hat{P}_{YY}(k,i-1)}{\frac{1}{K}\sum_{k=0}^{K-1}|Y(k,i)|^2}-1)^2} \tag{B.3}$$

and the result is limited to values larger than 0.7 and smoothed over time. This is

$$\alpha_c(i) = 0.7\alpha_c(i-1) + 0.3max(\tilde{\alpha}_c(i), 0.7) \tag{B.4}$$

The multiplication of the correction factor $\alpha_c(i)$ with the Eq. B.2 (bounded to values smaller than α_{max}) yield to

$$\hat{\alpha}(k,i) = \frac{\alpha_{max}\alpha_c(i)}{1+(\frac{\hat{P}_{YY}(k,i-1)}{\hat{P}_{WW}(k,i-1)}-1)^2} \tag{B.5}$$

Finally to improve the performance of the noise PSD estimator in high levels of nonstationary noise, a lower limit $\alpha_{min} = 0.3$ is applied to $\hat{\alpha}(k,i)$.

B.3 Bias Compensation

The minimum statistics noise tracking method determines the minimum of the noisy PSD estimate within a finite window of length D. Since often the minimum value is smaller than

the mean value, the minimum noise estimate is biased. In this section we present the formulas for the compensation of that bias. For the analytical derivation of those formulas see [31].

An unbiased estimator of the noise PSD is given by

$$\hat{P}_{WW}(k,i) = B_{min}(D, Q_{eq}(k,i))\hat{P}_{YY}^{min}(k,i) \tag{B.6}$$

where $\hat{P}_{YY}^{min}(k,i)$ is the minimum of D successive noisy PSD estimates $\hat{P}_{YY}(k,i)$, $i \in \{i_{cur}, \ldots, i_{cur} - D + 1\}$, $B_{min}(D, Q_{eq}(k,i))$ is the inverse mean of that minimum, i.e $B_{min}(D, Q_{eq}(k,i)) = \frac{1}{E\{\hat{P}_{YY}^{min}(k,i)\}}$ and $Q_{eq}(k,i)$ the inversed normalized variance of the smoothed noisy PSD estimate, i.e. $Q_{eq}(k,i) = \frac{2P_{WW}^2(k,i)}{var\{\hat{P}_{YY}(k,i)\}}$.

In practice, the inverse mean is approximated by

$$B_{min}(k,i) \approx 1 + (D-1)\frac{2}{\tilde{Q}_{eq}(k,i)}\Gamma(1 + \frac{2}{Q_{eq}(k,i)})^{H(D)} \tag{B.7}$$

where $\tilde{Q}_{eq}(k,i)$ is given by

$$\tilde{Q}_{eq}(k,i) = \frac{Q_{eq}(k,i) - 2M(D)}{1 - M(D)} \tag{B.8}$$

$M(D)$ and $H(D)$ are functions of D obtained by linear interpolation from Table. B.1 and $\Gamma(.)$ the Gamma function.

D	M(D)	H(D)	D	M(D)	H(D)
1	0	0	60	0.841	3.1
2	0.26	0.15	80	0.865	3.38
5	0.48	0.48	120	0.89	4.15
8	0.58	0.78	140	0.9	4.35
10	0.61	0.98	160	0.91	4.25
15	0.668	1.55	180	0.92	3.9
20	0.705	2.0	220	0.93	4.1
30	0.762	2.3	260	0.935	4.7
40	0.8	2.52	300	0.94	5

Table B.1: Parameters for the approximation of the mean of the minimum Eq. B.7 and Eq. B.8. Source [29]

In order to compute the inversed normalized variance of the smoothed noisy PSD estimate, first the variance has to be estimated. In [31] they propose to use a first order smoothing recursion for the approximation of the first moment ($E\{\hat{P}_{YY}(k,i)\}$) and the second moment ($E\{\hat{P}_{YY}^2(k,i)\}$) of $\hat{P}_{YY}(k,i)$. This is

$$\begin{aligned}
\overline{\hat{P}}_{YY}(k,i) &= \beta(k,i)\overline{\hat{P}}_{YY}(k,i-1) + (1-\beta(k,i))\hat{P}_{YY}(k,i) \\
\overline{\hat{P}^2}_{YY}(k,i) &= \beta(k,i)\overline{\hat{P}^2}_{YY}(k,i-1) + (1-\beta(k,i))\hat{P}_{YY}^2(k,i) \\
\widehat{var}\{\hat{P}_{YY}(k,i)\} &= \overline{\hat{P}^2}_{YY}(k,i) - \overline{\hat{P}}_{YY}^2(k,i)
\end{aligned} \tag{B.9}$$

where $\beta(k,i)$ is experimentally set to $\beta(k,i) = min(\hat{\alpha}^2(k,i), 0.8)$. Finally $Q_{eq}(k,i)$ is estimated by

$$Q_{eq}(k,i) \approx \frac{2\hat{P}^2_{WW}(k,i-1)}{\widehat{var}\{\hat{P}_{YY}(k,i)\}} \tag{B.10}$$

and is limited to a maximum of 0.5.

Fig. B.1 shows the clean amplitude spectrum, the noisy amplitude spectrum, the noise amplitude spectrum, the minimum statistics noise PSD estimation, the amplitude spectrum of the enhanced signal and the Wiener estimator (Eq. 2.10) for the frequency $f_k = 492\ Hz$ of a female speech utterance that is degraded with white Gaussian noise at 10 dB overall SNR. Note that the a priori SNR was estimated by the decision directed approach 2.18 and that the Wiener filter is scaled to make the figure clearer.

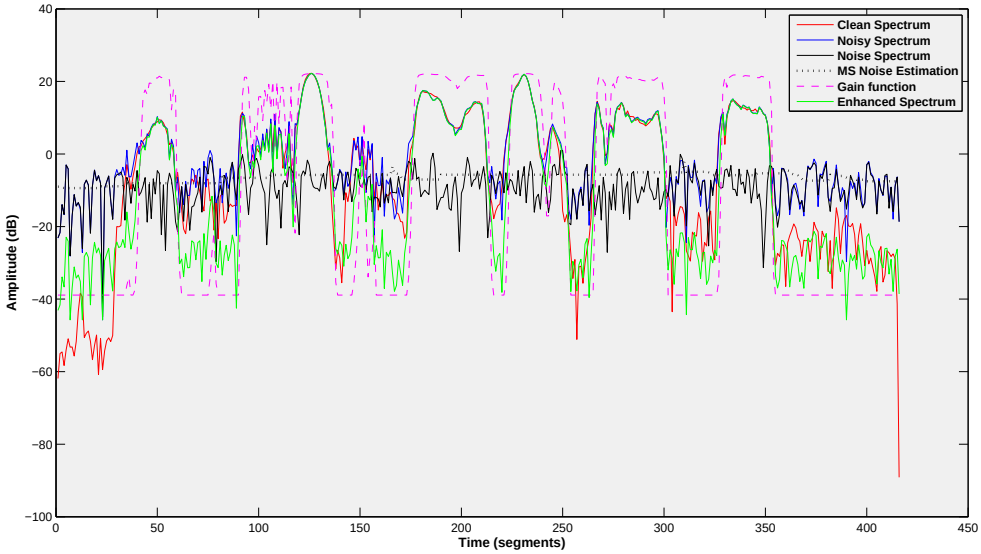

Figure B.1: Minimum statistics noise tracking example

Appendix C

Linde-Buzo-Gray (LBG) Algorithm

Vector Quantization (VQ) is a classical quantization technique that is used in many applications of speech processing, such as speech compression, speech enhancement, speech bandwidth extension, voice recognition and more. In this appendix we describe one of the most popular method of designing codebooks, known as the LBG VQ algorithm [26].

C.1 Vector Quantization

An N level k dimensional quantizer is a mapping q, that assigns to each input vector $\mathbf{x} = (x_0, \ldots, x_{k-1})$ a reproduction vector, $\hat{\mathbf{x}} = q(\mathbf{x})$ drawn from a finite reproduction alphabet (codebook), $\hat{A} = \{\mathbf{y}_i, i = 1, \ldots, N\}$. The quantizer is completely described by the codebook together with the partition, $S = \{S_i, i = 1, \ldots, N\}$, of the input vector space into the sets $S_i = \{\mathbf{x}|q(\mathbf{x}) = \mathbf{y}_i\}$ of input vectors mapping into the i^{th} codeword. The goal of a quantizer is to minimize the expected distortion introduced by the quantization, i.e.

$$D(q) = E\{d(\mathbf{X}, q(\mathbf{X}))\} \tag{C.1}$$

where $\mathbf{X} = (X_0, \ldots, X_{k-1})$ is the real random vector corresponding to the realization $\mathbf{x}$, $E\{.\}$ the expectation operator and $d(\mathbf{x}, \hat{\mathbf{x}})$ the distortion caused by reproducing an input vector $\mathbf{x}$ by a codeword $\hat{\mathbf{x}}$. Many such distortion measures have been proposed in the literature. Most of them, such as the squared error, i.e.

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i=0}^{k-1} |x_i - \hat{x}_i|^2 \tag{C.2}$$

have the property that they depend on the vectors $\mathbf{x}$ and $\hat{\mathbf{x}}$ only through the error vector $\mathbf{x} - \hat{\mathbf{x}}$. Such distortion measures are having the general form $d(\mathbf{x}, \hat{\mathbf{x}}) = L(\mathbf{x} - \hat{\mathbf{x}})$. Distortion measures not having this form but depending on $\mathbf{x}$ and $\hat{\mathbf{x}}$ in a more complicated fashion have also been proposed, such as

$$d(\mathbf{x}, \hat{\mathbf{x}}) = (\mathbf{x} - \hat{\mathbf{x}})\mathbf{R}(\mathbf{x})(\mathbf{x} - \hat{\mathbf{x}})^T \tag{C.3}$$

where $\mathbf{R}(\mathbf{x})$ is a positive definite $k \times k$ symmetric matrix. In this thesis two different distortion measures were used. Specifically, for training the codebook of LSF (section 4.1) the log spectral distortion were used (Eq. 4.1), which results in a distortion measure of the form of Eq. C.3. On the other hand, for training the harmonic amplitude and harmonicity codebooks (section 5.2.2) the square error distortion measure (Eq. C.2) was used.

C.2 LBG Algorithm

The LBG Algorithm for the design of a vector quantizer based on a long training sequence is summarized below

(Input) N: the number of levels (codewords), $\varepsilon \geq 0$: the distortion threshold, $\{\mathbf{x}_j,\, j = 1, \ldots, n-1\}$: a training sequence and $\hat{A}_0$: an initial codebook

(0) Set $m = 0$ and $D_{-1} = \infty$

(1) Given $\hat{A}_m = \{\mathbf{y}_i, i = 1, \ldots, N\}$ find the minimum distortion partition $P(A_m) = \{S_i, i = 1, \ldots, N\}$ of the training sequence according to

$$\mathbf{x}_j \in S_i \; if \; d(\mathbf{x}_j, \mathbf{y}_i) < d(\mathbf{x}_j, \mathbf{y}_l), \forall l \neq i \tag{C.4}$$

(2) Compute the average distortion

$$D_m = \frac{1}{n} \sum_{j=0}^{n-1} \min_{\mathbf{y} \in \hat{A}_m} d(\mathbf{x}_j, \mathbf{y}) \tag{C.5}$$

(3) If $\frac{D_{m-1} - D_m}{D_m} \leq \varepsilon$, halt with $\hat{A}_m$ the final codebook. Otherwise continue

(4) Find the optimal codebook $\hat{\mathbf{x}}(P(\hat{A}_m)) = \{\hat{\mathbf{x}}(S_i),\, i = 1, \ldots, N\}$. Set $\hat{A}_{m+1} \equiv \hat{\mathbf{x}}(P(\hat{A}_m))$, $m = m + 1$ and go to (1).

There are several ways to choose the initial codebook $\hat{A}_0$ required by the above algorithm. A technique which yields to good codebooks is the one called "splitting" technique. Considering a M level quantizer with $M = 2^R,\; R = 0, 1, \ldots$ the method is summarized below

(0) Set $M = 1$ and define $\hat{A}_0(1) = \hat{\mathbf{x}}(A)$ the centroid of the whole training sequence

(1) Given the codebook $\hat{A}_0(M)$ containing M codewords $\{\mathbf{y}_i, i = 1, \ldots, M\}$ "split" each vector $\mathbf{y}_i$ into two close vectors $\mathbf{y}_i + \epsilon$ and $\mathbf{y}_i - \epsilon$, where ϵ is a fixed perturbation vector. The collection $\tilde{A}$ of $\{\mathbf{y}_i + \epsilon, \mathbf{y}_i - \epsilon,\; i = 1, \ldots, M\}$ has $2M$ vectors. Replace M with $2M$.

(2) Is $M = N$?. If so, set $\hat{A}_0 = \tilde{A}(M)$ and halt. $\hat{A}_0$ is then the initial codebook for the above described N level quantization algorithm. If not, run the algorithm for an M level quantizer on $\tilde{A}(M)$ to produce a good codebook $\hat{A}_0(M)$ and then go to (1).

Bibliography

[1] ITU-T Recommendation P. 835. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. 2003.

[2] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2):113–120, Apr 1979.

[3] ITU-R Recommendation BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems. 2001-2003.

[4] D. Burshtein and S. Gannot. Speech enhancement using a mixture-maximum model. *Speech and Audio Processing, IEEE Transactions on*, 10(6):341–351, Sep 2002.

[5] A. Buzo, Jr. Gray, A., R. Gray, and J. Markel. Speech coding based upon vector quantization. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(5):562–574, Oct 1980.

[6] O. Cappe'. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *Speech and Audio Processing, IEEE Transactions on*, 2(2):345–349, Apr 1994.

[7] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[8] J. H. L. Proakis J. G. Deller, J. R. Hansen. *Discrete-Time Processing of Speech Signal*. John Wiley & Sons, 2000.

[9] Y. Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *Signal Processing, IEEE Transactions on*, 40(4):725–735, Apr 1992.

[10] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1109–1121, Dec 1984.

[11] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):443–445, Apr 1985.

[12] J. Erkelens, J. Jensen, and R. Heusdens. A data-driven approach to optimizing spectral speech enhancement methods for various error criteria. *Speech Commun.*, 49(7-8):530–541, 2007.

[13] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Transactions on Audio, Speech & Language Processing*, 15(6):1741–1752, 2007.

[14] J.S. Erkelens, R.C. Hendriks, and R. Heusdens. On the estimation of complex speech dft coefficients without assuming independent real and imaginary parts. *Signal Processing Letters, IEEE*, 15:213–216, 2008.

[15] J.S. Erkelens and R. Heusdens. Tracking of nonstationary noise based on data-driven recursive noise power estimation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1112–1123, Aug. 2008.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993.

[17] R. Gray, A. Buzo, Jr. Gray, A., and Y. Matsuyama. Distortion measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):367–376, Aug 1980.

[18] R. Hendriks. *Advances in DFT-Based Single-Microphone Speech Enhancement*. PhD thesis, 2008.

[19] R.C. Hendriks, J.S. Erkelens, and R. Heusdens. Comparison of complex-dft estimators with and without the independence assumption of real and imaginary parts. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4033–4036, 31 2008-April 4 2008.

[20] R.C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems. Fast noise psd estimation with low complexity. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3881–3884, April 2009.

[21] R.C. Hendriks, J. Jensen, and R. Heusdens. Noise tracking using dft domain subspace decompositions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):541–553, March 2008.

[22] A. Kindoz and A. M. Kondoz. *Digital Speech Coding for Low Bit Rate Communication Systems*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[23] W. B. Kleijn and K. K. Paliwal, editors. *Speech Coding and Synthesis*. Elsevier Science Inc., New York, NY, USA, 1995.

[24] J. Laroche, Y. Stylianou, and E. Moulines. Hns: Speech modification based on a harmonic+noise model. *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 2:550–553 vol.2, Apr 1993.

[25] E. R. Larsen and R. M. Aarts. *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. John Wiley & Sons, 2004.

[26] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95, Jan 1980.

[27] P. C. Loizou. *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*. CRC, 1 edition, June 2007.

[28] T. Lotter and P. Vary. Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP J. Appl. Signal Process.*, 2005:1110–1126, 2005.

[29] Rainer M. Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Process.*, 86(6):1215–1229, 2006.

[30] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[31] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on*, 9(5):504–512, Jul 2001.

[32] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *Speech and Audio Processing, IEEE Transactions on*, 13(5):845–856, Sept. 2005.

[33] C. Plapous, C. Marro, and P. Scalart. Speech enhancement using harmonic regeneration. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1:157–160, 18-23, 2005.

[34] C. Plapous, C. Marro, and P. Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):2098–2108, Nov. 2006.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP '01: Proceedings of the Acoustics, Speech, and Signal Processing, 200. on IEEE International Conference*, pages 749–752, Washington, DC, USA, 2001. IEEE Computer Society.

[36] F. Soong and B. Juang. Line spectrum pair (lsp) and speech data compression. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, volume 9, pages 37–40, Mar 1984.

[37] T.V. Sreenivas and P. Kirnapure. Codebook constrained wiener filtering for speech enhancement. *Speech and Audio Processing, IEEE Transactions on*, 4(5):383–389, Sep 1996.

[38] S. Srinivasan. *Knowledge-Based Speech Enhancement*. PhD thesis, 2005.

[39] S. Srinivasan, J. Samuelsson, and W.B. Kleijn. Codebook driven short-term predictor parameter estimation for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):163–176, Jan. 2006.

[40] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *Speech and Audio Processing, IEEE Transactions on*, 9(1):21–29, Jan 2001.

[41] S V Vaseghi. *Advanced digital signal processing and noise reduction; 2nd ed.* Wiley, New York, NY, 2000.

[42] D. L. Wang and J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-30(4):679, 1982.

[43] An-Tze Yu and Hsiao-Chuan Wang. New speech harmonic structure measure and it application to post speech enhancement. *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 1:I–729–32 vol.1, May 2004.

[44] E. Zavarehei, S. Vaseghi, and Q. Yan. Noisy speech enhancement using harmonic-noise model and codebook-based post-processing. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1194–1203, May 2007.