uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Frédéric Delâge

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Wideband Speech Enhancement Approaches Using a Kalman Filter and a Perceptual Post-Filter

TITRE DE LA THÈSE / TITLE OF THESIS

M. Bouchard

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

H. Ding

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

C. Giguère

R. Dasereau

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

WIDEBAND SPEECH ENHANCEMENT APPROACHES USING A KALMAN FILTER AND A PERCEPTUAL POST-FILTER

by

Frédéric Delâge

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

Master of Applied Science

School of Information Technology and Engineering

University of Ottawa

Ottawa, Ontario

Your file Votre référence
ISBN: 978-0-494-49308-3
Our file Notre référence
ISBN: 978-0-494-49308-3

ABSTRACT

This thesis discusses the issue of speech enhancement in the presence of white noise for wideband speech. Speech enhancement aims at improving one or more perceptual aspects of speech such as overall quality, intelligibility, or listener fatigue. For example, reducing the amount of background noise in an automatic recognition system might improve the speech recognition rate by a considerable amount improving the intelligibility for machine recognizers.

Most speech enhancement algorithms have been developed and tested for conventional narrowband speech systems (sampling rate of 8 kHz). This thesis shows that some narrowband algorithms can be extended to be used for noise reduction in wideband speech applications (sampling rate of 16 kHz) by presenting a slightly modified Kalman filter based approach that uses a perceptual post-filter. The results are compared with the performance of other speech enhancement algorithms. The results show that the proposed method has a good potential for speech enhancement in wideband, especially in lower SNRs environments.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr. Martin Bouchard and Dr. Heping Ding for their direction, support, and patience.

TABLE OF CONTENTS

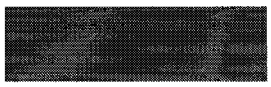

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation

We live in a noisy world. In a society where mobile telecommunications, conference calls, and automatic recognition systems have become an integral part of day-to-day activities, the role of noise reduction and speech enhancement algorithms has gained considerable importance. The main objective of speech enhancement is to improve one or more perceptual aspects of speech such as overall quality, intelligibility for human or machine recognizers, or degree of listener fatigue [1]. Although speech quality and intelligibility can be degraded by many factors (e.g. transmission through communication channels, low-cost loudspeakers, microphones), this thesis will focus on the speech degradation caused by additive background noise; unwanted sounds present in the environment of the speaker picked up by the microphone during a communication. The basic overview of a noise reduction system is shown in figure 1.1.

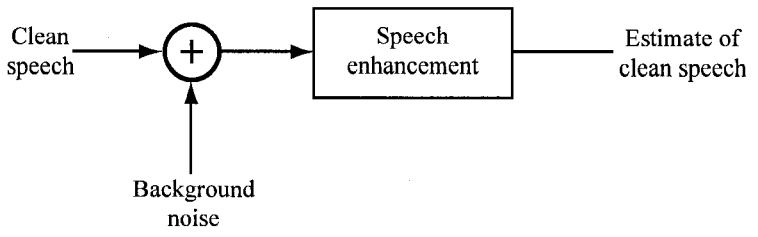

FIGURE 1.1 - Overview of noise reduction system

In telecommunications, applications of noise reduction are numerous ranging from the suppression of car engine noise while one is using a hands-free phone, to the suppression of street noise during a mobile communication. Furthermore, telecommunication systems do not perform particularly well in noisy environments. Spectral parameters cannot be properly estimated by the coding blocks in the presence of noise, resulting in a coded speech that sounds mechanical and distorted.

The applications of noise reduction are not limited to the telecommunication field. Hearing aids are devices that amplify ambient sounds for the hearing impaired. Amplifying the added background noise as much as the desired audio signal can strain the listener and be counter-productive. The restoration of audio archives is also an interesting application. In this case, the aim is to reduce the hisses, the cracks, and the pops introduced by the recording media to obtain an enhanced signal as close as possible to the original.

Although many of the applications are relatively new, noise reduction is not a new problem and algorithms have been around for quite a while. With the appearance of digital signal processing and the increase in computational power, many approaches have spawned new and powerful algorithms which have proven to be successful in improving the quality of speech signals even under harsh acoustic conditions. Currently, the improvement of speech intelligibility is limited to an increase in recognition in automatic recognition systems.

Furthermore, with this development of technology, the transmission of wideband speech over the telephone network has become a new reality. This increase in speech quality would allow face-to-face-like communications over the network [10].

1.2 Goal of The Thesis

The goal of the thesis is to investigate speech enhancement algorithms performance in wideband applications (applications where the speech is sampled at 16 kHz). Most of the

noise reduction methods published have been proposed and tested in narrowband environments. With wideband technologies becoming more and more available, wideband speech enhancement algorithms evaluation is required. Moreover, some modifications may be required for the algorithms to perform properly in a wideband environment, and this will also be considered when required.

This thesis focuses on a speech enhancement algorithm based on the Kalman filter. Perceptual masking can be achieved using a post-filter to improve the results in some conditions. The algorithm is based on the work of Ma, Bouchard and Goubran [3], [4], [5], [6] but is extended to wideband and further modified to improve the perceptual quality.

1.3 Thesis Organization

The thesis is organized into 6 chapters. Chapter 2 serves as background information for the remainder of the thesis. Selected speech enhancement algorithms are presented in Chapter 3. Chapter 4 presents the Kalman filter algorithm and its perceptual masking post-filter. Implementation details are also included. The results of the simulations are presented in chapter 5. Finally, chapter 6 concludes the work.

Note: in this work, speech enhancement and noise reduction are used interchangeably

1 INTRODUCTION

Thesis Organization

CHAPTER 2

BACKGROUND INFORMATION

2.1 Narrowband Versus Wideband Speech Transmission

2.1.1 Narrowband

The bandwidth of today's telephony network is limited to about 200-3400 Hz and sampled at a rate of 8 kHz, a limitation built-in the telephone system dating from its origin. This is referred to narrowband speech. Because most of the energy of the speech signal is present below 7 kHz (although it may extend to higher frequencies), a decent representation of the original speech is achievable.

2.1.2 Wideband

Wideband speech refers to speech that is sampled at 16 kHz and bandlimited to 50-7000 Hz. The added bandwidth results in an increased speech quality close to face-to-face. With the appearance and increased use of end-to-end digital networks such as wireless systems, ISDN, and voice over packet networks, wideband speech coding standards have been developed (e.g. AMR-WB [9]) and permit a much higher speech quality [8].

2.1.3 Effect on speech

Wideband speech results in major subjective improvements in speech quality compared to narrowband speech. These include increased naturalness, presence, comfort, better fricative differentiation and therefore higher intelligibility. It also adds a feeling of transparent communication and eases speaker recognition [8]. A comparison of both narrowband

and wideband speech is shown in figure 2.1. One can easily notice the quantity of information that is lost resulting in the aforementioned loss of quality.

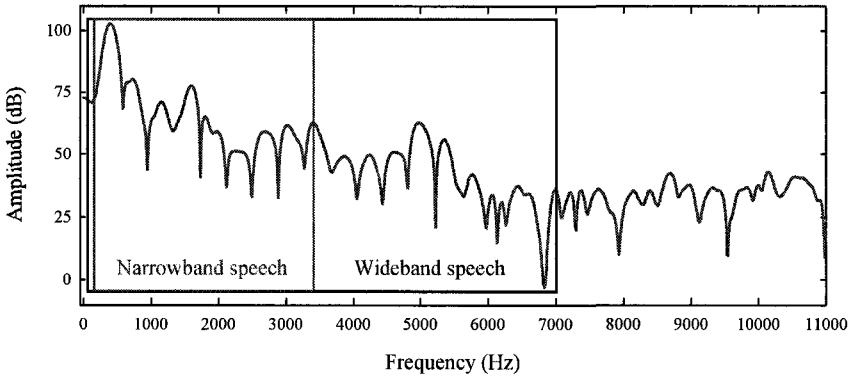

FIGURE 2.1 - Energy spectrum of speech segment

2.2 Noise Corruption

The objective of speech enhancement is to reduce the noise in speech signals to an acceptable level. In speech applications, a general model of noise is

$$y(n) = s(n)*h(n) + v(n) \qquad \text{(EQ 2.1)}$$

where *y(n)* is the noisy speech signal, *s(n)* is the clean speech, *h(n)* represents the transfer function of the transmission channel and is said to cause convolutional noise (i.e. unwanted filtering operations), and *v(n)* is the additive noise. The * operator denotes a convolution.

2.2.1 White Gaussian Noise Characteristics

In this work, we only concern ourselves with the case of additive noise

$$y(n) = s(n) + v(n) \qquad \text{(EQ 2.2)}$$

Furthermore, *v(n)* is assumed to be zero-mean white noise with a Gaussian distribution. White noise is a random signal with a flat power spectral density, i.e. every possible frequency is present in the same amount. The level of the power spectral density is σ^2, also

known as the variance of the process. A direct consequence of this flat spectrum is the uncorrelation between each noise sample. A sample cannot be predicted by other past and/ or future samples.

White noise is often assumed in order to simplify computations and increase mathematical tractability. Although noises in nature cannot necessarily be fully modelled by white noise, it serves as a proper starting point to speech enhancement in wideband conditions. Once this assumption has been made, we will see that the Kalman filter is a natural choice for speech enhancement.

2.2.2 Signal-To-Noise Ratio

The signal-to-noise ratio is a measure used to quantify the amount of noise in a signal. It is the power ratio between a signal and the background noise and is defined as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n v^2(n)} \ \mathrm{dB} \qquad \text{(EQ 2.3)}$$

where $s(n)$ is the speech signal and $v(n)$ the background noise. This measure is directly affected by the activity of the speech (i.e. speech versus silence) but will be computed consistently over the entire speech sample including speech and silent periods.

2.2.3 Clean vs. Noisy

To show the effect of white Gaussian noise on wideband speech, figure 2.2 shows the spectrogram of both a clean speech sample and noisy speech sample at a 6 dB SNR. As can be observed, most of the frequency features are "drowned" in the noise. The objective of the noise reduction algorithms presented in this thesis are to recover as much information as possible from the spectrogram on the right to get as close as possible to the spectrogram on the left.

2.3 Sound Loudness, Intensity and dB SPL

The human ear is able to hear an extremely wide variety of sound intensities and this fact justifies the use of the dB scale. Sounds are simply small, rapid changes in air pressure relative to the constant pressure of the atmosphere. The faintest perceivable sound has been measured to be around 20 μPa for a pure sine wave at 1000 Hz (this value varies in function of the frequency and from one listener to the other). Nevertheless, this has been chosen as the reference level or the absolute threshold of hearing. Pressure levels using this reference are termed sound pressure levels (SPL). The intensity of a sound is another common measure and it refers to the amount of energy transported past a given area per unit of time. It is proportional to the square of the pressure and is also used to compute the SPL [13]. The intensity reference level is 10^{-12} W/m^2.

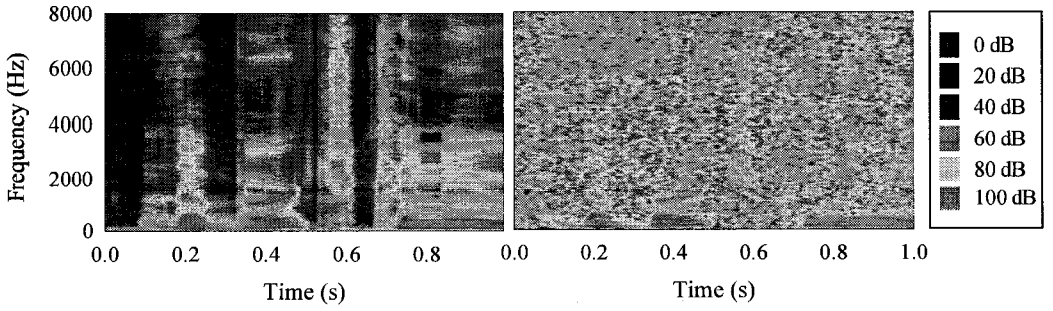

FIGURE 2.2 - Speech spectrogram for clean speech (left) and noisy speech (right)

$$\text{dB SPL} = 10\log\left(\frac{I}{I_0}\right) = 10\log\left(\frac{P^2}{P_0^2}\right) = 20\log\left(\frac{P}{P_0}\right) \quad \text{(EQ 2.4)}$$

where I is the intensity, I_0 the intensity reference level, P the pressure, and P_0 the reference pressure level. A few examples of dB SPL are shown in table 2.1 [11].

Sound example	dB SPL
Loud rock concert	120
Normal conversation	60-70
Rustling leaves	20
Reference level	0

Table 2.1: Approximate SPL for common sounds

The loudness of a sound is a subjective term describing the strength of the ear's perception to that sound. One unit of measure of the loudness is the sone. One sone is equal to the loudness of a 1 kHz tone at 40 dB SPL and the scale is chosen so that doubling the number of sones, doubles the loudness. Since the sensitivity of the ear varies according to the frequency, the conversion from the intensity of a sound to its loudness is done through equal loudness curves [12], [14]. An example of such a curve can be found in figure 2.5 on page 13.

2.4 Speech Production and Modelling

A brief overview of speech production will be given in this section. From there, a linear model for speech production will be presented. This model will be used in the enhancement algorithms.

2.4.1 Physical Speech Production

A speech signal is produced when the airflow, pushed out of the lungs is shaped at the larynx and further constricted by the shape of the vocal tract [15]. Each sound has its own positioning of the vocal tract articulators (vocal cords, tongue, lips, teeth, velum, and jaw).

At the base of the vocal tract, the airflow is either interrupted periodically by the vocal folds or forced into narrow constrictions resulting in turbulence. This air source passes through the vocal tract which acts as a time-varying filter. When the airflow is periodic, the sound is classified as voiced (mostly the vowels) and when it is noise-like, it is classified as unvoiced. There can also be a combination of both phenomenon.

The spectral domain representation of voiced speech consists of harmonics of the fundamental frequency (F_0) of the vocal cord vibration. The envelope of the spectrum of a voiced sound is characterized by a set of peaks which are called formants. The spectrum for unvoiced speech is less important perceptually [16].

2.4.2 Model of Speech Production

A basic model of speech production is given in figure 2.3 [2]. It models both the excitation source and the vocal tract shaping. The excitation of the vocal tract filter is modeled as a periodic impulse train for voiced speech and white noise for unvoiced speech. *H(z)* is a time varying filter modelling the vocal tract.

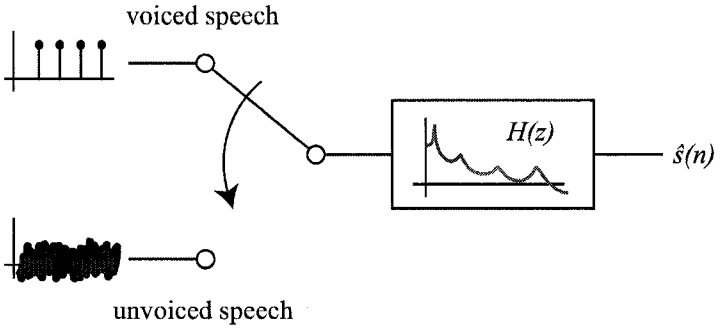

FIGURE 2.3 - Model of speech production

The speech is modelled as an autoregressive (AR) system for its simplicity and computational efficiency. The speech signal *s(n)* is predicted from a linear combination of the past outputs and the current input. It is expressed by this difference equation

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad \text{(EQ 2.5)}$$

where G is a gain factor, $\{a_k\}$ are the filter coefficients, and p the order of the linear prediction. By taking the z-transform of equation 2.5, the transfer function $H(z)$ of the all-pole system can be computed

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad \text{(EQ 2.6)}$$

The voiced sounds are modelled well enough by the poles but the unvoiced can only be crudely approximated by the all-pole model. However, solving an all pole model is fairly easy and justifies the use of the model. If $G = 1$, equation 2.6 simplifies to

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \qquad \text{(EQ 2.7)}$$

The error signal $e(n)$ is computed by the difference of the real speech value and the predicted value

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad \text{(EQ 2.8)}$$

In the z-domain, this in turn becomes

$$E(z) = S(z)A(z) \qquad \text{(EQ 2.9)}$$

The filter coefficients $\{a_k\}$ are called the linear prediction (LP) coefficients. They are derived in such a way that the energy of $e(n)$ is minimized. This signal is also known as the residual, i.e. the information that cannot be modelled by the filter. Ideally, the residual is a periodical impulse train for voiced speech and white noise for unvoiced speech. The filter coefficients are found by solving a set of linear equations which can be done efficiently using the Levinson-Durbin algorithm. Details can be found in [17].

From the residual, it is possible to obtain the original speech by filtering with $1/A(z)$; this is called synthesis. From an estimate of the residual, it is only possible to estimate the original speech and obtain $\hat{s}(n)$. Filtering the speech signal to get the residual is called analysis as shown in figure 2.4 [16].

The analysis and synthesis filters are only valid for a very short time because the vocal tract, which the filter $H(z)$ tries to approximate, is a time-varying filter and changes constantly. Nevertheless, speech is usually considered stationary (i.e. the speech statistics do not change and the vocal tract remains constant) for periods of around 20-30 ms. This implies that a speech signal must be separated into frames of 20 ms, often slightly overlapping, and that the filter parameters $\{a_k\}$ must be computed for each frame for a decent representation of the original [2].

2.5 Human Auditory Masking

The human auditory system is the assembly of the ear, the auditory nerve fibers, and section of the brain that converts acoustical energy into sounds that we can hear. The phenomenon of auditory masking occurs when the perception of one sound is obscured by the perception of another [18]. The masking sound is called the *masker* and the weaker signal is referred to the *maskee*. The threshold level above which a signal becomes audible in the presence of a masker is known as the *masking threshold* [19]. Masking is a complex phe-

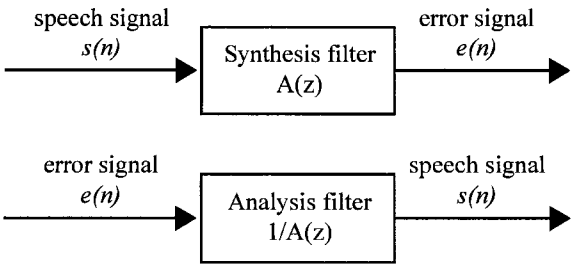

FIGURE 2.4 - LP analysis and synthesis model

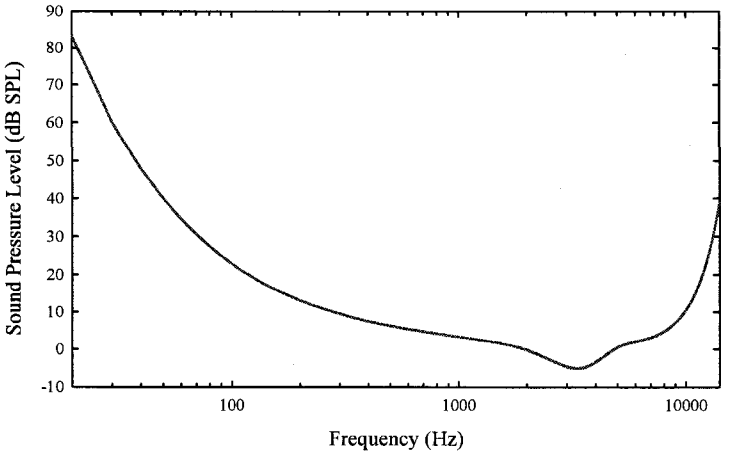

FIGURE 2.5 - Absolute threshold of hearing (equal loudness curve)

nomenon which is only partially understood. Nevertheless, models have been developed to emulate its behavior based on observations.

2.5.1 Absolute Threshold of Hearing

To be audible, sounds require a minimum pressure level. Due in part to filtering in the outer and middle ear, this minimum pressure varies considerably with frequency, individuals, and their age. This threshold is called the *Absolute Threshold of Hearing* (ATH) and for signal processing purposes, is approximated by

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \text{ (dB SPL)} \qquad \text{(EQ 2.10)}$$

[20]. This is shown in figure 2.5.

2.5.2 Critical Bands

For a given frequency, a *critical band* (CB) is the smallest band of frequencies around it which activates the same part of the basilar membrane in the ear. In other words, they are frequency ranges in which one pitch will mask another pitch. The nerve endings over the area of a critical band are excited simultaneously when a sound wave strikes a point of the membrane in that band. This overall excitation reduces its sensitivity to other stimuli.

The sensation of loudness is therefore not much affected by frequencies close to the original one [21]. There is some debate on the actual limits and width of the critical bands but the model by Zwicker is commonly used. These band limits are shown in table 2.2 [22].

Critical Band Number	Frequency Range (Hz)	Critical Band Number	Frequency Range (Hz)
1	0 - 100	13	1720 - 2000
2	100 - 200	14	2000 - 2320
3	200 - 300	15	2320 - 2700
4	300 - 400	16	2700 - 3150
5	400 - 510	17	3150 - 3700
6	510 - 630	18	3700 - 4400
7	630 - 770	19	4400 - 5300
8	770 - 920	20	5300 - 6400
9	920 - 1080	21	6400 - 7700
10	1080 - 1270	22	7700 - 9500
11	1270 - 1480	23	9500 - 12000
12	1480 - 1720	24	12000 - 15500

Table 2.2: Critical Band Limits

2.5.3 Bark Scale

The Bark scale ranges from 1 to 24 corresponding to the first 24 critical bands of hearing. A common function to convert from the frequency scale to the bark scale is given by

$$z(f) = 13\,\mathrm{atan}(0.00076f) + 3.5\,\mathrm{atan}\left[\left(\frac{f}{7500}\right)^2\right] \qquad \text{(EQ 2.11)}$$

[22].

2.5.4 Simultaneous Masking

For simplification, the masking effects are often classified into two different categories: simultaneous and temporal masking. Simultaneous masking occurs when the maskee and masker are present at the same time and masking occurs in the frequency domain. The masker can either be tone-like or noise-like, the latter being a more effective masker for

equal loudness maskers [24]. For tonal maskers, the slope towards lower frequencies becomes steeper when increasing the magnitude of the masker. Noise-like tone have been shown to have invariant low frequency slopes relative to the masker level. Furthermore, low frequencies tend to mask higher frequencies.

Masking between critical bands has also been observed; an effect called spread of masking. A masker within one critical band has some reduced but predictable effects on detection thresholds in other neighboring critical bands. It is modelled by a spreading function and often approximated by the following function [25]

$$SF_{dB}(k) = 15.81 + 7.5(k + 0.474) - 17.5\sqrt{1 + (k + 0.474)^2} \text{ dB} \qquad \text{(EQ 2.12)}$$

which is shown in figure 2.6. Both noise-like and tone-like maskers use the same spreading function but their different masking capabilities are taken into account by the coefficient of tonality in the Johnston model (next section).

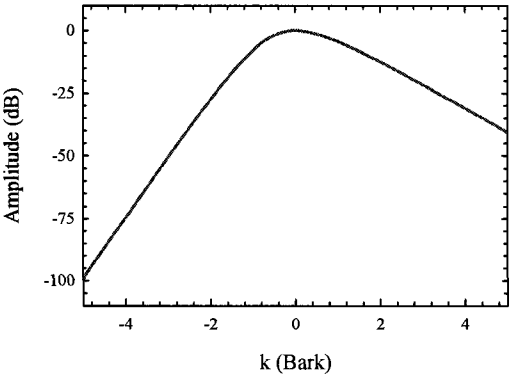

FIGURE 2.6 - Spreading function

The Johnston Model

A popular simultaneous masking model has been developed by Johnston [26] for coding audio sampled at 32 kHz. It was used to compute the auditory masking threshold, to figure out how much noise the coder can add before it becomes audible. The steps are:

1. Critical Band Analysis

The first step consists of calculating the energy present in each critical band. This is done from the power spectrum (more information on the actual computation of the power spectrum is given in section 4.4.2 on page 48). The frequency bins are assigned to their appropriate critical band according to the frequency limits presented in table 2.2 on page 14

$$P(i) = \sum_{m=b_l(i)}^{b_h(i)} P_{xx}(m) \qquad \text{(EQ 2.13)}$$

where $b_l(i)$ and $b_h(i)$ are the lower and upper boundaries of the i^{th} critical band and $P_{xx}(m)$ is the power spectrum of the signal.

2. Spreading Function

The masking between critical bands is taken into account by convolving the power spectrum with the spreading function $SF(i)$ of equation 2.12 (converted to the linear scale) to get the spread critical band spectrum $C(i)$

$$C(i) = 10^{0.1SF_{dB}(i)} * P(i) \qquad \text{(EQ 2.14)}$$

where i is the critical band number.

3. Calculation of the Noise Masking Threshold

Two masking thresholds are used to model both tone-like and noise-like maskers. The tone-like masking threshold is estimated as $14.5+k$ dB below $C(i)$. The noise-like masking threshold is estimated as being a uniform 5.5 dB below $C(i)$ across the whole critical band spectrum [27], [25]. The *Spectral Flatness Measure* (SFM) is used to determine which threshold to use

$$SFM_{dB} = 10\log 10\left(\frac{G}{A}\right) \text{ dB} \qquad \text{(EQ 2.15)}$$

where G is the geometric mean

$$G = \left(\prod_{i=1}^{Z} P(i) \right)^{1/Z} \quad \text{(EQ 2.16)}$$

and A is the arithmetic mean

$$A = \frac{1}{Z} \sum_{i=1}^{Z} P(i) \quad \text{(EQ 2.17)}$$

where Z is the maximum critical band number required. A coefficient of tonality α is generated as

$$\alpha = min\left(\frac{SFM_{dB}}{SFM_{dBmax}}, 1 \right) \quad \text{(EQ 2.18)}$$

where SFM_{dBmax} = -60 dB represents the SFM of an entirely tone-like signal. Therefore, if $\alpha = 1$, the signal is estimated as being entirely tone-like and if $\alpha = 0$, the signal is entirely noise-like. Anything in between is a combination of both. The threshold offset can then be computed from

$$O_{dB}(i) = \alpha(14.5 + i) + (1 - \alpha)5.5 \text{ dB} \quad \text{(EQ 2.19)}$$

The threshold offset $O(i)$ is then subtracted from the spread critical band spectrum in the linear domain given the simultaneous threshold estimate

$$T_s(i) = 10^{(10\log 10[C(i)] - O_{dB}(i))/10} \quad \text{(EQ 2.20)}$$

4. Renormalization

The renormalization aims at "undoing" the convolution done in equation 2.14 on the critical band energy $P(i)$. Given the shape of the spreading function, deconvolution is a very unstable operation and thus a renormalization is used instead. In the original model [26], the renormalization multiplies $T_s(i)$ by the inverse of the energy gain, assuming a uniform energy of 1 in each band. In [32], a simpler approach is

used. The renormalization consists of simply dividing each band of $T_s(i)$ by the number of FFT bins in that respective band. This is the method preferred in this work

$$T_n(i) = \frac{T_s(i)}{b_h(i) - b_l(i) + 1} \quad \text{(EQ 2.21)}$$

where again, $b_l(i)$ and $b_h(i)$ are the lower and upper boundaries of the i^{th} critical band.

5. Comparison to the ATH

The normalized threshold $T_n(i)$ is then compared to the absolute threshold of hearing $T_q(f)$ (equation 2.10). Because this operation takes place in the Bark scale, $T_q(f)$ must be converted to the Bark scale first using equation 2.13 giving $T_q(i)_{bark}$. Any critical band that has a calculated threshold below the absolute threshold of hearing is changed to the absolute threshold of hearing in that critical band.

$$T_n'(i) = max(T_n(i), T_q(i)_{\text{bark}}) \quad \text{(EQ 2.22)}$$

2.5.5 Temporal Masking

Temporal masking occurs when the masker and maskee have a slight time delay. There are two forms of this masking: forward masking and backward masking. Forward masking masks signal components that occur some short delay after the masker in the same critical band and has an effective duration up to 200 ms depending on the length of the masker, the delay, and the frequency [28]. Backward masking masks signal components that occur

before the masker. The different behaviors can be observed in figure 2.7 [29]. This work only considers forward masking.

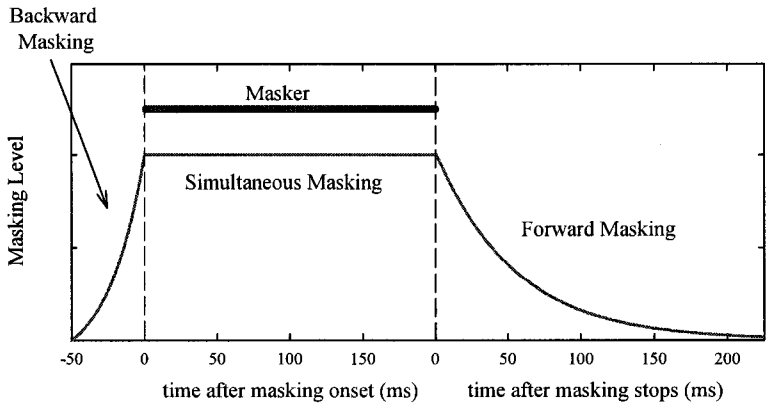

FIGURE 2.7 - Simultaneous and temporal masking

The Zwicker Model

The forward masking effect has been modelled by a non-linear circuit proposed in [30] to estimate the output specific loudness in each critical band. This circuit is shown in figure 2.8 and there is one per critical band.

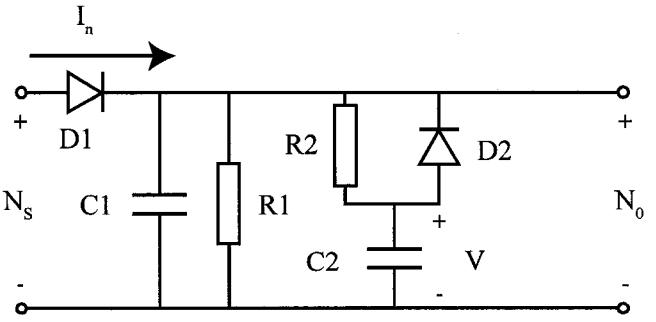

FIGURE 2.8 - Non-linear circuit for loudness estimation

The estimated specific loudness for the i^{th} critical band is given by

$$N_s(i) = 0.08\left(\frac{E_T(i)}{E_0}\right)^{0.23}\left[\left(0.5 + 0.5\frac{E(i)}{E_T(i)}\right)^{0.23} - 1\right] \qquad \text{(EQ 2.23)}$$

where $E_T(i)$ is the excitation at the absolute threshold of hearing, E_0 is the excitation for 0 dB SPL, and $E(i)$ is the excitation of the current power spectral density, i.e. the energy in the critical bands ($C(i)$ of equation 2.14). More details on the actual computation of these

values will be given in section 4.4.3. From this, the low-pass circuit is used to emulate the non-linear loudness processing in the neural system and generate the output specific loudness $N_0(s)$ [29].

The differential equations are linearized and converted into the following difference equations

$$I_n = \begin{cases} \frac{N_s - N_0^*}{R_{D1}}, & N_s > N_0^* \\ 0, & \text{otherwise} \end{cases} \quad \text{(EQ 2.24)}$$

where R_{D1} is the on resistance of diode D1, N_s the specific loudness of the current frame, and N_0^* the output specific loudness of the previous frame. The asterisk refers to the value of the previous frame. Referring to figure 2.8, N_0^* and V^* are the voltages across capacitors C1 and C2 in the previous frames respectively. If $N_0^* > V^*$, then

$$N_0 = \frac{I_n + C2V^*/(\Delta t + C2R2) + C1N_0^*/(\Delta t)}{1/R1 + C2/(\Delta t + C2 + R2) + C1/\Delta t} \quad \text{(EQ 2.25)}$$

$$V = N_0^* \Delta t/(\Delta t + R2 + C2) + V^* C2R2/(\Delta t + C2R2) \quad \text{(EQ 2.26)}$$

But if $N_0^* < V^*$, D2 is turned off and the following holds

$$N_0 = V = \frac{I_n + (C1 + C2)V^*/\Delta t}{1/R1 + (C1 + C2)/\Delta t} \quad \text{(EQ 2.27)}$$

where Δt is the time difference between two frames in ms. The total loudness, N, is defined as the sum of the output specific loudness in all critical bands and is directly related to the forward masking level [31] and is then normalized by a 60 dB uniform masking noise so that it lies between 0 and 1. If N is larger than one, it is set to one.

2.5.6 *Combining the Simultaneous and Temporal Masking*

The total masking levels can be determined according to the method proposed by Huang and Chiueh in [29]. The levels depend not only on the current frame but also on the previous frames

$$T_t(i) = max(T_n'(i), T_t(i)^* \cdot e^{-\Delta t/(\tau(i)\cdot N)}) \qquad \text{(EQ 2.28)}$$

where $T_t(i)$ and $T_t(i)^*$ are the total masking levels of the current and previous frame respectively, $T_n'(i)$ is the masking level computed from the Johnston simultaneous masking model [26] (equation 2.22), Δt is the time difference between two frames, $\tau(i)$ is the maximum decay time constant in each critical band [28] and is set to 10 ms, and N is the normalized total loudness.

2.5.7 *Filtering Operation*

Now that both simultaneous and temporal masking have been considered and before the perceptual filtering operation takes place, the final threshold $T_t(i)$ must be converted back to the linear frequency domain. This is done by assigning the value of the critical band to the frequency bins included in that band. This value is then divided by the number of bins in that critical band to keep the energy constant in each band. This results in the final threshold $T_{final}(n)$ in the linear frequency domain.

The filtering operation is done as a threshold operation as specified in [3] and follows the following equation

$$|\tilde{X(n)}|^2 = \begin{cases} |X(n)|^2 \cdot \alpha^{(c(n))/9}, & \frac{1}{N}|X(n)|^2 < T_{final}(n) \\ |X(n)|^2 \cdot \alpha^{(c(n)-1)/9}, & \text{otherwise} \end{cases} \qquad \text{(EQ 2.29)}$$

where $\tilde{X(n)}$ is the N-point discrete Fourier transform of the filtered signal, $X(n)$ is the N-point discrete Fourier transform of the original signal, $c(n)$ is the critical band number cor-

responding to the frequency bin n, and α is the tonality coefficient computed in equation 2.18. The reasoning behind this operation is that noise-like frames should be more attenuated than tone-like frames. Furthermore, the higher frequencies tend to be more noise-like than the lower ones and are therefore attenuated in consequence. Finally, because the perceptual mask is only an approximation of the real mask, the signal above the mask is also attenuated but to a lesser extent. Equation 2.29 was developed heuristically taken these into account.

Finally, the perceptually filtered signal can be reconstructed by taking the inverse discrete Fourier transform of the new computed magnitude and the original signal's phase

$$\tilde{x}(t) = \text{IFFT}\{|\tilde{X(n)}| \cdot e^{j\angle X(n)}\} \quad . \qquad \text{(EQ 2.30)}$$

Implementation specific details can be found in section 4.4 on page 47.

2.6 Quality Assessment

The assessment of the quality of an audio signal is not a trivial matter. Often, an audio signal will sound better to a certain listener than a different signal that might have been preferred by another listener. This merely implies that, in essence, the determination of speech quality is a measure which is completely subjective. Albeit this fact, there are measures and tools that can be used to facilitate the process of assessing speech quality. Some of these will be described below, focusing on the ones that will be implemented and used in this thesis.

2.6.1 Speech Quality vs. Speech Intelligibility

In the field of speech enhancement, the system usually aims at improving speech quality and/or speech intelligibility. This can be confusing and the difference between the two is subtile. The quality of speech is a subjective measure which reflects on the way the signal is perceived by listeners, whereas the intelligibility is an objective measure of the amount of information which can be extracted by listeners from the given signal [33].

The application of noise reduction systems will usually indicate if speech quality or speech intelligibility or both should be improved and this will have an important effect on the approach used. In speech recognition systems, one can deduce that the intelligibility is more important since extracting as much information as possible is key, while in telecommunication systems, the speech quality might be slightly more important to increase the enjoyment of the communication and reduce listener fatigue.

Unfortunately, speech quality and speech intelligibility are not proportional and improving one might improve, worsen, or leave unchanged the other. And so, it is important to determine the objective of the speech enhancement system. In this thesis, we are trying to improve the speech quality without affecting too much the speech intelligibility.

2.6.2 Subjective Evaluation

Subjective evaluation relies on the opinion of a group of human listeners to rate the quality or intelligibility of processed speech. Because the objective of noise reduction algorithms is to improve the perceived quality of a signal, in theory, subjective evaluation is essential.

MOS

The methods for formal subjective testing are rigidly specified. They are often time consuming and costly as they require proper training of listeners. In addition, a constant listening environment and identically tuned output are necessary to generate consistent results. In the telecommunications field, the most common is the Mean Opinion Score (MOS) testing as specified by the ITU-T recommendations P.800 and P.830 [34], [35]. It quantifies the performance of speech algorithms by assigning a value from 1 to 5 accord-

ing to table 2.3. The MOS ratings are reliable as they are based on human response and

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

Table 2.3: Verbal categories and the five point scale for MOS

perception, but a large number of listeners is required so that a reasonable assessment can be made about system performance.

MUSHRA

Due to the complexity, time and resources required in formal MOS testing, it is not feasible within the scope of this thesis. Instead, testing will be carried out based on the ITU-R MUSHRA procedure [39]. The MUSHRA method (MUlti Stimulus test with Hidden Reference and Anchors) has been designed to give a reliable and repeatable measure of the audio quality of intermediate-quality signals.

2.6.3 Objective Evaluation

Objective methods rely on a mathematically based measure between the original and degraded speech signal. The success of these measures rests with their correlation with subjective quality [41].

WPESQ

The Wideband Perceptual Evaluation of Speech Quality (WPESQ) is an algorithm very recently standardized under P.862.2 [37] by the ITU-T. It is based on the PESQ algorithm (P.862 [36]) and it returns a value correlated to the MOS. Differently stated, it is an estimate of the score that should be obtained when formal subjective tests are carried out.

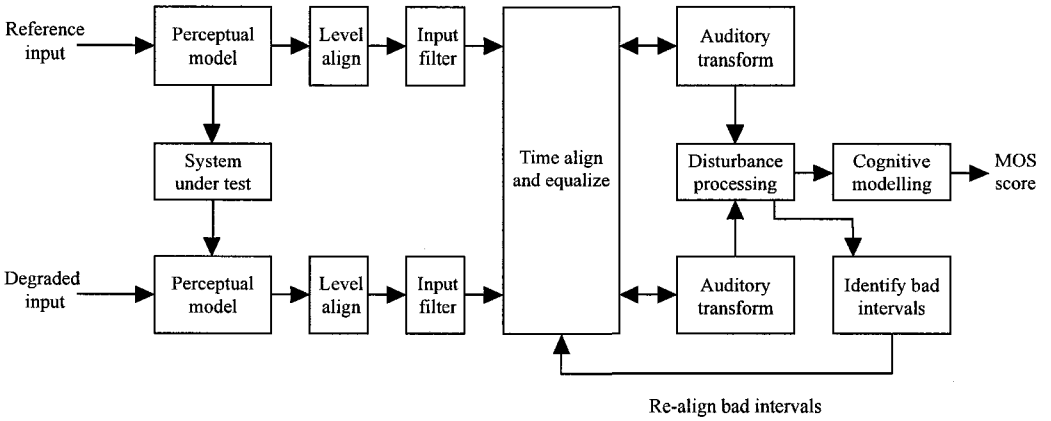

FIGURE 2.9 - Overview of the WPESQ algorithm

A high-level structure of the model is seen in figure 2.9 [40]. Initially, the reference signal and degraded (or enhanced) signal are level aligned such that they have identical playback levels. The signals are filtered by a wideband (50-7000 Hz) filter before being time-aligned and equalized. The auditory filters deployed by [38] are then used to process the ensuing signals. The auditory transformation also involves equalization for linear filtering in the system and gain variation. Two distortion parameters are extracted for the disturbance (the difference between the transforms of the two signals). They are then aggregated in frequency and time domain, and mapped to a prediction of subjective MOS [36]. The range for WPESQMOS mapping is from 1.0 to 4.5, in contrast to the range listed in table 2.3. This is mainly due to the fact that WPESQ simulates a listening test and is optimized to reproduce the average result of all listeners and statistically, 4.5 is around the best average one can generally expect from listening tests.

SNR

Signal-to-noise ratio is one of the most frequently used performance criterion when it comes to rating noise-corrupted signals but it is not the most appropriate for speech enhancement. A study by Quackenbush [42] shows that the ability of the SNR to predict

subjective quality is very poor and should probably not be used. Although the SNR is used to describe the level of corruption of a speech sample, it will not be used to rate the enhanced speech. The SNR is computed as follows

$$\text{SNR} = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n [\hat{s}(n) - s(n)]^2} \qquad \text{(EQ 2.31)}$$

and will be computed over the entirety of the speech sample.

Segmental SNR

Since the correlation of SNR with subjective quality is so poor, it is of little interest as a general objective measure of speech quality [42]. The frame-based segmental SNR is formed by averaging frame level SNR estimates and is a reasonable measure of speech quality. The entire speech signal is separated into M frames of N samples and the segmental SNR is computed as such

$$\text{SNR}_{\text{SEG}} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1} s^2(n)}{\sum_{n=Nm}^{Nm+N-1}[\hat{s}(n) - s(n)]^2} \qquad \text{(EQ 2.32)}$$

where $s(n)$ is the reference signal and $\hat{s}(n)$ is the processed speech. Frames with SNRs above 35 dBs do not reflect large perceptual differences and can be replaced by 35 dB in equation 2.32. Likewise, in periods of silence, SNR values can become very negative since signal energies are small. Again, these frames do not reflect the perceptual contributions of the signal and a lower threshold is often set. In this case, it has been set to -10 dB [41].

CHAPTER 3

SPEECH ENHANCEMENT OVERVIEW

3.1 Introduction

There are numerous and varied speech enhancement algorithms that have been proposed in the past and a fair amount of classification possibilities. Here are a few examples [1]:

- **General approach** - Most speech enhancement methods fall into a few general approaches i.e. spectral subtraction, short time spectral amplitude, Kalman filtering, subspace decomposition, etc. Some methods lie in multiple categories combining their benefits and reducing their ill-effects.
- **Single or multiple sensors** - Speech enhancement can be accomplished by using one or multiple microphones. In the former, noise characterization is required and is usually done during periods of silence, although in some cases it is done concurrently with the speech. In the latter, multiple microphones provide more information and can therefore result in better enhancement. Having multiple sensors is not as convenient and is more expensive.
- **Desired output** - In some applications, the desired output will be a standard waveform where the noise has been, hopefully, reduced to acceptable levels. In other applications, speech features might be the desirable output. Applications such as speech recognition require features for the recognition process to take place, and properly enhanced features might not be equivalent to enhanced waveforms.

In this work, we will only concern ourselves with the enhancement of speech waveforms in single-channel environments.

One challenge of speech enhancement is the balanced trade-off between noise reduction and speech distortion [43]. The reduction of noise levels is limited by the actual objective of the speech enhancement algorithms, removing most noise while leaving comfortable residual noise with minimal speech distortion. Removing more noise results in unnatural distorted speech and removing too little leaves you with noisy speech [44].

A common distortion introduced in enhanced speech is musical noise, a residual artefact that can sometimes be deemed more disturbing than the original noise [40]. It is characterized by randomly spaced peaks in the spectrum of the enhanced spectrum and is usually due to the crude estimation of noise spectrum used in many approaches. In the time-domain, this residual noise is unnatural, sounding like pure tones of random frequencies being turned on and off [45]. Several approaches have been proposed to reduce the problem such as: time averaging, noise floors, oversubtraction, and perceptual filtering [40].

Many algorithms require the use of noise estimation during the enhancement process. In some cases, this estimation is done at the beginning of the speech signal where it is assumed that there is no speech present. Noise statistics can be computed and averaged over a few frames. However, if the noise is non-stationary and its statistics vary over time, the enhancement process will degrade rapidly. To circumvent this problem, noise estimation can be done during the enhancement process but requires a *voice activity detector* (VAD). A VAD returns the value true if there is speech present and the value false if no speech is present. Although VADs work generally well in low noise environments, as the noise level increases, the detection deteriorates and the noise can be improperly estimated resulting in poor enhancement.

Considering that wideband speech is a fairly new reality, most speech enhancement methods have been understandably proposed and tested for narrowband speech. Most of these approaches require some kind of extension for them to be used in wideband.

The methods overviewed in this chapter are noise suppression techniques used to reduce the effects of additive noise in speech (equation 2.2). Although many of them can be extended for different kinds of noise, we will mainly focus on white Gaussian noise. The speech enhancement techniques presented in this section constitute by no means an exhaustive list of what has been published but rather, is meant for completeness, to give an idea of the different main approaches used for speech enhancement. More detail is given to the methods relevant to this work, i.e. the weighted MMSE STSA (section 3.2.5). Otherwise, the omitted information can be found in the literature.

3.2 Speech Enhancement Techniques

3.2.1 Spectral Subtraction

Spectral subtraction is one of the simplest and oldest method for reducing the effect of noise in speech. The power spectrum of the clean speech is estimated by subtracting an estimated noise power spectrum from the noisy speech power spectrum. The noise power spectrum is estimated from non-speech periods in the noisy signal and therefore, it is supposed that the noise characteristics change slowly with time [1].

Rewriting equation 2.2 on page 6, we have

$$y(n) = s(n) + v(n) \qquad \text{(EQ 3.1)}$$

where $s(n)$ is the clean speech, $v(n)$ is the undesired noise, and $y(n)$ is the noisy speech. Taking the discrete-time Fourier transform (DTFT) of equation 3.1 results in

$$Y(\omega) = S(\omega) + V(\omega) \qquad \text{(EQ 3.2)}$$

The noise power spectrum estimate $|\hat{V}(\omega)|^2$ is updated during the periods of non-speech activity using the formula

$$|\hat{V}(\omega)|^2 = \Gamma_n|\hat{V}_{old}(\omega)|^2 + (1-\Gamma_n)|Y(\omega)|^2 \quad \text{(EQ 3.3)}$$

where $|Y(\omega)|^2$ is the power spectrum of the current frame, $|\hat{V}_{old}(\omega)|^2$ is the power spectrum of the previous noise estimate, and Γ_n is a decay factor. The estimated speech power spectrum $|\hat{S}(\omega)|^2$ is

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{V}(\omega)|^2, & \text{if } |Y(\omega)|^2 > |\hat{V}(\omega)|^2 \\ 0, & \text{otherwise} \end{cases} \quad \text{(EQ 3.4)}$$

The estimated speech is obtained from the following relationship

$$\hat{s}(n) = \text{IDTFT}\{|\hat{S}(\omega)|e^{j\angle Y(\omega)}\} \quad \text{(EQ 3.5)}$$

where $\angle Y(\omega)$ is the phase of the original noisy speech and IDTFT is the inverse discrete-time Fourier transform. The phase is left untouched because it is not so important for speech intelligibility and quality [46] and would complicate the process greatly.

Apart from the assumption of slowly varying noise characteristics, there are a few issues that limit the usability of spectral subtraction. The subtraction operation can result in negative power spectrum values, which are then set to zero. This introduces musical noise, as described earlier. Furthermore, this method makes use of a VAD to determine the periods of inactivity. If this activity detector is not working properly, noise characteristics will be poorly estimated.

3.2.2 Wiener Filtering

The Wiener filter is a linear filter whose transfer function is chosen to result in an optimal estimate of the clean speech in the *Mean-Square Error* (MSE) sense when corrupted by white Gaussian noise. Even though Wiener filtering can be used in the spectral subtrac-

tion sense (and results in the same speech estimation), its real strength comes from exploiting prior knowledge of the clean speech which can be expressed as a Gaussian mixture model, hidden Markov model, or others [1]. Theoretically, the transfer function of the filter is

$$H(\omega) = \frac{|S(\omega)|^2}{|S(\omega)|^2 + |V(\omega)|^2} \quad \text{(EQ 3.6)}$$

However, in practice, both the clean speech and noise power spectrums are unknown and can only be estimated. The transfer function is therefore estimated on a frame-by-frame basis from

$$H(\omega) = \frac{|\hat{S}(\omega)|^2}{|\hat{S}(\omega)|^2 + |\hat{V}(\omega)|^2} \quad \text{(EQ 3.7)}$$

The estimated speech can be retrieved by filtering and doing an inverse discrete-time Fourier transform

$$\hat{s}(n) = \text{IDTFT}\{H(\omega)Y(\omega)\} \quad \text{(EQ 3.8)}$$

In Wiener filtering, accurate estimation of the estimated speech is key. By using the power spectrum estimation given by spectral subtraction, we would get the identical result as spectral subtraction. Rather, statistical descriptions of speech models are used as the prior information. Wiener filtering also runs into similar problems as spectral subtraction and to try to circumvent some of these, a noise-scaling factor and a power exponent can be added to generalize the operations [1].

3.2.3 Spectral Subtraction Using Auditory Masking

As it was mentioned earlier, spectral subtraction results in fair amounts of musical noise. In [47], Virag has proposed a spectral subtraction approach using some auditory masking characteristics to try and convert the residual noise to something "perceptually white", something less annoying to listeners. Only simultaneous masking is considered in

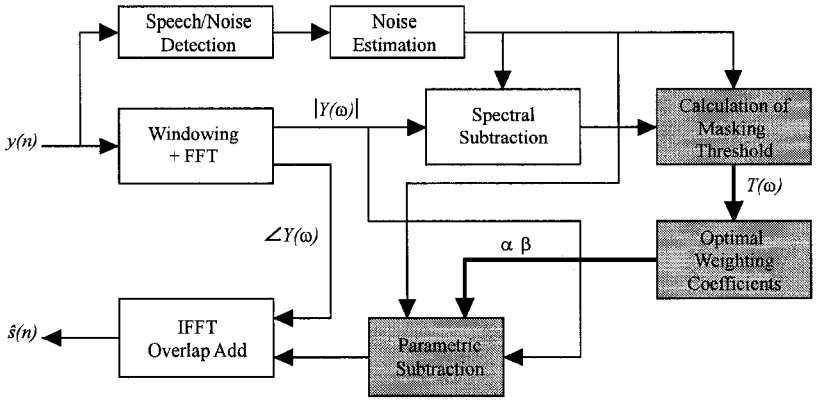

FIGURE 3.1 - Overview of perceptual spectral subtraction

this approach and the Johnston model is used (refer to section 2.5.4 on page 14). The generalized spectral subtraction method proposed by Berouti in [48] is used to have more control over the difference operation.

From [47], the perceptual enhancement scheme is shown in figure 3.1. The regular white boxes represent the simple spectral subtraction algorithm. The grey boxes are the perceptual aspects added to the original method. The main steps are:

1. Spectral Decomposition
2. Speech/noise detection and estimation of noise during speech pauses
3. Calculation of the noise masking threshold $T(\omega)$
4. Adaptation in time and frequency of the subtraction parameters α and β based on $T(\omega)$
5. Calculation of the enhanced speech spectral magnitude via parametric subtraction
6. Inverse transform.

Although the algorithm has been implemented for narrowband speech, the extension to wideband speech did not require much work, mainly the modification of the mapping from FFT bins to critical bands.

The spectral subtraction with auditory masking results in better speech enhancement because the subtraction parameters are adapted based on a criterion much more correlated with speech perception [47]. A trade-off between noise reduction, residual noise, and speech distortion is possible.

3.2.4 Minimum Mean-Square Error STSA

Many actual implementations of noise suppression systems in the real world are based on *Short Time Spectral Amplitude* (STSA) analysis. This is due in big part to its computational advantages [49]. The classic approach (and most popular) is the *Minimum Mean-Square Error* (MMSE) STSA proposed by Ephraim and Malah in 1984 [50] which minimizes the mean squared error of the short time spectral amplitude. Good noise suppression is achieved without the introduction of musical noise [51]. The equations from this section and the following one are rewritten from [49].

Again, we start from the noisy speech given by

$$y(n) = s(n) + v(n) \qquad \text{(EQ 3.9)}$$

where $s(n)$ is the clean speech and $v(n)$ is the undesired noise.

The speech signal is initially framed and windowed. The discrete Fourier transform is computed for each frame and is denoted by $Y_n(k)$ where n and k are the frame number and frequency bin index respectively. Modifications are done only to the spectral amplitude $|Y_n(k)|$ based on estimates of the *a priori* and *a posteriori* SNRs $\hat{\xi}_n(k)$ and $\hat{\gamma}_n(k)$ respectively. The a posteriori SNR is estimated by

$$\hat{\gamma}_n(k) = \frac{|Y_n(k)|^2}{\lambda_n(k)} \qquad \text{(EQ 3.10)}$$

where $\lambda_n(k)$ is obtained by averaging the power spectrum of the noisy speech in the first few frames of non-speech periods.

The a priori SNR is estimated by

$$\hat{\xi}_n(k) = \alpha\hat{\gamma}_{n-1}(k)G_{n-1}^2(k) + (1-\alpha)P[\hat{\gamma}_n(k)-1] \quad \text{(EQ 3.11)}$$

where α is a forgetting factor and *P[x]* is a rectifying function (i.e. negative values are set to zero). The spectral gain is computed with the estimated SNRs by

$$G_n(k) = \left[(1+c_n(k))I_0\left(\frac{c_n(k)}{2}\right) + c_n(k)I_1\left(\frac{c_n(k)}{2}\right)\right] \cdot \frac{\Lambda_n(k)}{1+\Lambda_n(k)}\frac{\sqrt{\pi c_n(k)}}{2\hat{\gamma}_n(k)}e^{-\frac{c_n(k)}{2}} \quad \text{(EQ 3.12)}$$

where $I_0(x)$ and $I_1(x)$ are the modified Bessel function of zero and first order respectively. $c_n(k)$ and $\Lambda_n(k)$ are defined by

$$c_n(k) = \frac{\eta_n(k)}{1+\eta_n(k)}\hat{\gamma}_n(k) \quad \text{(EQ 3.13)}$$

$$\Lambda_n(k) = \frac{1-q}{q}\frac{e^{v_n(k)}}{1+\eta_n(k)} \quad \text{(EQ 3.14)}$$

where

$$\eta_n(k) = \hat{\xi}_n(k)(1-q) \quad \text{(EQ 3.15)}$$

and where q is the probability of speech in that frame.

Finally, the enhanced speech spectrum is constructed by multiplying the spectral gain with the noisy spectrum

$$\hat{S}_n(k) = G_n(k)|Y_n(k)| \cdot e^{j\angle Y_n(k)} \quad \text{(EQ 3.16)}$$

and the enhanced speech $\hat{s}(n)$ is computed by taking the IFFT and doing an overlap and add operation.

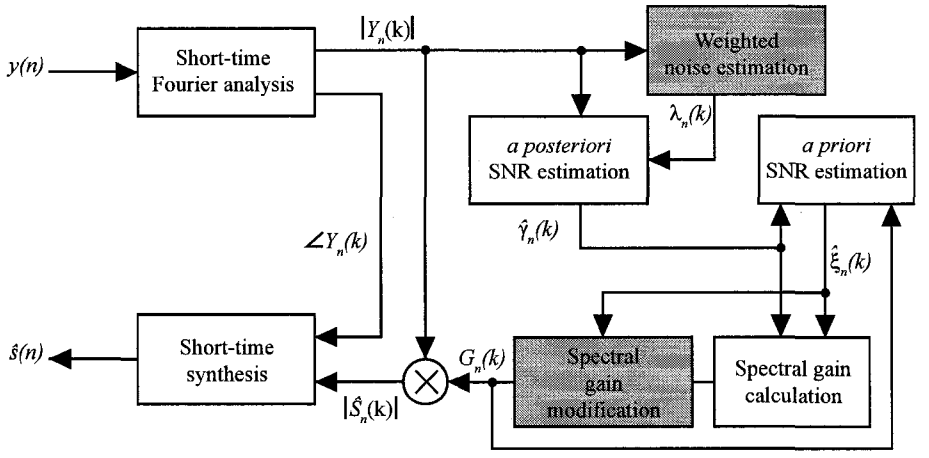

FIGURE 3.2 - Structure of the MMSE STSA based on weighted noise estimation

3.2.5 MMSE STSA Based on Weighted Noise Estimation

Kato et *al.* proposed a modification to the MMSE STSA approach by Ephraim. This method is implemented as noise suppressor in 3G handsets [49]. It improves the noise estimation and adds a spectral gain modification step to use this improved estimation more efficiently. Figure 3.2 shows an overview of the algorithm. It is almost identical to the structure of the MMSE STSA but the greyed out boxes have been added.

The *a posteriori* SNR is now computed using the noise estimate of the previous frame

$$\hat{\gamma}_n(k) = \frac{|Y_n(k)|^2}{\lambda_{n-1}(k)} \qquad \text{(EQ 3.17)}$$

To compute the estimate of the noise power spectrum, a weighting factor $W_n(k)$ based on a non-linear function (figure 3.3) is taken into account according to the SNR in the previous step, where $\tilde{\gamma}_1$, $\tilde{\gamma}_2$, and θ_z are constants. The resulting weighted noisy speech is given by

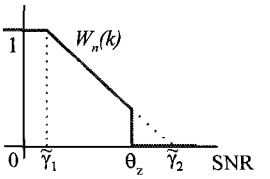

FIGURE 3.3 - Non-linear function

$$z_n(k) = W_n(k)|Y_n(k)|^2 \quad \text{(EQ 3.18)}$$

$$\lambda_n(k) = \frac{trace\{\boldsymbol{Z}_n(k)\}}{\psi(\boldsymbol{Z}_n(k))} \quad \text{(EQ 3.19)}$$

where

$$\boldsymbol{Z}_n(k) = \begin{cases} [z_n(k), \tilde{\boldsymbol{Z}}_{n-1}(k)], & n \le T_{init} \\ [z_n(k), \tilde{\boldsymbol{Z}}_{n-1}(k)], & \tilde{\gamma}_n(k) < \theta_z \\ \boldsymbol{Z}_n(k), & \text{otherwise} \end{cases} \quad \text{(EQ 3.20)}$$

$$\tilde{\boldsymbol{Z}}_0(k) = \boldsymbol{0}_{1\times(L_Z-1)} \quad \text{(EQ 3.21)}$$

$$\tilde{\boldsymbol{Z}}_n(k) = \boldsymbol{Z}_n(k)[\boldsymbol{I}_{L_Z-1}\boldsymbol{0}^T{}_{1\times(L_Z-1)}]^T \quad \text{(EQ 3.22)}$$

The *trace* function returns the sum of the diagonal elements and the $\Psi()$ function returns the number of non-zero elements.

Finally, the spectral gain is modified to reduce oversuppression for better enhanced speech quality according to

$$G_n(k) = \begin{cases} G_{\text{mod}}G_n(k), & \hat{\xi}_n(k) < \theta_G \\ G_n(k), & \text{otherwise} \end{cases} \quad \text{(EQ 3.23)}$$

and further modified as

$$G_n(k) = \begin{cases} G_n(k), & G_n(k) < G_{floor} \\ G_{floor}, & \text{otherwise} \end{cases} \quad \text{(EQ 3.24)}$$

where θ_G is the a priori SNR threshold, G_{mod} a scaling factor, and G_{floor} a flooring parameter. The enhanced speech is retrieved as it was done in the MMSE STSA case.

3.2.6 Hidden Markov Model

This class of speech enhancement methods rely on statistical models which are generated by training the algorithm with proper data. These statistical processes are called Hidden Markov Models (HMM) because it is assumed that the modelled system, in this case speech and noise, is of Markovian nature, i.e. that the conditional probability of future states depends only on the probability of the current state and not on past states. The actual system parameters are hidden and can only be estimated through observations.

Increase in signal-to-noise ratio can be achieved by re-estimating the clean speech through iterative generation of HMMs by maximizing an auxiliary function. Many similar methods can be found in the literature in [52] and [53]. The main disadvantages of HMMs are their necessity of training periods and their complexity.

3.2.7 Signal Subspace Decomposition

Signal subspace decomposition can be used to achieve noise reduction. An M-dimension space (from a vector built from M noisy speech samples) is decomposed into two subspaces: a signal subspace and a noise subspace [40]. Theoretically, the speech signal spans only the signal subspace while the noise component spans the entire space. The enhanced speech is estimated from the contents of the signal subspace. The orthogonality of the subspace allows a complete separation of the two components and, even though some noise remains in the signal space, it is greatly reduced.

The most common decomposition approach is the Karhunen-Loève (KL) expansion; it is the optimal basis for signal subspace decomposition [40]. The linear estimation is obtained by modifying the KL coefficients of the signal subspace according to an estimation criterion and by setting the other KL coefficients to 0. The estimated speech is obtained from the inverse KL transform.

Many subspace decomposition speech enhancement methods can be found in the literature. The readers are referred to [54] and [55]. Interestingly, since the KL transform and the DFT are closely related, spectral subtraction can be viewed as a form of signal subspace decomposition.

3.2.8 Kalman Filtering

The Kalman filter is a recursive filter which estimates the state of a dynamic system from noisy measurements in a way that minimizes the mean of the squared error. It can be used for estimation of past, present, and future states and was first proposed by R. E. Kalman in 1960 in his famous paper [56].

The Kalman filter is a set of mathematical equations that implements predictor-corrector type estimator, that is optimal in the sense that it minimizes the estimated error covariance [57] when the following three conditions are met: 1) the system can be described through a linear system, 2) the system and measurement noises are white, and 3) the system and measurement noises are of Gaussian distribution.

Since the work of this thesis mostly focuses on the Kalman filter, it will be covered in great length in the following chapter. The justification for using it as the noise reduction approach is that it is a fairly simple algorithm that allows good enhancement properties without the introduction of musical noise and without the use of a voice activity detector. Furthermore, previous work in our group had indicated some good potential for Kalman filters warranting further investigation. Finally, since the main focus of this thesis is noise reduction in white Gaussian noise, a condition very suitable, a priori, for the Kalman filter, we feel our choice is justified.

CHAPTER 4

KALMAN FILTERING BASED METHODS FOR WIDEBAND SPEECH ENHANCEMENT

4.1 Introduction

The Kalman filter was first proposed by R. E. Kalman in 1960 as a recursive solution to the discrete-data linear filtering problem. The output of the filter is the optimal estimation of the state of a stochastic process and is obtained by minimizing the mean of the squared error [57]. Due to the increase of computing power by computers, Kalman filtering has been extensively researched and used for many applications, mainly in the area of autonomous or assisted navigation [57] but definitely not restricted to that field.

Kalman filtering applied to speech enhancement has been first proposed by Paliwal and Basu in 1987 [58]. This algorithm was used to reduce the effect of white Gaussian noise in narrowband speech. In this implementation, the various parameters required by the filter are derived from the clean and the noise-only data resulting in an upper-bound of the enhancement process. In real life, these parameters are not known and can only be estimated but nevertheless, it gave an encouraging start. Many other papers have been published on the subject in which: 1) techniques for parameter estimation, 2) extension to colored noise or other types of noise, or 3) modification to the original algorithm, are proposed [59], [60], [3]-[7].

Ma has published a series of articles on speech enhancement for narrowband speech corrupted by white noise and narrowband speech corrupted by colored noise [3]-[7]. Her work is based upon [60] but she modifies the algorithm by making use of the perceptual properties of the human auditory system to improve the enhanced speech further. In [3], the update equations are truly constrained by the perceptual auditory masking resulting in an effective but computationally intensive algorithm. A heuristic approach is also proposed and used to reduce the computational load but still keep good speech enhancement. A perceptual auditory post-filter is used at the output of the Kalman filter. Other methods using neural networks have been proposed [7].

The work of this thesis is based on the heuristic aspect of [3]. The rationale for basing the work on this approach is that it has demonstrated good potential for speech enhancement in the narrowband case. Furthermore, it is as fast as other Kalman filtering schemes (i.e. the post-filter does not add too much computation). Some modifications have been made to the algorithm to extend it to the wideband case and to fine-tune and improve some aspects for better enhancement.

This chapter presents the proposed Kalman filtering algorithm with the various modifications starting with an overview of the filter.

4.2 Kalman Filter Algorithm

The Kalman filter algorithm deals with randomly time-varying signals or random processes. The random signal to be estimated can be modelled as a first-order recursive process driven by zero-mean white noise [61]. The signal evolves according to the dynamic equation

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}_k \boldsymbol{x}_k + \boldsymbol{G}_k \boldsymbol{w}_k \qquad \text{(EQ 4.1)}$$

where $\boldsymbol{x}_k$ is the state of the model at discrete time k (i.e. iteration k), $\boldsymbol{F}_k$ is a transition matrix that takes the state from time k to $k+1$, and $\boldsymbol{G}_k$ is a gain matrix. The process noise $\boldsymbol{w}_k$ is specified by

$$E[\boldsymbol{w}_k] = \boldsymbol{0}$$
$$E[\boldsymbol{w}_n \boldsymbol{w}_k^T] = \begin{cases} \boldsymbol{0} & \text{for } n \neq k \\ \boldsymbol{Q}_k & \text{for } n = k \end{cases} \qquad \text{(EQ 4.2)}$$

where the $E[x]$ operator denotes the expected value and T denotes matrix transposition. The dimension of the state is M (i.e. $\boldsymbol{x}_k$ is a $M \times 1$ matrix).

It is not possible to measure the states directly. The state observations are made through the following measuring equation

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{v}_k \qquad \text{(EQ 4.3)}$$

where $\boldsymbol{y}_k$ is the observed state, $\boldsymbol{H}_k$ is the measurement matrix, and $\boldsymbol{v}_k$ is the measurement noise. This noise is assumed to be additive zero-mean white Gaussian defined by

$$E[\boldsymbol{v}_k] = \boldsymbol{0}$$
$$E[\boldsymbol{v}_n \boldsymbol{v}_k^T] = \begin{cases} \boldsymbol{0} & \text{for } n \neq k \\ \boldsymbol{R}_k & \text{for } n = k \end{cases} \qquad \text{(EQ 4.4)}$$

where the dimension of the state $\boldsymbol{y}_k$ is N.

The objective of the Kalman filter is to estimate the state x_k by minimizing the mean-square error by using the observation data $y_1, y_2, ..., y_k$. The initial values are assumed to be zero (i.e. $x_k = 0$ and $w_k = 0$ for $k < 0$). The model is shown in figure 4.1.

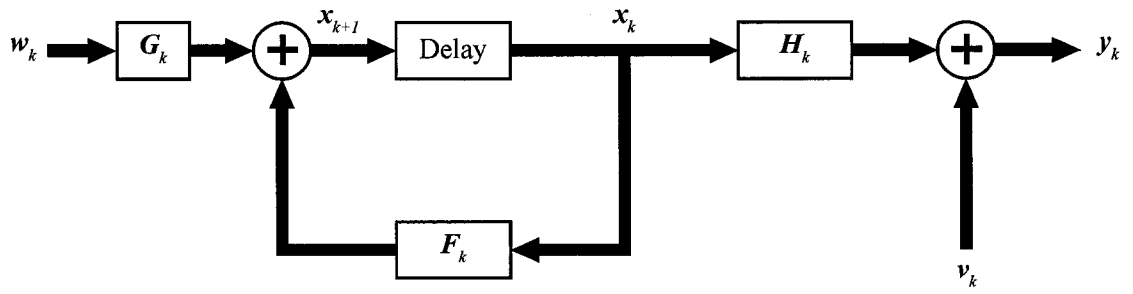

FIGURE 4.1 - Model of linear system given by equation 4.1 and equation 4.3

The state estimate at time k, $\hat{x}_k$, is computed from the previous state and the current observation vector; it therefore contains all of the system's information up until time k

$$\hat{x}_k = A_k \hat{x}_{k-1} + B_k y_k \qquad \text{(EQ 4.5)}$$

where the matrices A_k and B_k must be determined to minimize the mean-square error. The state estimate $\hat{x}_k$ is an unbiased estimate of the true state x_k and the state-error vector can be used to simplify the equations. This can be used to find a relationship between A_k and B_k

$$\tilde{x}_k = \hat{x}_k - x_k \qquad \text{(EQ 4.6)}$$

$$E[\tilde{x}_k] = 0 \qquad \text{(EQ 4.7)}$$

Replacing $\hat{x}_k$ by equation 4.5, x_k by equation 4.1, and y_k by equation 4.3

$$
\begin{aligned}
&E[A_k \hat{x}_{k-1} + B_k y_k - x_k] = 0 \\
&E[A_k \hat{x}_{k-1} + B_k (H_k x_k + v_k) - (F_{k-1} x_{k-1} + w_{k-1})] = 0 \\
&E[A_k(\hat{x}_{k-1} - x_{k-1}) + A_k x_{k-1} - F_{k-1} x_{k-1} - G_k w_{k-1} + B_k(H_k(F_{k-1} x_{k-1} + G_k w_{k-1}) + v_k)] = 0 \\
&E[A_k \tilde{x}_{k-1} + (B_k H_k F_{k-1} - F_{k-1} + A_k) x_{k-1} + (B_k H_k - I) G_k w_{k-1} + B_k v_k] = 0 \\
&(B_k H_k F_{k-1} - F_{k-1} + A_k) E[x_{k-1}] = 0
\end{aligned}
\qquad \text{(EQ 4.8)}
$$

since $E[\tilde{x}_{k-1}]$, $E[w_{k-1}]$, and $E[v_k]$ are all equal to 0. Knowing that $E[x_{k-1}]$ is not necessarily 0, we have the following relationship

$$A_k = (I - B_k H_k) F_{k-1} \qquad \text{(EQ 4.9)}$$

Replacing A_k in equation 4.5 and simplifying, the a posteriori estimate of the state at time k is expressed as

$$
\begin{aligned}
\hat{x}_k &= F_{k-1} \hat{x}_{k-1} + B_k (y_k - H_k F_{k-1} \hat{x}_{k-1}) \\
\hat{x}_k &= \hat{x}_k^- + B_k (y_k - H_k \hat{x}_k^-)
\end{aligned}
\qquad \text{(EQ 4.10)}
$$

where $\hat{x}_k^- = F_{k-1} \hat{x}_{k-1}$ is the a priori state estimate at time k. This estimate is based on all the observations up until and including time k-1 (but not time k) and it is the best estimate of x_k without the new current observations.

The matrix variable B_k is known as the Kalman gain and must be solved to minimize the mean-square error (or variance) of the estimate error $\tilde{x}_k$. The cost-function is defined as the estimation error covariance matrix

$$P_k = E[\tilde{x}_k \tilde{x}_k^T] \qquad \text{(EQ 4.11)}$$

Furthermore, the prediction error covariance matrix is defined similarly

$$P_k^- = E[\tilde{x}_k^- \tilde{x}_k^{-T}] \qquad \text{(EQ 4.12)}$$

where $\tilde{x}_k^- = \hat{x}_k^- - x_k$.

Starting from equation 4.12 and expanding, we get

$$\begin{aligned} \boldsymbol{P}_k^- &= E[(\hat{\boldsymbol{x}}_k^- - \boldsymbol{x}_k)(\hat{\boldsymbol{x}}_k^- - \boldsymbol{x}_k)^T] \\ &= E[(\boldsymbol{F}_{k-1}\hat{\boldsymbol{x}}_{k-1} - \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} - \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1})(\boldsymbol{F}_{k-1}\hat{\boldsymbol{x}}_{k-1} - \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} - \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1})^T] \\ &= E[(\boldsymbol{F}_{k-1}\tilde{\boldsymbol{x}}_{k-1} - \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1})(\boldsymbol{F}_{k-1}\tilde{\boldsymbol{x}}_{k-1} - \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1})^T] \\ &= \boldsymbol{F}_{k-1}E[\tilde{\boldsymbol{x}}_{k-1}\tilde{\boldsymbol{x}}_{k-1}^T]\boldsymbol{F}_{k-1}^T + \boldsymbol{G}_{k-1}E[\boldsymbol{w}_{k-1}\boldsymbol{w}_{k-1}^T]\boldsymbol{G}_{k-1}^T \\ &= \boldsymbol{F}_{k-1}\boldsymbol{P}_{k-1}\boldsymbol{F}_{k-1}^T + \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{G}_{k-1}^T \end{aligned} \quad \text{(EQ 4.13)}$$

From this result, an equation can be derived for $\boldsymbol{P}_k$. First, the estimation-error is developed using equation 4.3 and equation 4.10 as such

$$\begin{aligned} \tilde{\boldsymbol{x}}_k &= \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k \\ &= \hat{\boldsymbol{x}}_k^- + \boldsymbol{B}_k(\boldsymbol{y}_k - \boldsymbol{H}_k\hat{\boldsymbol{x}}_k^-) - \boldsymbol{x}_k \\ &= (\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)\hat{\boldsymbol{x}}_k^- + \boldsymbol{B}_k(\boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k) - \boldsymbol{x}_k \\ &= (\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)(\hat{\boldsymbol{x}}_k^- - \boldsymbol{x}_k) + \boldsymbol{B}_k\boldsymbol{v}_k \\ &= (\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)\tilde{\boldsymbol{x}}_k^- + \boldsymbol{B}_k\boldsymbol{v}_k \end{aligned} \quad \text{(EQ 4.14)}$$

Then, the estimation error covariance matrix is built from

$$\begin{aligned} \boldsymbol{P}_k &= E[\tilde{\boldsymbol{x}}_k\tilde{\boldsymbol{x}}_k^T] \\ &= (\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)E[\tilde{\boldsymbol{x}}_k^-\tilde{\boldsymbol{x}}_k^{-T}](\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)^T + \boldsymbol{B}_k E[\boldsymbol{v}_k\boldsymbol{v}_k^T]\boldsymbol{B}_k^T \\ &= (\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)\boldsymbol{P}_k^-(\boldsymbol{I} - \boldsymbol{B}_k\boldsymbol{H}_k)^T + \boldsymbol{B}_k\boldsymbol{R}_k\boldsymbol{B}_k^T \end{aligned} \quad \text{(EQ 4.15)}$$

Since we aim to minimize the estimation error variance, the trace of $\boldsymbol{P}_k$ is differentiated with respect to $\boldsymbol{B}_k$. $\boldsymbol{P}_k$ has been further expanded for this operation (where the subscripts and superscript have been left out to save space)

$$\begin{aligned} \mathrm{Tr}[\boldsymbol{P}] &= \mathrm{Tr}[\boldsymbol{P} - \boldsymbol{BHP} - \boldsymbol{PH}^T\boldsymbol{B}^T + \boldsymbol{BHPH}^T\boldsymbol{B}^T + \boldsymbol{BRB}^T] \\ &= \mathrm{Tr}[\boldsymbol{P}] - 2\mathrm{Tr}[\boldsymbol{BHP}] + \mathrm{Tr}[\boldsymbol{BHPH}^T\boldsymbol{B}^T] + \mathrm{Tr}[\boldsymbol{BRB}^T] \\ \frac{\partial \mathrm{Tr}[\boldsymbol{P}]}{\partial \boldsymbol{B}} &= -2\boldsymbol{PH}^T + 2\boldsymbol{KHPH}^T + 2\boldsymbol{KR} \end{aligned} \quad \text{(EQ 4.16)}$$

Setting the derivative to zero, we obtain the Kalman gain

$$\boldsymbol{B}_k = \boldsymbol{P}_k^-\boldsymbol{H}_k^T(\boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^T + \boldsymbol{R}_k)^{-1} \quad \text{(EQ 4.17)}$$

And, finally, to obtain the final simplified form for $\boldsymbol{P}_k$, equation 4.17 is inserted back into equation 4.15. Again, the subscripts and superscripts are omitted for simplicity

$$\begin{aligned} \boldsymbol{P}_k &= \boldsymbol{P} - \boldsymbol{BHP} - \boldsymbol{PH}^T\boldsymbol{B}^T + \boldsymbol{BHPH}^T\boldsymbol{B}^T + \boldsymbol{BRB}^T \\ &= \boldsymbol{P} - \boldsymbol{BHP} - (\boldsymbol{PH}^T - \boldsymbol{BHPH}^T - \boldsymbol{BR})\boldsymbol{B}^T \\ &= \boldsymbol{P} - \boldsymbol{BHP} - \boldsymbol{PH}^T(\boldsymbol{I} - (\boldsymbol{HPH}^T + \boldsymbol{R})^{-1}(\boldsymbol{HPH}^T + \boldsymbol{R}))\boldsymbol{B}^T \\ &= \boldsymbol{P}_k^- - \boldsymbol{B}_k\boldsymbol{H}_k\boldsymbol{P}_k^- \end{aligned} \qquad \text{(EQ 4.18)}$$

The Kalman filter equations are summarized in figure 4.2.

$$\begin{aligned} \boldsymbol{B}_k &= \boldsymbol{P}_k^- \boldsymbol{H}_k^T(\boldsymbol{H}_k\boldsymbol{P}_k^- \boldsymbol{H}_k^T + \boldsymbol{R}_k)^{-1} \\ \hat{\boldsymbol{x}}_k &= \hat{\boldsymbol{x}}_k^- + \boldsymbol{B}_k(\boldsymbol{y}_k - \boldsymbol{H}_k\hat{\boldsymbol{x}}_k^-) \\ \boldsymbol{P}_k &= \boldsymbol{P}_k^- - \boldsymbol{B}_k\boldsymbol{H}_k\boldsymbol{P}_k^- \\ \hat{\boldsymbol{x}}_{k+1}^- &= \boldsymbol{F}_k\hat{\boldsymbol{x}}_k \\ \boldsymbol{P}_{k+1}^- &= \boldsymbol{F}_k\boldsymbol{P}_k\boldsymbol{F}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^T \end{aligned}$$

FIGURE 4.2 - Kalman filter algorithm

4.3 Application to Speech Enhancement

Now that the Kalman filter algorithm has been developed and presented, it will be shown how it can be applied to speech enhancement [60].

Firstly, the clean speech signal *s(n)* is modeled as an auto-regressive process as it was presented in section 2.4.2

$$\begin{aligned} s(n) &= \sum_{i=1}^{p} a_i(n)s(n-i) + u(n) \\ y(n) &= s(n) + v(n) \end{aligned} \qquad \text{(EQ 4.19)}$$

where $a_i(n)$ is the i-th filter coefficient, p is the order of the model, and $u(n)$ and $v(n)$ are the process and measurement noises as presented in the previous section. The time-index n

of the filter coefficients implies that these are not constant and will vary to model the non-stationarity characteristics of speech.

The system is further modeled as a state-space model

$$\begin{aligned} \boldsymbol{x}(n) &= \boldsymbol{F}(n)\boldsymbol{x}(n-1) + \boldsymbol{G}u(n) \\ y(n) &= \boldsymbol{H}\boldsymbol{x}(n) + v(n) \end{aligned} \qquad \text{(EQ 4.20)}$$

where :

1. $\boldsymbol{x}(n)$ is the $p \times 1$ state vector

$$\boldsymbol{x}(n)_{p\times 1} = \left[s(n-p+1)\ s(n-p+2)\ \ldots\ s(n)\right]^T,$$

2. $\boldsymbol{F}(n)$ is the $p \times p$ transition matrix

$$\boldsymbol{F}(n)_{p\times p} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ a_p(n) & a_{p-1}(n) & a_{p-2}(n) & \ldots & a_1(n) \end{bmatrix},$$

3. $\boldsymbol{G}$ and $\boldsymbol{H}$ are the $p \times 1$ input vector and the $1 \times p$ observation row vector respectively

$$\boldsymbol{H}_{1\times p} = \boldsymbol{G}^T_{p\times 1} = \left[0\ 0\ \ldots\ 1\right],$$

4. $u(n)$ and $v(n)$ are the process and measurement noises respectively; zero-mean White Gaussian noise sequences with variance σ_u^2 and σ_v^2.

The Kalman filter algorithm is shown again below with the new variable definitions and a slight modification to simplify the notation

$$\begin{aligned}
\boldsymbol{K}(n)_{p\times 1} &= \boldsymbol{P}(n|n-1)\boldsymbol{H}^T[\boldsymbol{H}\boldsymbol{P}(n|n-1)\boldsymbol{H}^T+\sigma_v^2]^{-1} \\
\hat{\boldsymbol{x}}(n|n)_{p\times 1} &= \hat{\boldsymbol{x}}(n|n-1)+\boldsymbol{K}(n)[y(n)-\boldsymbol{H}\hat{\boldsymbol{x}}(n|n-1)] \\
\boldsymbol{P}(n|n)_{p\times p} &= [\boldsymbol{I}-\boldsymbol{K}(n)\boldsymbol{H}]\boldsymbol{P}(n|n-1) \\
\hat{\boldsymbol{x}}(n+1|n)_{p\times 1} &= \boldsymbol{F}(n)\hat{\boldsymbol{x}}(n|n) \\
\boldsymbol{P}(n+1|n)_{p\times p} &= \boldsymbol{F}(n)\boldsymbol{P}(n|n)\boldsymbol{F}(n)^T+\boldsymbol{G}\boldsymbol{G}^T\sigma_u^2
\end{aligned} \qquad \text{(EQ 4.21)}$$

The covariance matrices $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ have become scalars because of the definitions of $\boldsymbol{H}$ and $\boldsymbol{G}$.

Normal Kalman Filter

The actual estimated speech sample is usually retrieved from the state vector by taking its last element

$$\hat{s}(n) = \boldsymbol{H}\hat{\boldsymbol{x}}(n|n) \qquad \text{(EQ 4.22)}$$

At time n, we have an estimated speech value for time n.

Delayed Kalman Filter

In [58], it was suggested that a better estimate could be obtained at time n by taking the first element of the state vector instead, $\hat{s}(n-p+1)$. This means that at instant n, the speech sample at time *n-p+1* is computed. We have a better estimation because the latter is based upon *p-1* extra observations (*y(k-p+2), ..., y(k)*). Although, this approach introduces a delay of *p-1* iterations, which translates to 1.1875 ms, this delay is short enough to be acceptable.

This is the preferred estimation for this work and it differs from most publications on the subject.

4.4 Auditory Masking

Based on [29], an auditory post-filter is used after the Kalman filtering operation to further improve the perceptual quality of the enhanced speech. This filter is based upon the

simultaneous and forward temporal masking approach presented in section 2.5 on page 12. Since it was presented in great length, only a few implementation details will be specified here.

4.4.1 *Windowing*

The speech signal is filtered by the Kalman filter in its entirety before it is treated by the perceptual post-filter. Therefore, at the input of the perceptual filter, we have a speech signal which is enhanced. We are trying to enhance it further by using the perceptual filter. The signal is separated into frames of 320 samples long (20 ms) with 50% overlap and windowed by a simple rectangular window.

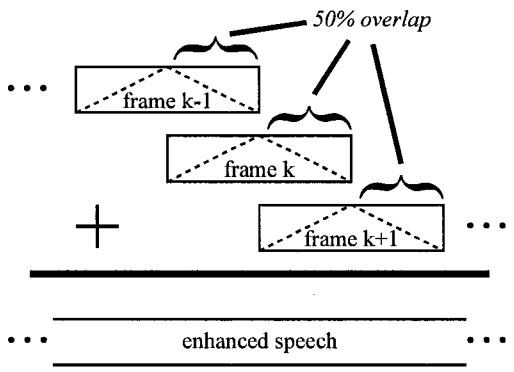

FIGURE 4.3 - Unwindowing operation

Each frame is processed separately (although the loudness from the previous frame is used) by the post-filter and the perceptually enhanced is "unwindowed" to generate the finalized enhanced speech. The overlapped sections are recombined by using a weighted linear average as shown by the dotted line in figure 4.3. By multiplying the frames by a triangular window and adding them with the proper time offset, the averaging operation is carried out and allows a smooth transition between the frames.

4.4.2 *Power Spectrum*

The computation of the power spectrum is a fairly straightforward operation. It is the scaled squared-amplitude of a N-point discrete Fourier transform and is given in the following equation

$$P_{xx}(n) = \begin{cases} \frac{1}{N}|X(n)|^2, \text{ for } n = 0 \text{ and } n = \frac{N}{2} \\ \frac{2}{N}|X(n)|^2, \text{ for } n = 1, \ldots, \frac{N}{2} - 1 \end{cases} \quad \text{(EQ 4.23)}$$

where $X(n)$ is the N-point discrete Fourier transform. The energy is doubled for the time index $n = 1,...,N/2-1$ because when we combine the complex exponentials with their matching exponentials (indexes $N/2+1,...,N-1$) to form actual sine waves, the amplitude is increased and ends up doubling the power. In this work, N has been chosen to be 512 resulting in a power spectrum vector of length 257.

4.4.3 Excitation Functions

In section 2.5.5, an equation for the specific loudness was presented. This equation makes use of 3 different functions: the excitation of the power spectral density of the current frame, the excitation at the absolute threshold of hearing, and the excitation for 0 dB SPL.

1. **Excitation of power spectral density - *E(i)***

   This excitation is in fact the function computed in equation 2.14, i.e. the energy present in each critical band convolved with the spreading function (equation 2.12).

2. **Excitation at threshold of hearing - $E_T(i)$**

   This function is computed by first computing the power spectrum of the absolute threshold of hearing (equation 2.10), computing the energy in each band according to equation 2.13, and then convolving with the spreading function.

3. **Excitation at 0 dB SPL - E_0**

   A sound pressure level of 0 dB is by definition the faintest sound perceivable and is set to 20 μPa for a 1 kHz sine wave. To compute this excitation, a 1 kHz sine wave is generated and calibrated to be at 0 dB SPL (see more details in section 4.7.1). The power spectrum of this signal is computed and convolved with the spreading function as it was done for the other excitation functions. Since the total excitation at 0

dB SPL is required and that is not a function of the critical band i, the final value is obtained by adding the energy level of each critical band.

4.5 LP Coefficients Estimation

An important aspect of the Kalman filter algorithm for speech enhancement is the estimation of the model parameters. This step is crucial because poor modelling degrades the enhanced speech quality.

The LP coefficients are computed using the well-known Levinson-Durbin algorithm. This is an efficient recursive algorithm which solves the normal equations [17]

$$\sum_{k=0}^{p} a_k \gamma_{ss}(l-k) = 0 \quad l = 1, 2, \ldots, p \qquad \text{(EQ 4.24)}$$

where

$$\gamma_{ss}(k) = E[s(n)s(n-k)] \qquad \text{(EQ 4.25)}$$

is the auto-correlation of the speech signal. The algorithm outputs the filter coefficients which minimize the mean-square error of the residual (equation 2.8 and section 2.4.2).

To compute the LP coefficients, the speech data used (refer to following sections for various cases) is first separated into frames of 320 samples (20 ms). It is assumed that this is short enough to ensure speech quasi-stationarity. These frames are then multiplied by a von Hann window to reduce the effect of the edge of the frames. The von Hann window is given by

$$w(n) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N+1}\right)\right), \quad n = 1, \ldots, N \qquad \text{(EQ 4.26)}$$

where N is the length of the window (320 in our case). This form might look slightly different from other von Hann function definitions, but this form does not include the zero-weighted window samples at each end.

Once the speech is windowed, the Levinson-Durbin algorithm is applied separately to each frame resulting in a distinct set of LP coefficients for each of them. These sets serve as input to the Kalman filter and are used in the transition matrix $\boldsymbol{F}(n)$. In this case, the time index n refers to the frame number. After the filter has finished processing a frame, before it starts processing the next one, the transition matrix is updated to reflect the AR properties of the new speech frame. The filtering operation of the algorithm smooths out small jumps that could be encountered at the boundaries of frames.

4.5.1 Choice of p

The dimension of the AR system is an important variable and must be chosen carefully. There is a trade-off between the model precision and the model complexity. A small value results in a small and simple system which do not model the data very well. A higher value for p makes the model more complex but improves its precision. Referring to equation 4.21, we see that the Kalman filter algorithm involves many matrix operations. The increase of dimension increases the amount of computation required in each step and lengthens the time of execution.

In most of the narrowband speech enhancement work involving Kalman filtering, the dimension of the model is set to 10. The choice of this value can be justified by the fact that 2 poles are required to model a formant. With a dimension of 10, up to 5 formants can be modelled giving a decent representation of narrowband speech.

In this thesis, the dimension used is 20. The rationale behind this choice is that since we are doubling the frequency covered by the model by extending from narrowband speech to wideband speech, more poles are required to keep a good quality. However, the increase in quality comes at a price of complexity. Nevertheless, it was found that a dimension of 20 was still of acceptable complexity. Figure 4.4 shows some results where the value of p is

modified. These results are based on an average of 5 enhanced sample files which were corrupted at 10 dB SNR and serve to justify the choice of p.

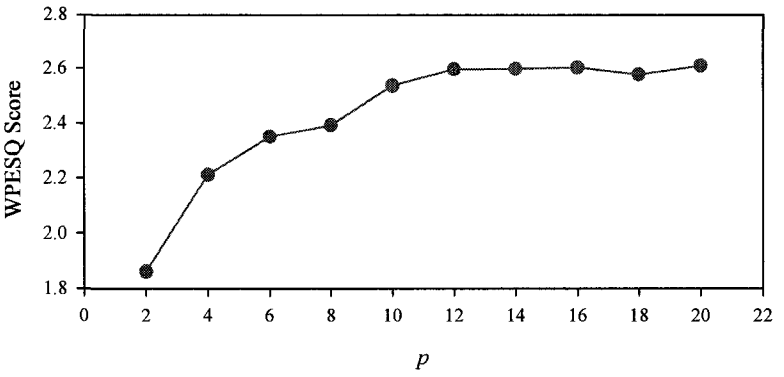

FIGURE 4.4 - Effect of varying p

4.5.2 Optimal Estimation

When the model parameters are computed from the clean data, the speech is properly modelled by an auto-regressive system. Unfortunately, the addition of noise degrades the estimation of the filter coefficients of the underlying process. In the first case, the coefficients are computed from the speech prior to noise corruption. This results in a set of parameters that are optimal. These coefficients are then used in the Kalman filter algorithm to enhance the noisy speech. This situation is obviously not possible in real life because we do not have access to the clean speech information to estimate the parameters. However, the results obtained still give us valid information on the capabilities of the algorithm by providing some sort of upper-bound on what can be achieved. The results can be found in section 5.3 and section 5.4.

4.5.3 Noisy Estimation

Estimating the LP coefficients from the noisy speech results in a worse quality. Even though the AR system is supposed to model speech, a fair amount of noise is modelled

into the AR parameters making it hard to extract it during the filtering operation. The results can be found in section 5.3 and section 5.4.

4.5.4 Iterative Estimation

This approach aims at improving the estimation of the LP coefficients by making iterative runs with the Kalman filter. Initially, the coefficients are computed from the noisy speech and used in the Kalman filter. The resulting enhanced speech is then used to compute a new and improved set of coefficients. The speech is then refiltered using this new set of parameters resulting in a better enhanced speech. The filter's input speech is always the original noisy data and not the enhanced speech from the previous iteration. To reduce the amount of distortion, the perceptual filter is only used during the last iteration where the LP coefficients do not need to be estimated from the enhanced speech.

Although many iterations can be done, one must be careful not to perform too many as a high number of filtering operations tends to introduce distortion in the resulting enhanced speech, not to mention the added complexity of running the Kalman filter algorithm multiple times.

In this work, the number of iterations has been fixed to 2. Adding this extra iteration to the noisy case can improve the quality of the speech but doubles the complexity. Performing more than 2 iterations introduces speech distortion which is probably caused by the limitations of the AR model and the accuracy of the MMSE estimate. The results can be found in section 5.3 and section 5.4.

4.5.5 MMSE STSA Based on Weighted Noise Estimation

The adequate estimation of the LP coefficients is truly an important step of the algorithm. By using another speech enhancement technique to get an initial enhancement, the estimation of the LP coefficients can be improved.

The MMSE STSA based on weighted noise estimation was presented in section 3.2.5. It is a low-complexity speech enhancement method that works well although distinct residual noise remains in the enhanced speech. By estimating the LP coefficients from this noise-reduced speech, it is hoped that further speech quality improvement will be possible. The results can be found in section 5.3 and section 5.4.

4.5.6 Note on Other Approaches

There are many published methods that aim to reduce the effect of noise in the estimation of AR parameters [62], [63], [64]. Some of these methods were implemented to test their adequacy. It was found that none of those implemented improved the results of the LP coefficients. We suspect that this discrepancy is caused by the fact that speech is not a purely AR process, unlike the processes normally considered in these references.

4.6 Variance Estimation

The Kalman algorithm requires the statistics of the driving process and the additive measurement noise. The actual values are not known and need to be estimated during the filtering operation. The processes $u(n)$ and $v(n)$ are assumed to be zero-mean white Gaussian sequences. Therefore, only their variances need to be estimated.

A variance estimation technique is proposed in [60] and has also been used in [3]. The variance is computed by using an estimate of the noise or driving process at each step and averaging over a certain amount of samples. By keeping this number high enough to get a good representation of the noise variance but low enough to allow tracking of the noise statistics if they happened to change, an adequate estimation can be obtained.

In order to obtain the best results, it is important that the estimated variances be close to the actual values. An over-estimation will cause speech distortion as the Kalman filter removes more noise than necessary. An under-estimation will leave hearable residual noise.

In this work, the variance of the measurement noise is estimated according to the approach suggested in [60]. However, the estimation of the driving process statistics is done differently. The adopted approach is simpler and results in good estimates. More details follow.

4.6.1 Driving Process Variance

Referring back to section 2.4.2 on speech modelling and comparing equation 2.5 and equation 2.8, we see that the residual *e(n)* is the same as *Gu(n)*. And, as it was mentioned, the residual should be similar to an impulse train for voiced speech and noise-like for unvoiced speech. Even though the Kalman filter algorithm requires that this process be white and Gaussian and that it is not always necessarily the case, the modelling still ends up being adequate as demonstrated in the many papers where the Kalman filter has been applied to speech enhancement. By computing the variance of the residual that we assume to be zero-mean white Gaussian (because of the Kalman filter), an estimate of σ_u^2 can be obtained.

Once the LP coefficients are computed, the speech *s(n)* is filtered by the analysis filter *A(z)* to obtain *e(n)* (equation 2.9). The variance of *e(n)* is computed as

$$\hat{\sigma}_u^2 = \frac{1}{N-1}\sum_{n=0}^{N-1}[e(n)-\bar{e}]^2$$
$$\text{where } \bar{e} = \frac{1}{N}\sum_{n=0}^{N-1} e(n) \qquad \text{(EQ 4.27)}$$

where N is the number of elements in the residual (which is 320 in this case) and $\bar{e}$ is the mean of the residual. Although we do assume that *u(n)* has a mean of zero for the Kalman filter, the actual mean is included in the computation of the variance to get a better estimate of the actual value.

Each set of LP coefficients have their own variance estimate and these are updated in the Kalman filter algorithm at the end of a frame, just like it was done with the filter parameters in the transition matrix.

4.6.2 *Measurement Noise Variance*

As opposed to the driving process variance, the measurement noise is computed on the fly. It is derived under the assumption of a constant mean and variance over N samples, *v(n), v(n-1), ..., v(n-N+1)* [60].

From equation 4.20, we have

$$v(n) = y(n) - \boldsymbol{H}\boldsymbol{x}(n) \qquad \text{(EQ 4.28)}$$

However, the actual state ***x(n)*** can only be estimated and the best option at time n is to use the a priori state estimation $\hat{x}(n|n-1)$ computed during the previous step (since we need the variance to compute the a posteriori state estimate). The measurement noise *v(n)* is estimated by $\alpha(n)$

$$\begin{aligned} \alpha(n) &= y(n) - \boldsymbol{H}\hat{\boldsymbol{x}}(n|n-1) \\ &= \boldsymbol{H}\tilde{\boldsymbol{x}}(n|n-1) + v(n) \end{aligned} \qquad \text{(EQ 4.29)}$$

where $\tilde{x}(n|n-1) = x(n) - \hat{x}(n|n-1)$ is the predicted state error.

The mean and variance of $\alpha(n)$ are computed from the last N samples, $\alpha(n)$, $\alpha(n-1)$, ..., $\alpha(n-N+1)$. The estimation of the mean is

$$\hat{\bar{\alpha}}(n) = \frac{1}{N}\sum_{i=0}^{N-1}\alpha(n-i) \qquad \text{(EQ 4.30)}$$

An unbiased estimator of the variance is

$$\hat{\sigma}_{\alpha}^{2} = \frac{1}{N-1}\sum_{i=0}^{N-1}[\alpha(n-i) - \hat{\bar{\alpha}}(n)]^{2} \qquad \text{(EQ 4.31)}$$

Taking the expected value of equation 4.31, we have

$$\begin{aligned} E[\sigma_\alpha^2] &= \frac{1}{N-1}\sum_{i=0}^{N-1} E[\alpha(n-i)-\hat{\bar{\alpha}}(n)]^2 \\ &= \frac{1}{N-1}\sum_{i=0}^{N-1} (E[\alpha^2(n-i)]-2E[\alpha(n-i)\hat{\bar{\alpha}}(n)]+E[\hat{\bar{\alpha}}^2(n)]) \end{aligned} \quad \text{(EQ 4.32)}$$

By replacing $\alpha(n)$ by equation 4.29 and expanding, we get

$$\begin{aligned} E[\sigma_\alpha^2] &= \frac{1}{N-1}\sum_{i=0}^{N-1} A(i) - \frac{2}{N(N-1)}\sum_{i=0}^{N-1} A(i) + \frac{1}{N^2(N-1)}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} A(i) \\ &= \frac{1}{N}\sum_{i=0}^{N-1} A(i) \\ \text{where } A(i) &= \boldsymbol{HP}(n-i|n-i-1)\boldsymbol{H}^T+\sigma_v^2 \\ \therefore E[\sigma_\alpha^2] &= \frac{1}{N}\sum_{i=0}^{N-1} \boldsymbol{HP}(n-i|n-i-1)\boldsymbol{H}^T+\sigma_v^2 \end{aligned} \quad \text{(EQ 4.33)}$$

Since we have an unbiased estimator in $\hat{\sigma}_\alpha^2$, it is equal to the expected value. Therefore, our variance estimator can be written as

$$\sigma_v^2 = \hat{\sigma}_\alpha^2 - \frac{1}{N}\sum_{i=0}^{N-1} \boldsymbol{HP}(n-i|n-i-1)\boldsymbol{H}^T \quad \text{(EQ 4.34)}$$

At every new step, a new estimate is computed based on the new variance and the prediction error covariance matrix through a moving average operation. The value of N is set to 80, the double of the value that was suggested in [3] for the narrowband case, in order to cover the same time span. This value is high enough to give a decent variance representation and low enough to allow tracking in the case of noise change.

4.7 Other Implementation Details

This section will present a few extra details that are required to complete the implementation.

4.7.1 Normalization of Speech Input

The algorithm makes use of a perceptual filter in which the input speech power spectrum is compared to various thresholds. Therefore, it is of the utmost importance to make sure that the input speech and threshold of hearing be calibrated so they can be compared.

The input speech is normalized to be in dB SPL to correspond to the units of the threshold of hearing. It is first scaled by the reference level of 20 μPa and then multiplied by a constant computed such that the root mean square of the speech is 70 dB SPL (refer to table 2.1). In order to normalize, we had to make an assumption on the level of recording. For normal conversation, we have assumed that it was recorded around 70 dB SPL. A snippet of Matlab code is shown in figure 4.5 where the input speech is represented by x and the properly 70 dB SPL normalized speech by *xcal* [65].

```
x_scaled = x./20e-6;
RMS = sqrt(mean(x_scaled.^2));
calcSpl = 20*log10(RMS);
c = 10^((70 - calcSpl)/20);

xcal = c*(x_scaled);
```

FIGURE 4.5 - Normalization of speech input to dB SPL

4.7.2 Algorithm Initialization

Before the Kalman filter algorithm starts, it must be initialized properly. The iterations actually start at $n = p$ because we do not have information on the previous states. The noisy speech starts at $n = 0$. The filtered state estimation, the measurement noise variance, and the estimation error covariance matrix require special notes.

1. **Filtered State Estimation - $\hat{x}(n|n)$**

   The initial state estimation is initialized with the noisy data because it is the best that can be done with the information available.

$$\hat{x}(p-1|p-1) = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(p-1) \end{bmatrix},$$

   where *y(n)* is the normalized noisy speech.

2. **Measurement Noise Variance - σ_v^2**

   The measurement noise variance is roughly estimated from the initial $6N = 480$ noisy samples, an arbitrary value. Theoretically, for the variance estimate to be precise, one would need to assume that there is no speech yet in the first few frames and that by computing the variance of those samples, we get the variance of the noise. Even if this supposition is only partly true (i.e. there could be speech in the initial frames), the initial variance estimate is not too critical and will rapidly steady itself once the filter runs through a few iterations.

$$\sigma_v^2 = \frac{1}{L-1}\sum_{n=0}^{L-1}[y(n)-\bar{y}]^2 ,$$

   where *y(n)* is the normalized noisy speech, $\bar{y}$ is the mean of the first L samples of *y(n)*, and $L = 480$.

3. **Estimation Error Covariance Matrix - *P(n|n)***

   The estimation error covariance matrix is estimated by an identity matrix multiplied by the measurement noise variance. This makes sense because the error between the initial state which is the noisy speech and the actual state which is the clean speech should be the variance of the measurement noise.

$$P(p-1|p-1) = \sigma_v^2 I_{pxp} ,$$

The driving process variance σ_u^2 and the transition matrix ***F(n)*** are initialized with the variance of the residual (section 4.6.1) and LP coefficients (section 4.5) computed from the first frame respectively. The predicted state estimation $\hat{x}(p|p-1)$ and predicted error covariance matrix ***P(p|p-1)*** are computed according to the Kalman filter equations (equation 4.21).

Finally, using these initialized states and parameters, the Kalman filter algorithm can be executed starting at iteration $n = p$ as it was executed for this work.

4.7.3 MMSE STSA Based on Weighted Noise Estimation

The MMSE STSA based on weighted noise estimation algorithm used for the computation of the LP coefficients requires many different parameters. This section briefly summarizes the values used in this implementation. All of the values are the ones recommended in [49]. They are summarized in table 4.1.

Parameter	Value	Parameter	Value
α	0.98	θ_Z	7 dB
q	0.20	T_{init}	4 frames
L_Z	20	θ_G	10 dB
$\tilde{\gamma}_1$	0 dB	G_{mod}	-1.0 dB
$\tilde{\gamma}_2$	10 dB	G_{floor}	-6.8 dB

Table 4.1: Parameters for MMSE STSA based on weighted noise estimation

During the algorithm's windowing operation, the speech is separated into frames of 256 samples with 50% overlap. Each frame is multiplied by a von Hann window without the zero-weighted samples (refer to section 4.5). At the end of the algorithm, this operation is undone by dividing each enhanced frame by the von Hann window and "unwindowing" the data as it was done in section 4.4.1 resulting in the enhanced speech.

4.8 Summary of Kalman Filtering Based Method

A summary of the different steps required in the proposed Kalman filtering method is presented in figure 4.6 which serves as an overview. The reader is redirected to the proper sections for more details.

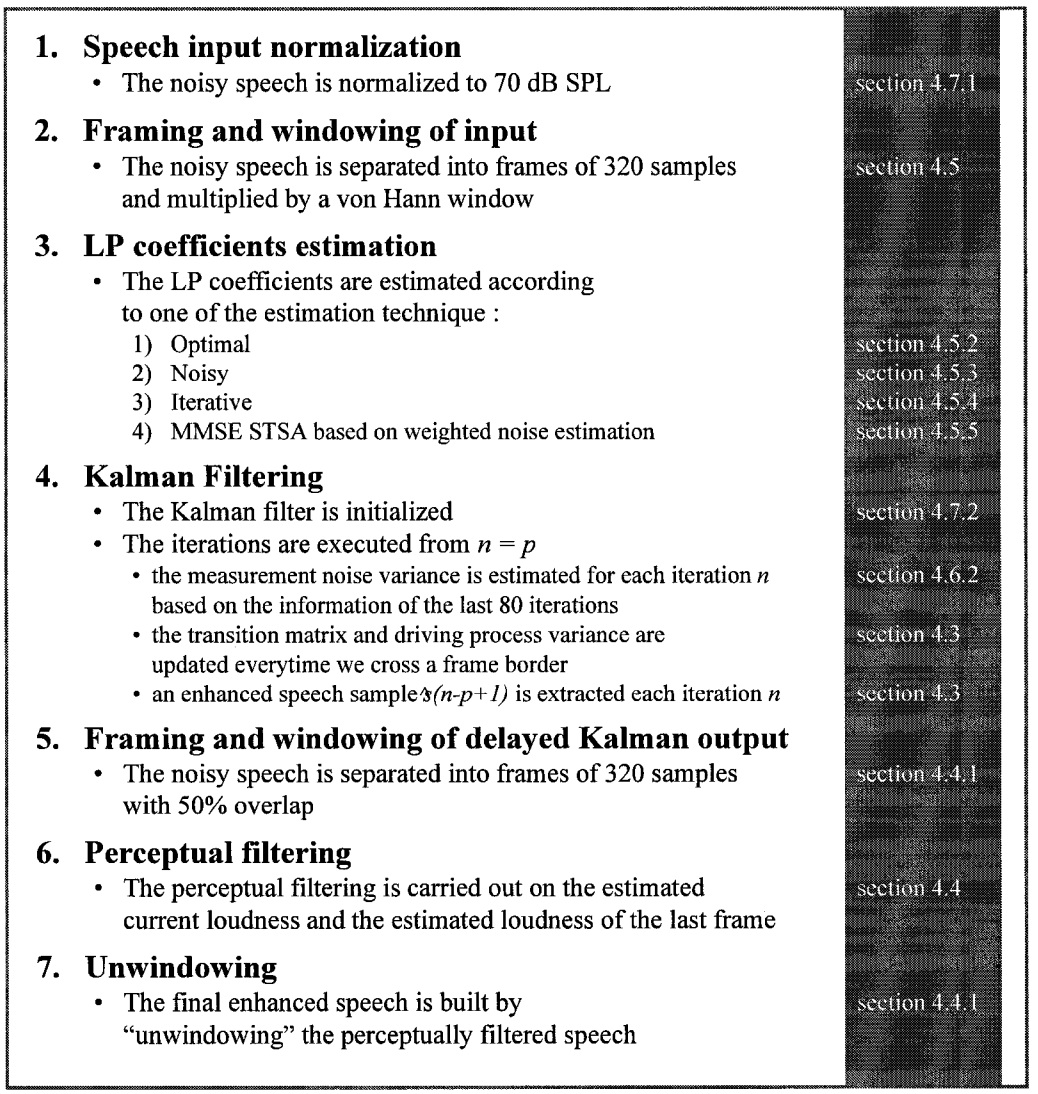

FIGURE 4.6 - Summary of proposed Kalman filter algorithm

4 KALMAN FILTERING BASED METHODS FOR WIDEBAND SPEECH ENHANCEMENT

CHAPTER 5

SIMULATION AND RESULTS

5.1 Introduction

This chapter presents the results of the experimental simulations as well as detailed explanations of the tests carried out. The tests are first explained, followed by the results. Analysis is given along the way where warranted and required.

In this chapter, only the overall results (i.e. the averaged results) are presented. The tables containing the detailed results for every sample file can be found in appendix A.

5.2 Setup

5.2.1 Environment

The various algorithms have been implemented in Matlab and were tested using Matlab 7.0. Matlab functions have been used to save time and increase efficiency where it was possible, such as for the computation of the LP coefficients using the function *lpc* and the Fast Fourier Transform computations using the function *fft*.

The implementation of the WPESQ algorithm (P.862.3) has been taken directly from the ITU-T standard and compiled with a C compiler. The segmental SNR is computed directly in Matlab by using the clean speech and the enhanced speech.

5.2.2 *Sample Files*

The simulations were conducted on 10 sample .wav files taken from the TIMIT speech database (table A.1 on page 85). The complete selection is made of 5 different male voices and 5 different female voices from different dialect regions. The sample files are around 3 seconds in length and start with 0.1 second of silence. The total percentage of silence versus speech spawns from 10% to 20% but averages around 14%. Each sample file has been sampled at 16 kHz.

5.2.3 *White Gaussian Noise*

The white Gaussian noise sequences *v(n)* are generated in Matlab using the function *randn*. The noise amplitude is modified so that the equation

$$\text{SNR} = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n v^2(n)} \qquad \text{(EQ 5.1)}$$

gives the desired signal-to-noise ratio. The corruption of the clean speech is done by simple addition of the noise sequence to the clean speech, resulting in noisy speech at the desired SNR.

5.2.4 *Quality Measures*

A few speech quality measures were presented in section 2.6. The algorithms are evaluated using the following quality measures :

- WPESQ measure
- Segmental SNR
- MUSHRA-like Subjective Evaluation

In a few cases, some speech assessment evaluation is done with the help of the spectrogram to show visually a specific result or effect. Although perceptual evaluation done through visual inspection is limited, useful information can nevertheless be extracted.

- Spectral Subtraction with perceptual masking ---------- SSV
- MMSE STSA based on weighted noise estimation -------- MMSE BEW
- Kalman filter - Optimal ---------- KFO
- Kalman filter - Optimal - Masked ---------- KFOM
- Kalman filter - Noisy ---------- KFN
- Kalman filter - Noisy - Masked ---------- KFNM
- Kalman filter - Iterative ---------- KFI
- Kalman filter - Iterative - Masked ---------- KFIM
- Kalman filter - MMSE STSA WNE ---------- KFM
- Kalman filter - MMSE STSA WNE - Masked ---------- KFMM

FIGURE 5.1 - Algorithms used during simulation

5.2.5 Description of Tests

The proposed Kalman filter algorithms are compared with other speech enhancement methods for validation. A list of the algorithms used for the simulation is given in figure 5.1. This figure also includes corresponding labels that are used in tables and legends to save a bit of space.

The different algorithms are used to enhance noisy speech at various signal-to-noise ratios. Seven different levels of noise are used : 20, 15, 10, 5, 0, -5, and -10 dB. Corrupting each of the 10 aforementioned clean speech samples at the different SNRs, 70 noisy samples are generated. Each of these is run through all of the different speech enhancement algorithms.

The WPESQ and segmental SNR quality measures are computed on every resulting enhanced speech and are averaged for each SNR.

MUSHRA-like Subjective Testing

The subjective test is based on the MUSHRA standard [39] as an informal evaluation with a limited number of participants (6 males and 2 females). These results are provided as reference only and should be taken as such.

There are three tests in which the participants are asked to rate the various speech files. All three tests consist of speech corrupted by white noise at 0 dB SNR for test 1, 5 dB SNR for test 2, and 15 dB SNR for test 3. Each test has 5 conditions consisting of the original noisy speech and the enhanced speech produced by the KFOM, KFNM, KFIM, and KFMM algorithms. During one test, the listener is asked to rate the quality of the speech based on the long term comfort and lack of annoyance on a scale from 1 to 5 (based on the MOS scale). A reference signal (the original clean speech) is provided for comparison and can be heard as many times as desired. The sample files are played in a random order and the listener does not know which speech he/she is listening to. Once the listener has rated the 5 samples in one test, he/she can move on to the next one.

5.3 Speech Enhancement Performance

The results are separated into two categories to increase readability and ease of comparison:

1. Selected speech enhancement approaches
2. Kalman filter algorithms with and without perceptual masking

5.3.1 Selected Speech Enhancement Approaches

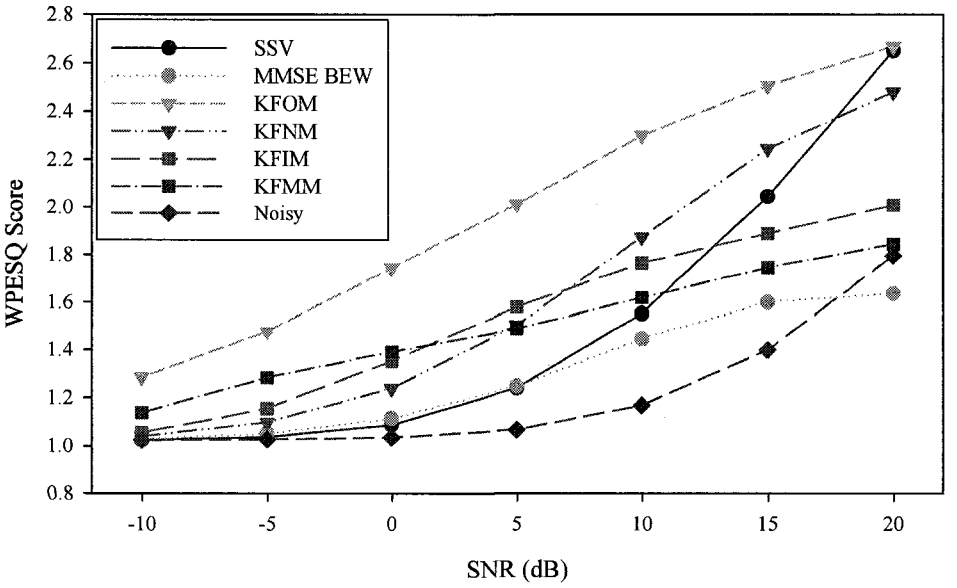

FIGURE 5.2 - Averaged WPESQ scores obtained in white Gaussian noise

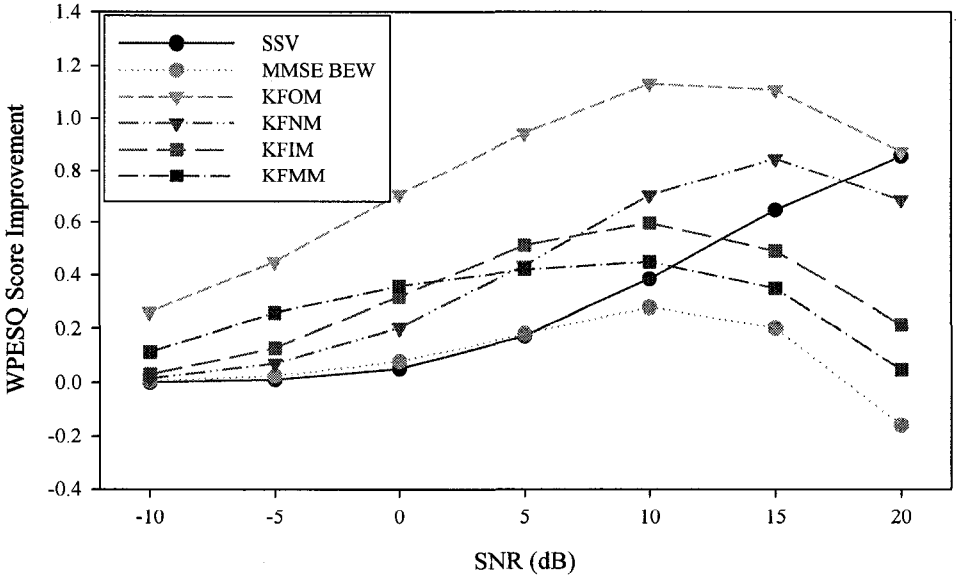

FIGURE 5.3 - Averaged WPESQ scores improvement in white Gaussian noise

Figure 5.2 shows the absolute WPESQ performance of the perceptually masked Kalman filter algorithms at various signal-to-noise ratios. The plot also includes the original noisy speech as well as the other two enhancement algorithms used for comparison. As it can be observed, all of the Kalman filter algorithms offer improvement in the perceptual speech quality compared to the noisy speech. Furthermore, the Kalman filter algorithms outperform the SSV and MMSE BEW algorithms for low to medium SNRs (-10 to 10 dB). For higher SNRs, the improvement is less noticeable for KFIM and KFMM but is still better for KFOM and KFNM.

Comparing the various Kalman filter algorithms, we see that the KFOM approach gives the highest WPESQ results. This is not too surprising since the LP coefficients are computed from the clean speech. The KFNM gives good results for high SNRs because the speech is relatively clean and gives decent LP coefficients estimations. For lower SNRs, this approach is the worst. The KFIM gives good WPESQ results for high SNRs as well and diminishes as it lowers. Finally, the KFMM gives higher results for low SNRs but not as good for high SNRs. This curve follows the MMSE BEW curve. And because at high SNRs, the MMSE BEW enhancement introduces some distortion, the LP coefficients estimation is not as good.

Figure 5.3 presents the information of figure 5.2 differently; it shows the WPESQ score improvement for ease of analysis.

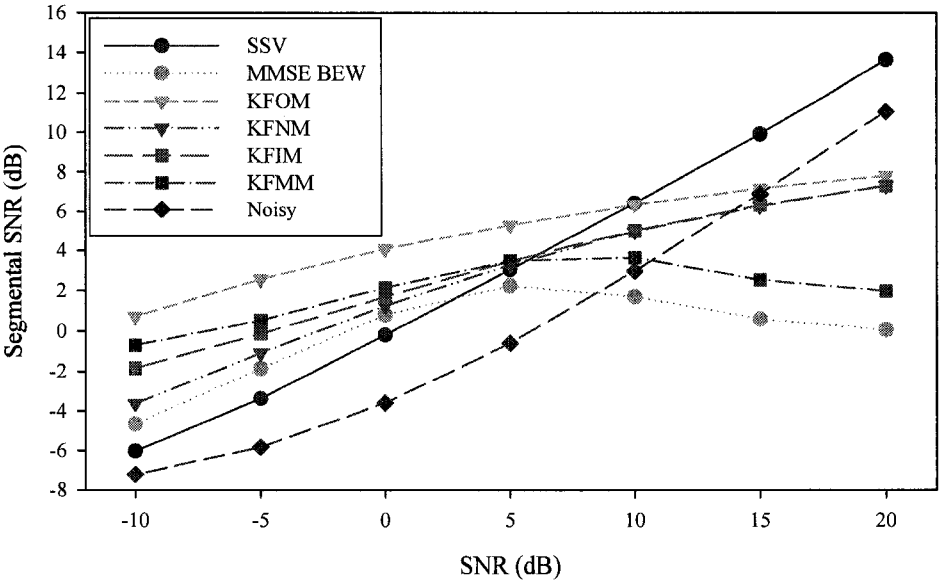

FIGURE 5.4 - Averaged segmental SNR obtained in white Gaussian noise

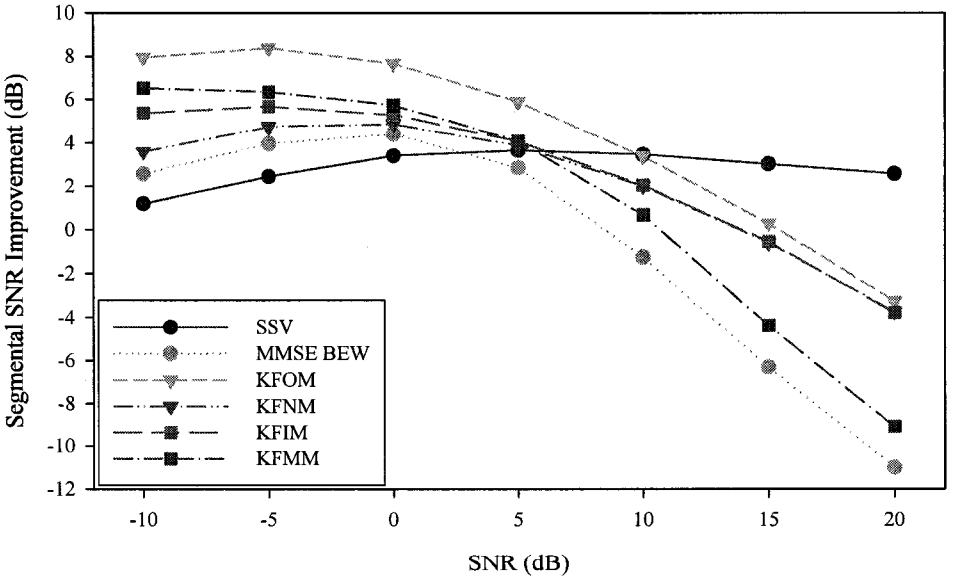

FIGURE 5.5 - Averaged segmental SNR improvement obtained in white Gaussian noise

Figure 5.4 shows the absolute segmental SNR performance of the same previous algorithms. For signal-to-noise ratios of 5 dB and below, the Kalman filter algorithms outperform SSV and MMSE BEW. SSV seems to perform well in high SNRs. Furthermore, for signal-to-noise ratios of 10 dB and lower, the Kalman filters algorithm improve the segmental SNRs. For higher SNRs, even though the segmental SNR is lower than the original noisy, there are still improvements in perceptual scores, as seen in figure 5.2 and figure 5.3. Again, the relative improvement of the algorithms is shown in figure 5.5.

5.3.2 Perceptual Masking

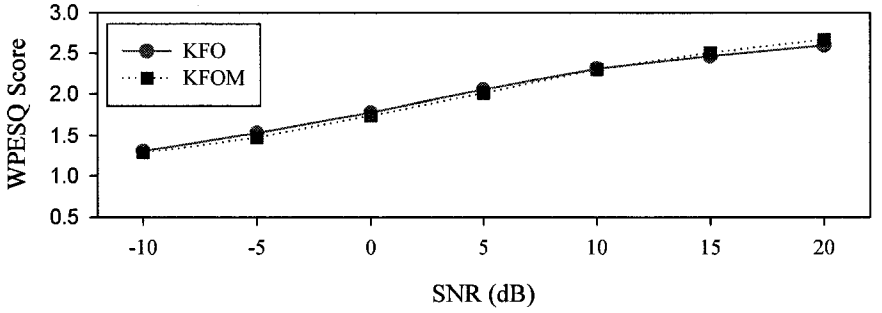

FIGURE 5.6 - Averaged WPESQ score of the optimal Kalman filter algorithm and the perceptually masked algorithm

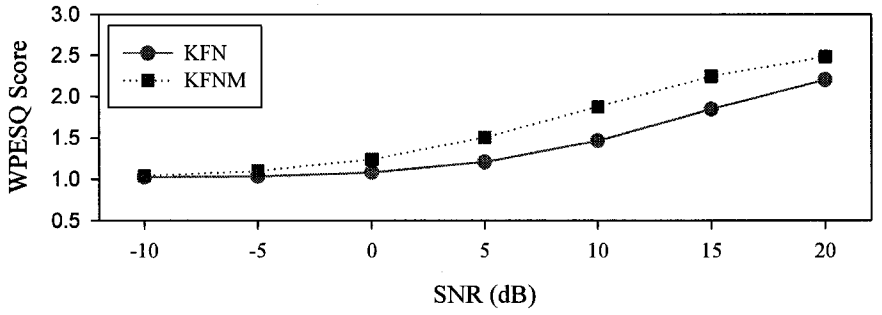

FIGURE 5.7 - Averaged WPESQ score of the noisy Kalman filter algorithm and the perceptually masked algorithm

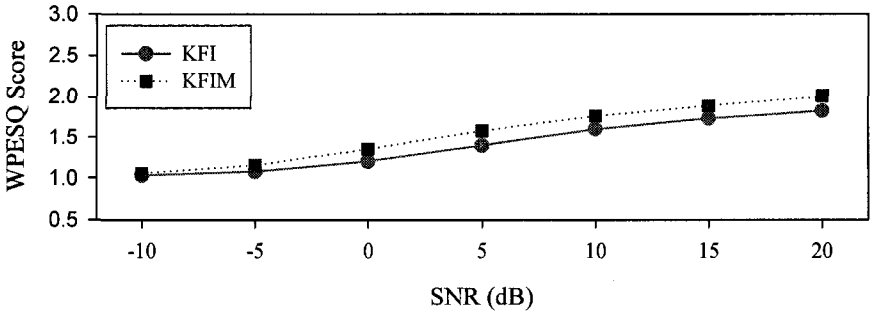

FIGURE 5.8 - Averaged WPESQ score of the iterative Kalman filter algorithm and the perceptually masked algorithm

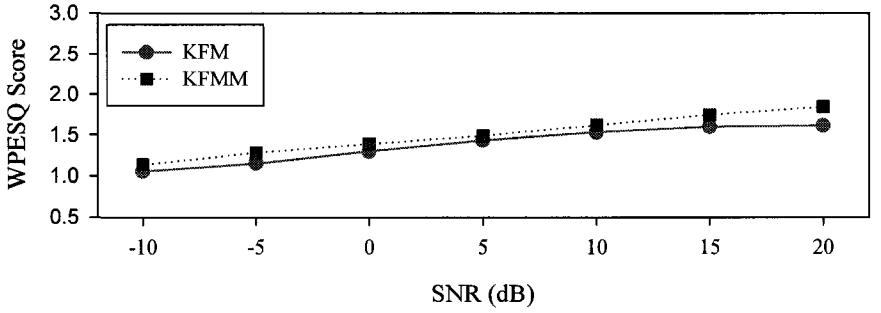

FIGURE 5.9 - Averaged WPESQ score of the MMSE BEW Kalman filter algorithm and the perceptually masked algorithm

Figures 5.6 to 5.9 show the WPESQ performance of the various Kalman filters with and without the perceptual masking applied. As it can be seen, the WPESQ score improvement is less noticeable when the LP coefficient computation uses the clean speech. In some cases, it even slightly worsens the score. Since the estimation is good to begin with, the perceptual masking seems to remove some good information. The improvement is most noticeable when the LP coefficients estimation is done from the

noisy speech. Since the estimation is corrupted, the perceptual masking removes some corrupted information and improves the overall enhanced speech.

From these observations, we can draw the conclusion that the more accurate the LP coefficients estimation is, the less the effect of the perceptual masking. The opposite is obviously true whereas the worse the estimation, the better the improvement of the perceptual masking.

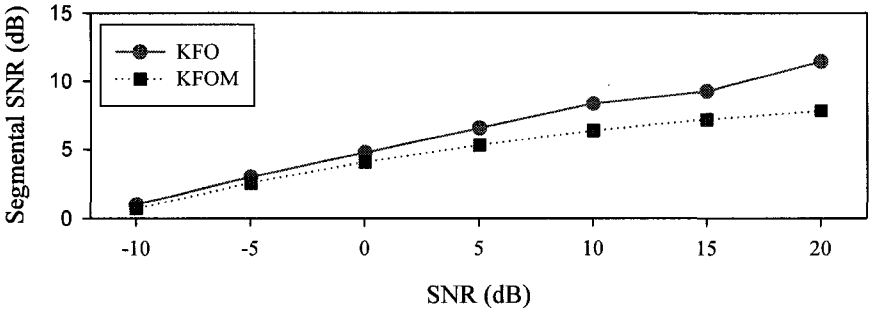

FIGURE 5.10 - Averaged segmental SNR score of the optimal Kalman filter algorithm and the perceptually masked algorithm

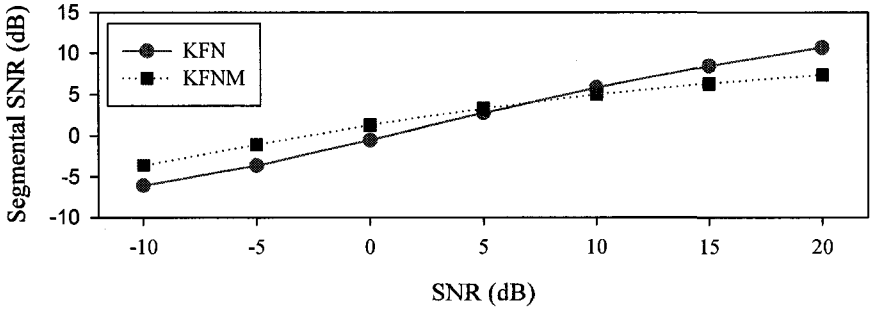

FIGURE 5.11 - Averaged segmental SNR score of the noisy Kalman filter algorithm and the perceptually masked algorithm

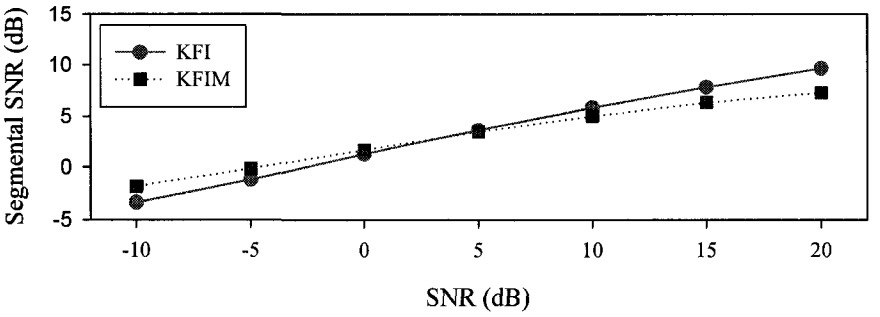

FIGURE 5.12 - Averaged segmental SNR score of the iterative Kalman filter algorithm and the perceptually masked algorithm

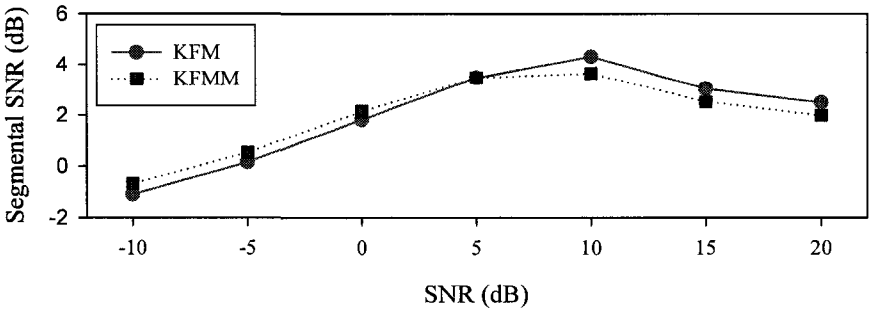

FIGURE 5.13 - Averaged segmental SNR score of the MMSE BEW Kalman filter algorithm and the perceptually masked algorithm

Figures 5.10 to 5.13 show the segmental SNR performance of the various Kalman filters with and without the perceptual masking applied. In the optimal case, the algorithm without the perceptual masking yields better results. The segmental SNR measure compares signal energies and does not take the sound perception into account which explains the differences. The same results are observed for the other algorithms for input SNRs of 10 dB and above. For lower SNRs, it seems that the speech is corrupted enough to benefit from the perceptual masking.

The use of the perceptual filter does not tend to increase the segmental SNR scores as it did with the WPESQ scores.

5.3.3 Subjective Test Results on a Subset of the Experiments

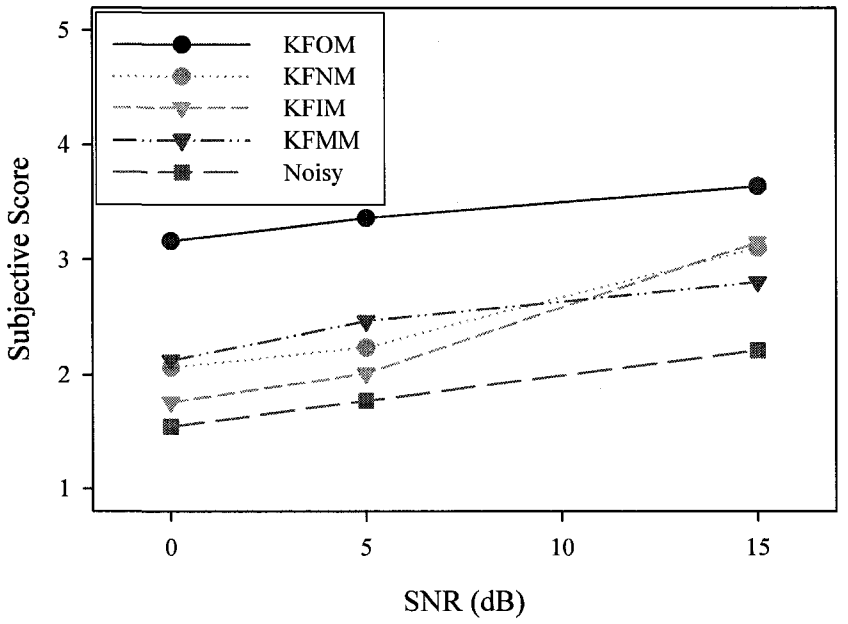

FIGURE 5.14 - Subjective scores in white noise

Algorithms	0 dB SNR	5 dB SNR	15 dB SNR
Noisy	1.55	1.78	2.21
KFIM	1.76	2.01	3.15
KFMM	2.13	2.46	2.80
KFNM	2.06	2.24	3.10
KFOM	3.16	3.36	3.64

Table 5.1: Mean of subjective scores

Algorithms.	0 dB SNR	5 dB SNR	15 dB SNR
Noisy	0.45	0.67	0.93
KFIM	0.51	0.72	1.35
KFMM	0.62	0.76	1.21
KFNM	0.42	0.56	0.99
KFOM	0.86	0.91	1.27

Table 5.2: Standard deviation of subjective results

Figure 5.14 shows the subjective results based on the MUSHRA procedure. Although the scores are slightly higher than the WPESQ ones, we see that the relationships between the curves remain somewhat similar. Again, KFOM has the best performance and the original noisy speech has the worst. KFMM performs well in lower SNRs but its performance compared to the other algorithms decrease as the SNR increases. KFNM does poorly in low SNRs but gets better as the input noise decreases. KFIM seems to do fairly good on the entire range.

The results are also included in table 5.1. Table 5.2 shows the standard deviation of the results. They are relatively high, especially for KFOM. This goes to show that the evaluation of speech quality is a truly subjective matter.

5.3.4 Note on Wideband Speech

Although the tests carried out give a good basis for evaluating and ranking the proposed methods, it is not necessarily the best to evaluate the remaining wideband characteristics. What is meant by that is that although a speech enhancement approach provides decent noise reduction, increasing speech quality, some high frequency components are filtered resulting in a loss or reduction of some wideband characteristics. The lower the input signal-to-noise ratio, the more information is lost. Figures 5.15 to 5.17 show the

spectrogram of one of the clean speech sample, its corresponding corrupted spectrogram at 5 dB SNR, and the spectrogram of the enhanced speech using KFMM respectively.

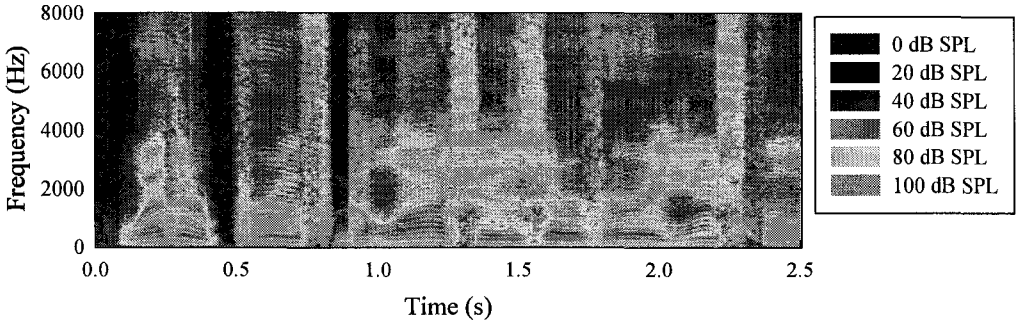

FIGURE 5.15 - Spectrogram of clean speech

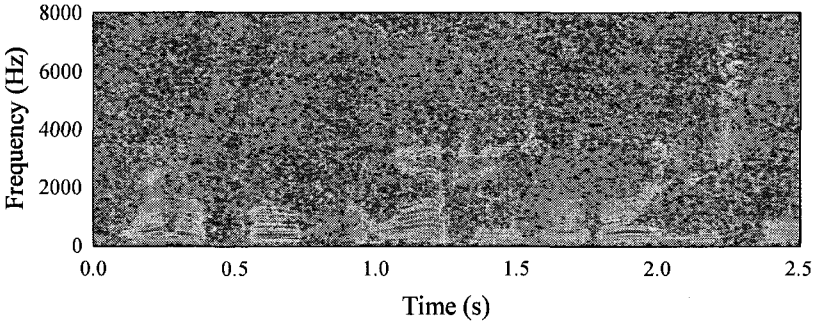

FIGURE 5.16 - Spectrogram of noisy speech at 5 dB SNR

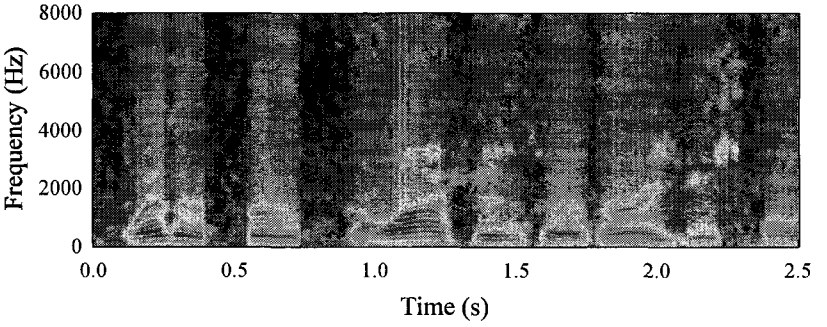

FIGURE 5.17 - Resulting spectrogram of KFMM enhanced speech

The energy of the high frequencies is usually much weaker compared to the lower frequencies but it still adds much intelligibility to the overall speech. As it can be observed in figure 5.16, the high frequencies are completely drowned in the noise and are difficult to extract and to model by the LP coefficients. The spectrogram of the enhanced speech, in this case by KFMM, shows that some of the wideband content of the original clean speech has been lost. This seems to be unavoidable in high levels of noise with the schemes discussed in this thesis.

5.4 Other Types of Noise

Although the Kalman filter algorithm was specifically developed to be used in white Gaussian noise environments, it is of interest to test its performance with other types of noise. Again, the ten sample files are corrupted at different SNRs and run through the different algorithms.

5.4.1 Colored Noise

The colored noise is generated by filtering a white Gaussian noise sequence by the multi-bandpass filter [3]

$$z(n) = -0.0851z(n-1) + 0.19126z(n-2) + 0.0458z(n-3) + 0.0229z(n-4) + 0.1097z(n-5) + 0.1553z(n-6) - 0.132z(n-7) - 0.76z(n-8) + w(n) \quad \text{(EQ 5.2)}$$

where *w(n)* is a zero-mean white noise sequence with variance of one. The frequency response of equation 5.2 is shown in figure 5.18.

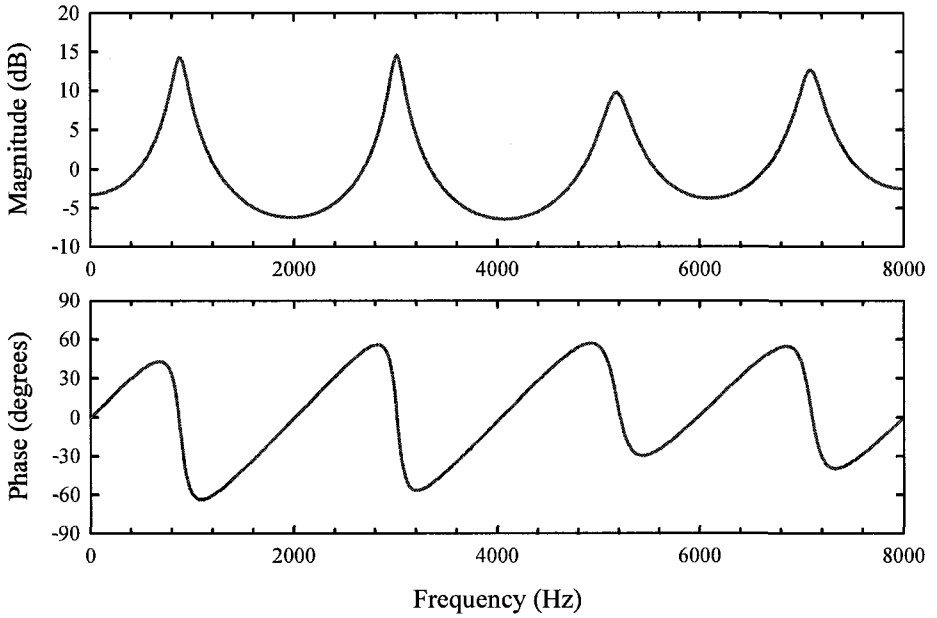

FIGURE 5.18 - Frequency response of filter

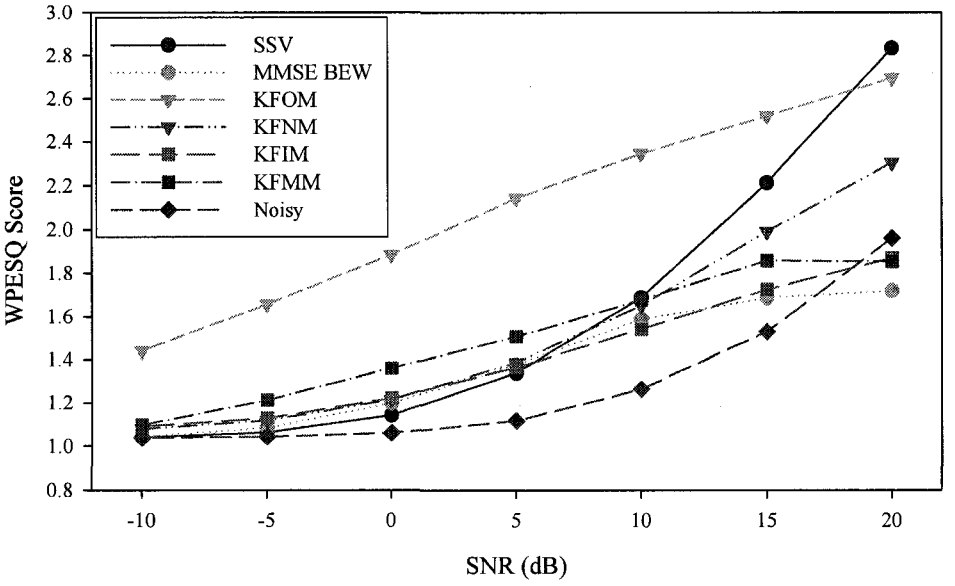

FIGURE 5.19 - Averaged WPESQ scores in colored noise

Figure 5.19 shows the performance of the various algorithms in colored noise. The performance of KFOM seems to remain unchanged. The KFNM, KFIM, and KFMM algorithms do not perform as well and KFMM is the most affected probably because of worse LP coefficients estimation. For SNRs lower than 5 dB, the Kalman filter algorithms outperform SSV and MMSE BEW.

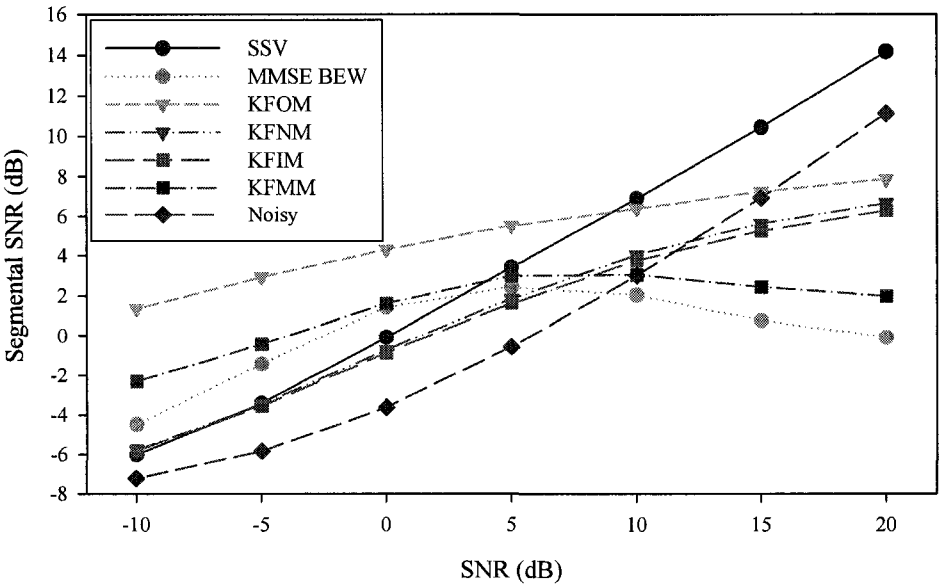

FIGURE 5.20 - Averaged segmental SNR in colored noise

Figure 5.20 shows the segmental SNR performance of the algorithms in colored noise. For signal-to-noise ratios at 10 dB and below, the Kalman filter algorithms improve the segmental SNR of the noisy speech. For higher SNRs, there is no more improvements. SSV does well in high input SNRs.

5.4.2 Street Noise

The street noise has been taken from the ITU-T P.23 Speech Database and its average power spectrum density is shown in figure 5.21.

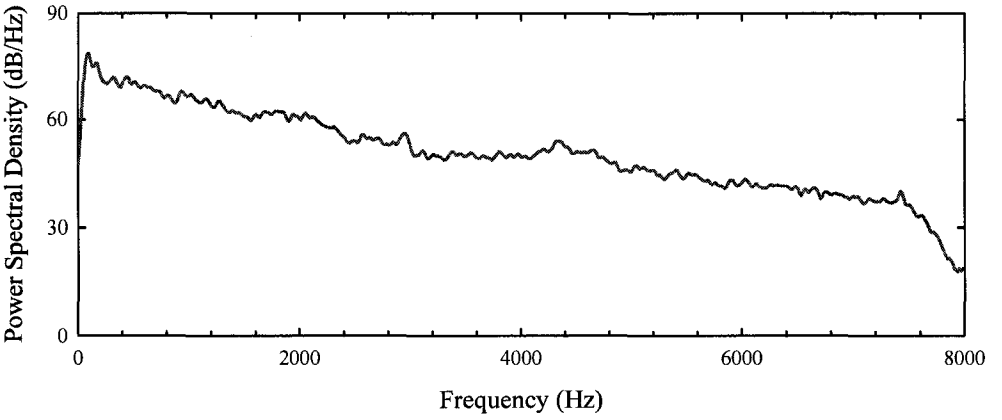

FIGURE 5.21 - Average power spectral density of street noise used in simulation

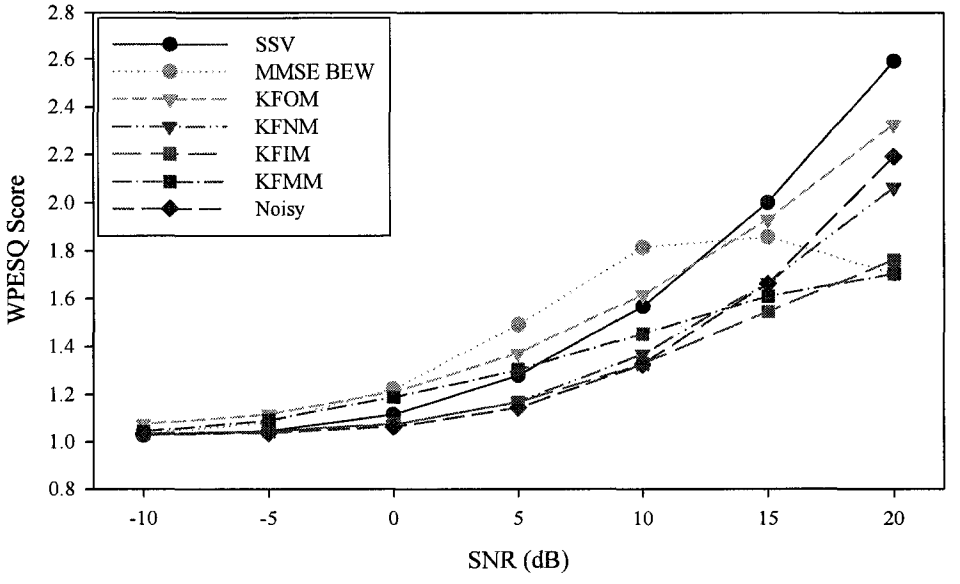

FIGURE 5.22 - Averaged WPESQ scores in street noise

Figure 5.22 shows the performance of the various algorithms in street noise. As we can see, for low SNRs, the improvement of the various algorithms is very limited. As the signal-to-noise ratio increases, the improvement also increases. Again, KFOM improves the WPESQ score at every SNR. From mid to low signal-to-noise ratios, KFMM also improves the score. KFIM and KFNM only provide slight improvements, none at higher SNRs. MMSE BEW and SSV give the best improvement over the 5 to 10 dB and 15 to 20 dB range respectively.

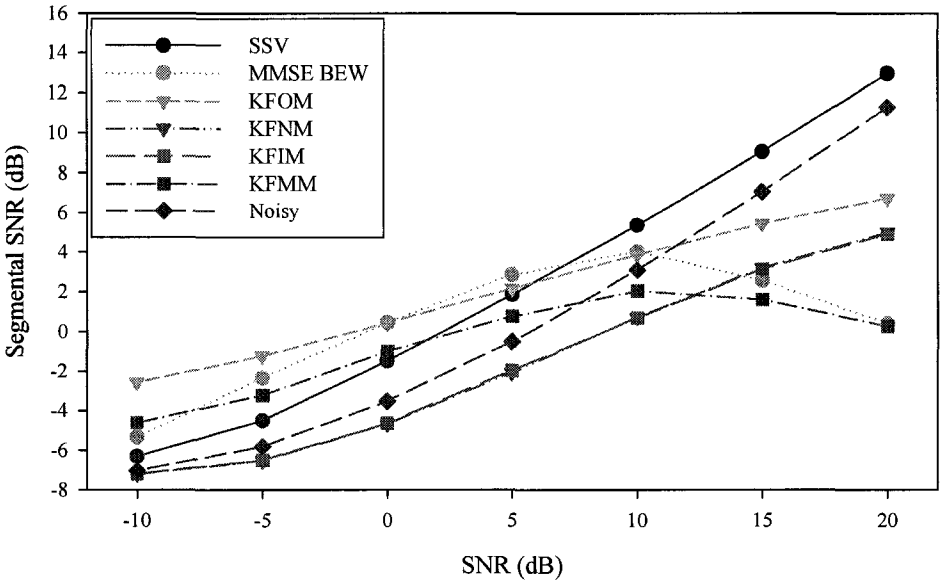

FIGURE 5.23 - Averaged segmental SNR in street noise

Figure 5.23 shows the segmental SNR performance of the various algorithms in street noise. KFOM provides improvement up until 10 dB. KFNM improves the segmental SNR up until 5 dB. KFMM and KFIM follow each other very closely and do not improve the segmental SNR at all over the entire range. SSV performs the best in high SNRs and MMSE BEW performs the best at 0 dB and 5 dB but does not improve the segmental SNR at higher SNRs.

CHAPTER 6

CONCLUSION

6.1 Summary of Work

In this present work, speech enhancement algorithms have been applied to wideband noisy speech. Specifically, perceptual Kalman filters have been shown to improve significantly speech quality in wideband environments.

As it was shown, some of the wideband information is lost in the corruption process; an effect that seems unavoidable with the schemes used in this thesis. However, improving the overall speech quality is still possible.

The non-practical KFO and KFOM algorithms provide the best speech quality. In this case, where the LP coefficients are computed from the clean speech, the perceptual masking does not improve the results further. For low signal-to-noise ratios, the KFMM provides better enhancement. As the SNR increases, the KFIM and KFNM approaches seem to yield better results. In these cases, the perceptual masking increases the perceptual quality further.

Although the Kalman filter algorithms have been developed specifically for white Gaussian noise, they can nevertheless, in some conditions, improve the quality of speech corrupted by colored noise and street noise.

6.2 Future Work

As it was observed, the Kalman filter algorithm performance is directly linked to quality of the estimation of the LP coefficients. By computing the LP coefficients directly from the clean speech, the optimal results are obtained. However, this estimation is not feasible in real life since we do not have the clean speech to work with. By improving the estimation of the LP coefficients, the resulting speech quality can also be improved. Although some research has been done in that field, LP coefficients estimation approaches tailored to Kalman filtering could be investigated.

APPENDIX

A

COMPLETE SIMULATION RESULTS

A.1 Simulation Files

Simulation Files	TIMIT Files
male1.wav	TEST/DR1/MSTK0/SA2.wav
male2.wav	TEST/DR2/MABW0/SA1.wav
male3.wav	TEST/DR3/MCSH0/SX289.wav
male4.wav	TEST/DR6/MESD0/SX282.wav
male5.wav	TEST/DR7/MDVC0/SX36.wav
female1.wav	TEST/DR1/FAKS0/SX43.wav
female2.wav	TEST/DR2/FJRE0/SX396.wav
female3.wav	TEST/DR3/FPKT0/SX278.wav
female4.wav	TEST/DR6/FDRW0/SX23.wav
female5.wav	TEST/DR7/FCAU0/SX407.wav

Table A.1: Files used during simulation

A.2 WPESQ Scores in White Noise

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	2.01	1.56	1.57	1.86	1.77	1.64	1.96	1.81	2.12	1.67
SS	2.37	1.86	1.90	2.24	2.06	1.92	2.36	2.15	2.66	2.02
SSV	2.95	2.36	2.40	2.65	2.47	2.37	2.89	2.71	3.15	2.58
MMSE	1.95	1.74	1.64	1.81	1.64	1.40	1.69	1.42	1.58	1.50
KFO	2.83	2.33	2.63	2.59	2.68	2.53	2.51	2.64	2.67	2.56
KFOM	3.01	2.55	2.69	2.63	2.76	2.61	2.54	2.61	2.73	2.55
KFN	2.31	2.01	1.95	2.16	2.06	2.14	2.27	2.29	2.58	2.26
KFNM	2.78	2.41	2.26	2.44	2.35	2.45	2.52	2.49	2.70	2.42
KFI	2.20	1.83	1.77	1.79	1.72	2.05	1.64	1.67	1.83	1.76
KFIM	2.47	2.11	1.97	1.95	1.89	2.14	1.82	1.85	1.98	1.93
KFM	2.19	2.02	1.66	1.63	1.59	1.38	1.63	1.33	1.36	1.37
KFMM	2.54	2.38	1.97	1.74	1.98	1.55	1.58	1.51	1.64	1.56

Table A.2: WPESQ scores in white noise at 20 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.57	1.27	1.27	1.43	1.43	1.32	1.46	1.39	1.53	1.31
SS	1.86	1.47	1.46	1.69	1.64	1.51	1.76	1.64	1.92	1.51
SSV	2.25	1.81	1.89	2.04	1.95	1.87	2.22	1.99	2.50	1.94
MMSE	1.83	1.52	1.51	1.72	1.58	1.50	1.72	1.42	1.73	1.47
KFO	2.78	2.47	2.69	2.52	2.72	2.28	2.28	2.34	2.36	2.23
KFOM	2.90	2.62	2.69	2.57	2.79	2.31	2.29	2.31	2.40	2.21
KFN	1.93	1.66	1.66	1.79	1.73	1.71	1.96	1.91	2.28	1.81
KFNM	2.44	2.20	2.06	2.24	2.15	2.17	2.36	2.25	2.46	2.12
KFI	2.06	1.76	1.70	1.76	1.63	1.84	1.61	1.62	1.70	1.64
KFIM	2.29	2.03	1.90	1.89	1.79	1.95	1.75	1.75	1.79	1.74
KFM	2.14	1.95	1.71	1.70	1.62	1.37	1.45	1.33	1.42	1.32
KFMM	2.37	2.25	1.92	1.78	1.83	1.46	1.51	1.34	1.56	1.44

Table A.3: WPESQ scores in white noise at 15 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.27	1.11	1.10	1.19	1.22	1.13	1.18	1.16	1.21	1.13
SS	1.46	1.22	1.20	1.34	1.34	1.25	1.35	1.28	1.42	1.23
SSV	1.72	1.39	1.45	1.53	1.56	1.46	1.66	1.49	1.81	1.45
MMSE	1.56	1.29	1.33	1.53	1.42	1.36	1.64	1.33	1.66	1.34
KFO	2.77	2.59	2.50	2.48	2.72	2.05	1.99	2.02	2.01	1.98
KFOM	2.84	2.63	2.42	2.43	2.74	2.02	1.98	1.96	2.01	1.95
KFN	1.59	1.36	1.39	1.42	1.45	1.35	1.53	1.49	1.68	1.38
KFNM	2.05	1.84	1.77	1.92	1.89	1.77	1.95	1.83	1.98	1.72
KFI	1.84	1.61	1.56	1.65	1.51	1.66	1.55	1.54	1.61	1.51
KFIM	2.07	1.87	1.78	1.85	1.68	1.80	1.65	1.65	1.68	1.61
KFM	1.82	1.73	1.61	1.66	1.57	1.46	1.38	1.32	1.42	1.37
KFMM	1.98	1.94	1.73	1.73	1.71	1.45	1.42	1.31	1.49	1.42

Table A.4: WPESQ scores in white noise at 10 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.14	1.04	1.04	1.08	1.09	1.05	1.06	1.06	1.07	1.05
SS	1.21	1.08	1.07	1.14	1.15	1.09	1.12	1.10	1.16	1.09
SSV	1.33	1.17	1.17	1.24	1.30	1.23	1.28	1.20	1.32	1.16
MMSE	1.33	1.17	1.20	1.25	1.31	1.19	1.38	1.18	1.32	1.16
KFO	2.48	2.41	2.19	2.39	2.35	1.78	1.75	1.72	1.71	1.76
KFOM	2.43	2.40	2.12	2.30	2.36	1.71	1.73	1.65	1.71	1.72
KFN	1.31	1.16	1.16	1.21	1.26	1.15	1.23	1.19	1.26	1.16
KFNM	1.67	1.48	1.43	1.52	1.56	1.41	1.57	1.46	1.57	1.37
KFI	1.52	1.40	1.38	1.39	1.37	1.36	1.42	1.36	1.46	1.34
KFIM	1.78	1.66	1.59	1.64	1.56	1.54	1.54	1.48	1.54	1.48
KFM	1.58	1.51	1.44	1.52	1.46	1.41	1.38	1.33	1.38	1.36
KFMM	1.69	1.62	1.53	1.61	1.54	1.45	1.40	1.32	1.38	1.35

Table A.5: WPESQ scores in white noise at 5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.08	1.03	1.02	1.04	1.04	1.03	1.03	1.03	1.03	1.03
SS	1.11	1.03	1.03	1.06	1.06	1.04	1.05	1.04	1.05	1.03
SSV	1.15	1.06	1.06	1.10	1.12	1.08	1.09	1.07	1.10	1.06
MMSE	1.15	1.09	1.07	1.12	1.17	1.09	1.14	1.10	1.15	1.09
KFO	2.08	2.03	1.86	2.17	2.07	1.52	1.56	1.49	1.50	1.53
KFOM	2.05	2.05	1.81	2.08	2.06	1.52	1.52	1.41	1.48	1.47
KFN	1.16	1.06	1.05	1.09	1.12	1.06	1.08	1.07	1.09	1.06
KFNM	1.35	1.22	1.20	1.25	1.28	1.18	1.28	1.21	1.26	1.15
KFI	1.26	1.19	1.18	1.20	1.24	1.17	1.22	1.19	1.20	1.17
KFIM	1.46	1.40	1.33	1.38	1.37	1.30	1.37	1.30	1.34	1.28
KFM	1.35	1.33	1.27	1.28	1.33	1.29	1.33	1.27	1.29	1.27
KFMM	1.50	1.51	1.37	1.42	1.42	1.40	1.38	1.30	1.31	1.31

Table A.6: WPESQ scores in white noise at 0 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.05	1.02	1.02	1.03	1.03	1.03	1.03	1.02	1.02	1.02
SS	1.06	1.02	1.02	1.03	1.03	1.03	1.03	1.02	1.03	1.02
SSV	1.07	1.03	1.03	1.04	1.05	1.04	1.03	1.03	1.04	1.03
MMSE	1.10	1.04	1.03	1.06	1.06	1.05	1.04	1.05	1.05	1.04
KFO	1.76	1.74	1.57	1.88	1.81	1.33	1.38	1.25	1.31	1.30
KFOM	1.72	1.69	1.54	1.75	1.76	1.25	1.31	1.21	1.29	1.27
KFN	1.08	1.03	1.03	1.04	1.05	1.03	1.03	1.03	1.04	1.03
KFNM	1.17	1.10	1.07	1.13	1.14	1.07	1.09	1.07	1.09	1.05
KFI	1.12	1.06	1.06	1.09	1.11	1.07	1.07	1.07	1.08	1.07
KFIM	1.20	1.18	1.14	1.15	1.20	1.13	1.16	1.14	1.15	1.11
KFM	1.16	1.18	1.11	1.14	1.24	1.14	1.16	1.13	1.17	1.13
KFMM	1.30	1.40	1.24	1.25	1.35	1.26	1.28	1.22	1.29	1.24

Table A.7: WPESQ scores in white noise at -5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.04	1.02	1.02	1.02	1.03	1.03	1.03	1.02	1.02	1.02
SS	1.04	1.02	1.02	1.02	1.03	1.03	1.03	1.02	1.02	1.02
SSV	1.04	1.02	1.02	1.02	1.03	1.03	1.03	1.02	1.02	1.02
MMSE	1.06	1.02	1.02	1.03	1.03	1.03	1.03	1.02	1.03	1.03
KFO	1.51	1.45	1.31	1.57	1.50	1.16	1.20	1.12	1.15	1.14
KFOM	1.44	1.43	1.28	1.53	1.51	1.15	1.16	1.11	1.13	1.12
KFN	1.05	1.02	1.02	1.03	1.03	1.03	1.03	1.02	1.02	1.02
KFNM	1.07	1.04	1.03	1.06	1.05	1.03	1.04	1.03	1.03	1.03
KFI	1.05	1.03	1.03	1.03	1.05	1.04	1.03	1.03	1.03	1.03
KFIM	1.08	1.05	1.05	1.07	1.08	1.05	1.05	1.05	1.05	1.04
KFM	1.08	1.05	1.04	1.06	1.09	1.05	1.04	1.05	1.06	1.05
KFMM	1.16	1.17	1.12	1.14	1.18	1.13	1.11	1.13	1.15	1.10

Table A.8: WPESQ scores in white noise at -10 dB

A.3 Segmental SNRs in White Noise

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	11.27	12.69	12.23	10.08	9.98	9.23	10.58	11.54	12.42	10.63
SS	12.66	14.43	13.95	11.32	11.38	10.77	11.77	12.80	13.63	12.33
SSV	14.15	15.60	14.88	12.15	12.75	12.17	12.92	13.63	14.27	13.93
MMSE	0.35	2.00	2.37	-1.62	-1.67	-1.57	0.65	0.90	0.66	-1.44
KFO	11.32	12.17	13.39	8.68	11.48	11.39	10.60	11.46	12.40	11.82
KFOM	8.17	9.25	8.76	5.42	8.88	8.35	4.83	7.15	8.10	9.18
KFN	10.97	12.87	13.19	7.82	10.49	9.24	9.33	10.66	11.32	10.63
KFNM	7.40	9.42	8.48	5.10	8.39	7.20	4.14	6.72	7.41	8.98
KFI	10.27	11.81	12.13	7.22	10.26	7.92	8.51	8.91	9.81	9.85
KFIM	7.58	9.25	8.47	5.23	8.32	7.30	4.36	6.54	7.26	8.58
KFM	6.09	1.71	3.88	2.00	2.41	2.21	-0.55	2.52	3.68	0.91
KFMM	4.14	1.68	1.34	2.01	2.03	1.03	0.53	2.26	3.30	1.43

Table A.9: Segmental SNRs in white noise at 20 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	7.24	8.24	7.94	5.84	6.00	5.39	6.50	7.30	8.24	6.25
SS	8.90	10.26	9.85	7.41	7.56	7.09	7.94	8.92	9.85	8.20
SSV	10.35	11.55	11.26	8.61	9.03	8.69	9.04	10.01	10.63	10.06
MMSE	2.16	2.59	1.97	-0.51	-1.68	-1.33	1.26	1.50	0.91	-0.95
KFO	9.91	6.94	8.20	7.49	9.97	9.95	9.17	10.05	10.80	10.11
KFOM	7.30	8.04	8.50	4.72	8.07	7.73	4.50	6.83	7.64	8.37
KFN	8.75	10.37	10.26	6.11	8.03	7.29	7.02	8.73	9.35	8.36
KFNM	5.94	8.22	7.57	4.35	7.18	6.20	3.09	6.01	6.57	7.63
KFI	8.43	9.35	9.61	5.92	8.19	6.57	6.65	7.54	8.00	8.02
KFIM	6.26	8.02	7.45	4.23	7.38	6.43	3.62	6.00	6.42	7.49
KFM	5.39	3.59	2.99	3.46	4.36	-0.68	1.71	4.20	4.63	0.72
KFMM	4.40	3.23	2.66	3.16	3.45	-0.69	1.60	2.90	4.22	0.26

Table A.10: Segmental SNRs in white noise at 15 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	3.41	4.03	3.83	1.82	2.30	1.76	2.68	3.39	4.30	2.18
SS	5.24	6.14	5.90	3.62	4.05	3.65	4.19	5.14	6.17	4.21
SSV	6.77	7.88	7.59	5.15	5.74	5.50	5.33	6.72	7.32	6.33
MMSE	3.64	2.64	3.38	0.77	-0.79	-2.76	2.58	2.34	4.27	0.89
KFO	8.65	9.25	9.34	6.40	8.38	8.23	7.69	8.33	8.89	8.19
KFOM	5.54	7.72	7.45	4.27	7.24	6.85	4.01	6.31	6.91	7.13
KFN	6.09	7.42	7.16	3.80	5.31	4.71	4.73	6.24	6.94	5.46
KFNM	4.36	6.74	6.23	3.32	5.66	4.97	1.89	5.25	5.52	5.83
KFI	6.15	7.37	7.24	4.38	5.98	4.89	4.70	5.93	6.15	5.84
KFIM	4.11	6.71	6.27	3.12	6.10	5.06	2.37	5.10	5.39	5.89
KFM	6.19	6.05	5.75	4.05	5.47	-1.23	3.03	4.74	6.13	3.01
KFMM	4.97	4.93	5.03	3.96	5.62	-0.40	0.81	3.92	4.58	3.04

Table A.11: Segmental SNRs in white noise at 10 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-0.16	0.11	0.02	-1.83	-0.96	-1.60	-0.72	-0.11	0.61	-1.40
SS	1.68	2.28	2.12	0.00	0.77	0.31	0.54	1.63	2.70	0.58
SSV	3.40	4.32	3.89	1.78	2.67	2.32	1.68	3.45	4.17	2.71
MMSE	3.58	2.53	3.47	1.44	1.12	-0.77	2.82	3.11	4.28	0.78
KFO	6.73	7.25	7.26	5.25	6.52	6.40	6.14	6.57	6.94	6.28
KFOM	4.84	6.39	6.25	3.79	5.96	5.66	3.01	5.45	5.86	5.65
KFN	2.90	4.11	3.91	0.90	2.48	1.69	1.59	3.27	4.03	2.22
KFNM	2.46	4.92	4.44	1.92	3.70	3.09	0.38	3.95	4.01	3.80
KFI	3.55	5.03	4.69	2.62	3.67	2.87	2.15	4.01	4.20	3.47
KFIM	2.84	4.97	4.52	2.09	4.34	3.26	0.94	3.80	3.91	3.97
KFM	3.87	4.37	4.66	2.73	4.43	0.03	3.69	4.01	4.31	2.66
KFMM	3.67	4.21	4.61	2.93	4.76	0.62	2.71	4.06	4.32	2.87

Table A.12: Segmental SNRs in white noise at 5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-3.22	-3.45	-3.22	-4.73	-3.66	-4.34	-3.74	-3.04	-2.48	-4.12
SS	-1.70	-1.39	-1.34	-3.31	-2.12	-2.72	-2.87	-1.49	-0.60	-2.51
SSV	0.06	0.99	0.53	-1.55	-0.25	-0.64	-1.61	0.36	1.00	-0.82
MMSE	0.70	1.90	1.78	-0.11	0.63	-1.68	0.13	1.72	2.09	0.90
KFO	5.04	5.29	5.33	3.91	4.86	4.65	4.02	4.86	5.15	4.49
KFOM	4.18	4.85	4.80	2.99	4.55	4.19	2.07	4.33	4.54	4.13
KFN	-0.64	0.58	0.56	-2.32	-0.45	-1.32	-1.68	0.10	0.81	-1.12
KFNM	0.62	2.73	2.39	0.05	1.48	0.88	-1.39	2.03	2.12	1.55
KFI	0.86	2.59	2.31	0.46	1.39	0.73	-0.86	1.91	2.20	1.08
KFIM	0.96	3.00	2.66	0.90	2.35	1.37	-1.09	2.38	2.21	2.10
KFM	1.24	2.73	2.71	1.23	2.26	0.20	0.64	2.58	2.32	2.19
KFMM	1.67	3.04	2.95	1.63	3.05	0.89	0.25	2.87	2.66	2.35

Table A.13: Segmental SNRs in white noise at 0 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-5.77	-5.74	-5.64	-6.80	-5.74	-6.48	-6.09	-5.27	-4.74	-5.96
SS	-4.65	-4.57	-4.19	-5.93	-4.52	-5.25	-5.60	-4.12	-3.41	-5.01
SSV	-3.98	-2.56	-2.51	-4.64	-3.00	-3.65	-4.77	-2.67	-2.12	-3.88
MMSE	-2.67	-1.25	-0.93	-2.97	-1.74	-2.17	-3.21	-1.31	-0.58	-1.81
KFO	3.11	3.32	3.47	2.34	3.21	2.88	2.27	3.14	3.34	2.80
KFOM	2.54	3.11	3.26	1.79	3.02	2.56	0.86	2.94	2.92	2.62
KFN	-3.98	-3.23	-2.59	-4.93	-3.31	-4.28	-4.70	-2.87	-2.35	-4.06
KFNM	-1.93	0.10	0.17	-2.08	-0.85	-1.44	-3.69	-0.22	0.04	-1.04
KFI	-2.26	-0.13	0.08	-2.40	-0.78	-1.57	-3.46	-0.30	0.04	-1.20
KFIM	-1.30	0.80	0.94	-0.75	0.38	-0.52	-2.67	0.78	0.59	0.36
KFM	-0.71	0.87	0.90	-0.24	0.50	-0.20	-1.67	0.76	0.69	0.67
KFMM	-0.53	1.09	1.15	0.00	1.33	0.31	-1.31	1.13	0.92	1.21

Table A.14: Segmental SNRs in white noise at -5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-7.43	-7.13	-7.00	-8.22	-7.17	-7.95	-7.49	-6.54	-6.08	-7.15
SS	-7.06	-6.76	-6.41	-7.86	-6.33	-7.14	-7.40	-5.83	-5.35	-6.72
SSV	-6.61	-5.95	-5.21	-6.88	-5.61	-6.53	-7.28	-5.20	-4.76	-6.26
MMSE	-5.59	-4.08	-4.03	-5.77	-4.22	-4.74	-5.72	-3.96	-3.95	-4.38
KFO	0.91	1.21	1.44	0.33	1.32	1.05	-0.02	1.38	1.49	1.14
KFOM	0.40	0.98	1.40	-0.01	1.35	0.66	-0.95	1.35	0.99	1.11
KFN	-6.40	-5.96	-5.38	-7.21	-5.82	-6.67	-6.95	-5.23	-5.11	-6.17
KFNM	-4.44	-2.84	-2.50	-4.88	-3.26	-3.89	-5.93	-2.56	-2.33	-3.67
KFI	-3.31	-2.84	-2.39	-5.18	-3.17	-4.08	-4.43	-2.54	-2.06	-3.49
KFIM	-2.57	-1.10	-0.74	-2.96	-1.65	-2.40	-3.66	-0.94	-1.02	-1.40
KFM	-1.42	-0.36	-0.47	-1.87	-1.40	-1.95	-1.90	-0.32	-0.67	-0.72
KFMM	-1.52	-0.35	-0.01	-1.40	-0.41	-1.25	-1.74	0.11	-0.45	0.00

Table A.15: Segmental SNRs in white noise at -10 dB

A.4 WPESQ Scores in Colored Noise

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	2.18	1.74	1.78	1.97	1.92	1.79	2.12	2.00	2.29	1.84
SS	2.52	2.10	2.10	2.36	2.23	2.10	2.55	2.38	2.77	2.25
SSV	3.01	2.58	2.55	2.80	2.70	2.58	3.06	2.93	3.30	2.82
MMSE	1.89	1.76	1.75	1.91	1.89	1.54	1.83	1.57	1.56	1.51
KFO	2.84	2.16	2.69	2.54	2.64	2.61	2.47	2.63	2.61	2.68
KFOM	3.03	2.36	2.69	2.61	2.80	2.72	2.62	2.71	2.73	2.71
KFN	2.23	1.96	1.91	2.00	1.96	2.01	2.03	2.10	2.31	2.02
KFNM	2.57	2.26	2.15	2.25	2.19	2.27	2.31	2.32	2.50	2.23
KFI	2.02	1.68	1.59	1.63	1.55	1.84	1.52	1.58	1.69	1.59
KFIM	2.24	1.96	1.81	1.82	1.73	2.00	1.74	1.76	1.88	1.77
KFM	1.87	1.71	1.93	1.76	1.82	1.70	1.62	1.44	1.47	1.64
KFMM	2.02	1.98	2.11	1.88	2.06	1.81	1.69	1.68	1.56	1.75

Table A.16: WPESQ scores in colored noise at 20 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.74	1.40	1.41	1.53	1.57	1.43	1.59	1.53	1.67	1.43
SS	2.02	1.67	1.62	1.81	1.80	1.67	1.91	1.78	2.06	1.69
SSV	2.40	2.04	2.01	2.20	2.14	2.04	2.37	2.21	2.60	2.13
MMSE	1.84	1.61	1.62	1.80	1.80	1.50	1.79	1.45	1.90	1.57
KFO	2.70	2.29	2.52	2.45	2.70	2.33	2.29	2.30	2.30	2.31
KFOM	2.87	2.46	2.59	2.51	2.81	2.43	2.39	2.37	2.42	2.34
KFN	1.88	1.63	1.62	1.62	1.68	1.63	1.71	1.75	1.92	1.64
KFNM	2.20	1.96	1.90	1.94	1.93	1.92	2.00	2.01	2.18	1.90
KFI	1.78	1.55	1.47	1.46	1.46	1.61	1.42	1.50	1.58	1.47
KFIM	2.01	1.80	1.67	1.66	1.62	1.80	1.62	1.68	1.77	1.64
KFM	2.09	1.87	1.83	1.79	1.77	1.51	1.60	1.39	1.59	1.53
KFMM	2.26	2.20	2.05	1.94	1.98	1.62	1.68	1.54	1.75	1.56

Table A.17: WPESQ scores in colored noise at 15 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.40	1.19	1.18	1.27	1.32	1.22	1.27	1.24	1.31	1.22
SS	1.60	1.35	1.31	1.43	1.49	1.35	1.48	1.40	1.54	1.35
SSV	1.84	1.60	1.56	1.64	1.72	1.62	1.77	1.66	1.91	1.58
MMSE	1.76	1.44	1.44	1.62	1.65	1.54	1.73	1.34	1.83	1.54
KFO	2.66	2.56	2.47	2.39	2.68	2.11	1.97	1.95	1.98	1.97
KFOM	2.77	2.57	2.53	2.43	2.75	2.18	2.08	2.03	2.10	2.02
KFN	1.57	1.36	1.37	1.36	1.46	1.38	1.41	1.41	1.51	1.36
KFNM	1.84	1.64	1.61	1.62	1.68	1.57	1.62	1.64	1.74	1.56
KFI	1.55	1.38	1.35	1.32	1.38	1.38	1.31	1.37	1.44	1.34
KFIM	1.72	1.57	1.50	1.50	1.52	1.54	1.46	1.53	1.58	1.49
KFM	1.84	1.71	1.56	1.57	1.57	1.42	1.47	1.37	1.43	1.43
KFMM	1.98	1.94	1.75	1.76	1.74	1.51	1.58	1.44	1.54	1.55

Table A.18: WPESQ scores in colored noise at 10 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.23	1.08	1.07	1.13	1.16	1.09	1.11	1.09	1.12	1.10
SS	1.32	1.17	1.13	1.21	1.26	1.16	1.21	1.17	1.25	1.16
SSV	1.44	1.29	1.27	1.31	1.42	1.30	1.37	1.30	1.42	1.28
MMSE	1.52	1.27	1.26	1.40	1.47	1.32	1.45	1.27	1.46	1.35
KFO	2.45	2.58	2.21	2.38	2.42	1.89	1.72	1.74	1.72	1.75
KFOM	2.53	2.63	2.27	2.36	2.54	1.92	1.82	1.80	1.80	1.79
KFN	1.36	1.19	1.16	1.21	1.27	1.20	1.21	1.20	1.25	1.20
KFNM	1.54	1.39	1.36	1.38	1.45	1.31	1.36	1.35	1.42	1.31
KFI	1.37	1.23	1.21	1.20	1.29	1.24	1.21	1.21	1.28	1.23
KFIM	1.51	1.36	1.35	1.32	1.40	1.32	1.32	1.32	1.42	1.32
KFM	1.55	1.47	1.36	1.40	1.44	1.40	1.30	1.28	1.34	1.32
KFMM	1.71	1.66	1.52	1.55	1.58	1.49	1.41	1.35	1.41	1.41

Table A.19: WPESQ scores in colored noise at 5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.14	1.04	1.04	1.07	1.08	1.05	1.05	1.05	1.06	1.05
SS	1.18	1.06	1.05	1.10	1.11	1.07	1.08	1.07	1.10	1.07
SSV	1.21	1.13	1.10	1.13	1.19	1.13	1.15	1.12	1.17	1.11
MMSE	1.28	1.17	1.15	1.20	1.29	1.18	1.22	1.15	1.20	1.17
KFO	2.17	2.19	1.94	2.19	2.17	1.66	1.56	1.57	1.55	1.54
KFOM	2.24	2.27	1.98	2.16	2.25	1.66	1.59	1.58	1.58	1.55
KFN	1.22	1.09	1.08	1.12	1.15	1.09	1.10	1.08	1.11	1.09
KFNM	1.35	1.21	1.18	1.23	1.27	1.16	1.19	1.17	1.21	1.16
KFI	1.23	1.12	1.10	1.11	1.17	1.12	1.11	1.11	1.14	1.13
KFIM	1.35	1.22	1.21	1.19	1.27	1.19	1.19	1.17	1.23	1.19
KFM	1.36	1.28	1.24	1.23	1.35	1.28	1.20	1.22	1.23	1.22
KFMM	1.49	1.40	1.36	1.36	1.47	1.36	1.30	1.28	1.30	1.29

Table A.20: WPESQ scores in colored noise at 0 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.10	1.03	1.03	1.05	1.05	1.04	1.04	1.03	1.04	1.04
SS	1.11	1.03	1.03	1.06	1.06	1.04	1.04	1.04	1.05	1.04
SSV	1.11	1.05	1.04	1.06	1.08	1.06	1.05	1.05	1.07	1.05
MMSE	1.14	1.06	1.05	1.09	1.11	1.07	1.08	1.08	1.09	1.07
KFO	1.94	1.96	1.72	2.03	1.98	1.42	1.41	1.37	1.37	1.32
KFOM	1.98	1.99	1.75	1.98	2.03	1.40	1.38	1.36	1.38	1.32
KFN	1.15	1.05	1.05	1.08	1.09	1.06	1.05	1.05	1.06	1.06
KFNM	1.25	1.11	1.09	1.15	1.15	1.09	1.10	1.08	1.10	1.08
KFI	1.15	1.07	1.06	1.08	1.11	1.08	1.06	1.06	1.08	1.07
KFIM	1.23	1.13	1.10	1.13	1.16	1.11	1.10	1.09	1.11	1.11
KFM	1.18	1.23	1.11	1.12	1.25	1.14	1.12	1.13	1.14	1.12
KFMM	1.24	1.24	1.20	1.21	1.30	1.19	1.20	1.18	1.20	1.17

Table A.21: WPESQ scores in colored noise at -5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.08	1.03	1.03	1.03	1.05	1.04	1.03	1.03	1.03	1.03
SS	1.08	1.03	1.03	1.04	1.05	1.04	1.04	1.03	1.04	1.03
SSV	1.08	1.03	1.03	1.04	1.05	1.04	1.04	1.04	1.04	1.04
MMSE	1.10	1.03	1.03	1.05	1.04	1.04	1.04	1.04	1.04	1.04
KFO	1.71	1.71	1.47	1.86	1.78	1.24	1.28	1.21	1.22	1.16
KFOM	1.71	1.70	1.47	1.73	1.77	1.22	1.25	1.19	1.22	1.16
KFN	1.12	1.04	1.04	1.06	1.07	1.06	1.04	1.04	1.05	1.05
KFNM	1.18	1.07	1.06	1.10	1.10	1.07	1.06	1.05	1.06	1.05
KFI	1.13	1.05	1.06	1.07	1.09	1.06	1.08	1.05	1.05	1.06
KFIM	1.19	1.09	1.07	1.10	1.11	1.08	1.07	1.06	1.07	1.07
KFM	1.11	1.09	1.05	1.07	1.09	1.08	1.05	1.06	1.06	1.07
KFMM	1.13	1.12	1.08	1.10	1.11	1.11	1.07	1.08	1.09	1.08

Table A.22: WPESQ scores in colored noise at -10 dB

A.5 Segmental SNRs in Colored Noise

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	11.30	12.75	12.26	10.11	10.01	9.38	10.61	11.64	12.44	10.69
SS	12.91	15.02	14.11	11.64	11.86	11.08	12.12	13.22	13.85	12.89
SSV	14.61	15.80	15.20	12.90	13.45	12.64	13.57	14.41	14.76	14.47
MMSE	0.42	0.18	1.07	0.24	-1.23	-0.69	0.23	0.38	0.42	-1.92
KFO	11.82	11.38	13.33	8.84	11.48	11.40	10.77	11.46	12.35	11.88
KFOM	8.13	9.57	8.72	5.56	8.93	8.34	5.03	7.04	8.12	9.29
KFN	10.32	12.16	12.14	6.93	9.37	8.29	8.47	10.06	10.66	9.37
KFNM	6.71	8.98	8.02	4.36	7.50	6.32	3.43	6.42	6.83	8.15
KFI	8.44	10.85	11.20	6.11	8.52	6.67	7.11	8.27	8.81	8.35
KFIM	6.59	8.51	7.68	4.35	7.13	5.80	3.13	5.78	6.17	7.81
KFM	3.93	1.42	3.58	2.59	2.02	0.97	-1.12	2.70	2.46	1.84
KFMM	4.07	1.89	2.41	2.40	2.16	0.56	-0.09	2.36	2.60	1.42

Table A.23: Segmental SNRs in colored noise at 20 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	7.27	8.27	7.96	5.88	6.03	5.52	6.50	7.38	8.29	6.28
SS	9.12	10.99	10.00	7.70	8.00	7.41	8.23	9.31	10.04	8.66
SSV	10.77	12.20	11.60	9.16	9.73	9.06	9.71	10.80	11.02	10.44
MMSE	1.01	0.20	2.48	0.30	0.35	-2.10	0.79	1.34	3.46	-0.28
KFO	9.86	9.52	11.28	7.57	9.97	9.93	9.38	10.02	10.71	10.13
KFOM	7.41	8.74	8.05	4.88	8.25	7.66	4.66	6.81	7.65	8.53
KFN	7.72	9.09	8.76	4.68	6.54	5.88	6.05	7.68	8.31	6.52
KFNM	5.74	7.78	6.99	3.54	6.21	5.25	2.47	5.59	6.07	6.75
KFI	6.97	8.44	8.35	4.47	6.15	5.13	5.16	6.55	7.05	6.19
KFIM	5.33	7.43	6.67	3.40	5.91	4.81	2.41	4.85	5.31	6.46
KFM	2.88	3.71	4.57	4.55	5.08	-0.54	-0.38	3.53	4.93	1.85
KFMM	1.12	3.43	3.25	3.94	4.63	-1.28	-0.14	2.93	4.03	2.46

Table A.24: Segmental SNRs in colored noise at 15 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	3.44	4.06	3.88	1.89	2.31	1.88	2.67	3.41	4.41	2.15
SS	5.41	7.02	6.01	3.81	4.41	3.92	4.50	5.42	6.33	4.65
SSV	7.25	8.75	7.91	5.64	6.28	5.75	6.03	7.21	7.69	6.67
MMSE	2.58	2.18	2.65	0.98	0.51	-3.27	2.36	3.18	5.60	3.73
KFO	8.58	9.00	8.80	6.46	8.20	8.15	7.71	8.30	8.82	8.09
KFOM	5.64	7.49	7.39	4.34	7.31	6.81	4.38	6.35	6.94	7.28
KFN	4.58	5.55	5.23	1.75	3.49	2.89	2.90	4.54	5.35	3.16
KFNM	4.34	6.09	5.41	2.10	4.44	3.58	0.74	4.25	4.86	4.67
KFI	4.54	5.57	5.28	2.15	3.67	2.67	2.75	4.18	4.77	3.31
KFIM	4.01	5.61	5.22	1.89	4.28	3.12	0.97	3.79	4.26	4.42
KFM	3.17	5.59	4.80	3.51	4.72	-1.46	2.20	3.57	5.37	4.33
KFMM	1.79	4.86	3.98	3.38	4.98	-1.81	0.78	3.62	4.44	4.50

Table A.25: Segmental SNRs in colored noise at 10 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-0.06	0.13	0.11	-1.71	-1.04	-1.52	-0.79	-0.14	0.82	-1.49
SS	1.89	3.24	2.17	0.19	1.02	0.58	0.95	1.78	2.88	0.93
SSV	3.84	5.15	4.14	2.20	2.96	2.56	2.37	3.71	4.33	2.88
MMSE	3.96	2.07	3.18	1.83	2.38	-2.64	1.67	3.15	4.59	4.05
KFO	6.96	7.23	7.19	5.27	6.29	6.26	6.19	6.58	6.82	6.03
KFOM	5.82	6.68	6.38	3.90	6.02	5.66	3.45	5.60	5.93	5.71
KFN	1.21	1.88	1.62	-1.51	0.30	-0.46	-0.63	1.00	2.01	-0.46
KFNM	2.28	3.78	3.27	-0.06	2.27	1.32	-1.70	2.42	3.02	1.98
KFI	1.50	2.40	2.07	-0.75	0.87	-0.15	-1.38	1.14	2.04	-0.09
KFIM	1.50	3.45	3.20	0.29	2.20	1.07	-2.13	2.12	2.65	1.78
KFM	3.95	4.39	4.35	2.35	3.55	0.39	1.18	3.21	3.73	3.66
KFMM	3.73	4.09	4.24	2.54	3.88	0.40	0.22	3.34	3.44	4.05

Table A.26: Segmental SNRs in colored noise at 5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-3.04	-3.29	-3.17	-4.70	-3.91	-4.36	-3.89	-3.16	-2.35	-4.29
SS	-1.39	-0.40	-1.20	-3.07	-2.03	-2.47	-2.35	-1.50	-0.37	-2.37
SSV	0.21	1.70	0.39	-1.44	-0.43	-0.71	-0.97	0.19	0.90	-0.66
MMSE	1.86	2.68	2.23	0.27	1.12	0.41	0.46	1.57	2.33	1.42
KFO	5.11	5.35	5.31	3.93	4.64	4.42	4.56	4.99	4.91	4.16
KFOM	4.92	5.17	5.06	3.28	4.59	4.31	2.45	4.63	4.59	4.12
KFN	-2.07	-1.73	-1.80	-4.69	-2.79	-3.48	-4.00	-2.43	-1.49	-3.72
KFNM	-0.50	1.02	0.61	-2.57	-0.31	-1.20	-3.86	-0.08	0.66	-0.99
KFI	-2.02	-1.08	-1.27	-3.96	-2.16	-3.29	-4.71	-1.98	-1.13	-3.29
KFIM	-0.82	0.88	0.61	-2.06	-0.34	-1.61	-4.31	-0.28	0.44	-1.18
KFM	1.50	2.43	2.43	0.36	1.20	0.82	-0.70	2.11	1.73	1.31
KFMM	1.52	2.72	2.75	1.11	1.90	1.27	-1.40	2.43	1.97	1.95

Table A.27: Segmental SNRs in colored noise at 0 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-5.42	-5.89	-5.49	-6.78	-6.17	-6.42	-5.91	-5.45	-4.64	-6.18
SS	-4.44	-3.76	-4.13	-5.83	-4.80	-4.87	-4.87	-4.34	-3.52	-4.89
SSV	-3.28	-2.07	-2.92	-4.75	-3.55	-3.83	-3.92	-3.11	-2.83	-3.84
MMSE	-1.60	-0.44	-0.61	-3.01	-1.85	-1.84	-2.40	-0.34	-0.71	-1.55
KFO	3.38	3.55	3.55	2.46	3.08	2.70	2.36	3.40	3.08	2.49
KFOM	3.37	3.55	3.65	2.13	3.07	2.73	1.72	3.44	3.00	2.56
KFN	-5.05	-4.74	-4.75	-6.88	-5.49	-5.89	-6.95	-5.18	-4.51	-5.99
KFNM	-3.42	-2.25	-2.37	-5.10	-3.05	-3.75	-5.88	-2.98	-2.22	-3.68
KFI	-5.21	-4.33	-4.30	-6.40	-4.88	-5.88	-6.89	-4.74	-4.23	-5.74
KFIM	-3.96	-2.40	-2.34	-4.72	-3.09	-4.16	-5.78	-3.01	-2.30	-3.78
KFM	-1.77	0.19	0.07	-2.07	-0.95	-1.50	-3.16	-0.18	-0.49	-0.94
KFMM	-1.34	0.83	0.66	-1.21	-0.17	-0.76	-2.93	0.58	0.07	-0.03

Table A.28: Segmental SNRs in colored noise at -5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-7.18	-7.60	-6.81	-8.02	-7.43	-7.84	-7.28	-6.62	-5.97	-7.44
SS	-6.34	-6.44	-6.28	-7.86	-6.96	-6.85	-6.90	-5.93	-5.36	-6.66
SSV	-5.93	-5.39	-5.51	-7.37	-6.00	-6.35	-6.10	-5.76	-5.29	-6.41
MMSE	-4.75	-4.10	-4.14	-5.68	-4.69	-4.46	-4.93	-3.37	-4.12	-4.49
KFO	1.63	1.66	1.76	0.85	1.44	1.04	0.75	1.73	1.30	0.85
KFOM	1.67	1.82	2.05	0.77	1.39	0.96	0.48	2.00	1.33	0.99
KFN	-6.83	-6.88	-6.60	-8.19	-7.15	-7.56	-7.61	-6.32	-5.85	-7.04
KFNM	-5.92	-5.28	-5.18	-7.10	-5.60	-5.92	-6.98	-5.33	-4.74	-5.82
KFI	-6.64	-6.72	-6.46	-7.97	-7.04	-7.60	-7.38	-6.13	-5.54	-6.83
KFIM	-6.02	-5.46	-5.34	-7.05	-5.70	-6.17	-6.71	-5.39	-4.62	-5.74
KFM	-3.26	-2.49	-2.21	-4.88	-3.07	-3.58	-4.24	-2.28	-2.30	-2.89
KFMM	-3.58	-1.31	-1.33	-3.81	-2.08	-2.67	-3.86	-1.20	-1.67	-1.70

Table A.29: Segmental SNRs in colored noise at -10 dB

A.6 WPESQ Scores in Street Noise

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	2.62	2.25	2.26	2.12	2.28	2.03	2.07	2.06	2.33	1.89
SS	2.97	2.47	2.50	2.40	2.49	2.22	2.24	2.19	2.47	2.04
SSV	3.11	2.81	2.69	2.63	2.66	2.42	2.44	2.43	2.53	2.20
MMSE	2.22	1.94	1.65	1.90	1.83	1.50	1.50	1.35	1.72	1.43
KFO	2.84	2.87	2.51	2.92	2.71	1.99	1.99	2.04	2.02	1.92
KFOM	2.80	2.77	2.36	2.67	2.69	1.93	2.06	2.05	2.07	1.86
KFN	2.49	2.38	2.24	2.16	2.22	1.98	1.87	2.00	2.04	1.85
KFNM	2.41	2.23	2.09	2.00	2.14	1.95	1.93	1.99	2.09	1.80
KFI	2.17	1.91	1.69	1.70	1.77	1.74	1.42	1.51	1.40	1.48
KFIM	2.16	2.03	1.77	1.74	1.82	1.79	1.57	1.65	1.54	1.54
KFM	2.17	1.88	1.74	1.70	1.83	1.43	1.53	1.46	1.48	1.44
KFMM	2.19	2.03	1.84	1.65	1.87	1.43	1.53	1.51	1.52	1.47

Table A.30: WPESQ scores in street noise at 20 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.99	1.68	1.73	1.61	1.75	1.55	1.59	1.57	1.71	1.44
SS	2.24	1.85	1.92	1.77	1.88	1.71	1.77	1.70	1.88	1.55
SSV	2.41	2.13	2.04	1.94	2.01	1.95	1.95	1.92	2.02	1.64
MMSE	2.47	2.09	1.73	1.99	2.14	1.66	1.69	1.50	1.79	1.51
KFO	2.27	2.33	2.10	2.32	2.34	1.60	1.71	1.73	1.67	1.59
KFOM	2.28	2.31	2.02	2.28	2.25	1.53	1.75	1.69	1.67	1.51
KFN	1.97	1.82	1.80	1.65	1.75	1.59	1.57	1.61	1.70	1.46
KFNM	1.95	1.76	1.72	1.69	1.72	1.55	1.58	1.59	1.68	1.44
KFI	1.82	1.63	1.52	1.45	1.52	1.55	1.33	1.40	1.33	1.33
KFIM	1.86	1.68	1.57	1.54	1.56	1.53	1.44	1.48	1.42	1.36
KFM	1.94	1.91	1.61	1.62	1.72	1.35	1.49	1.40	1.36	1.42
KFMM	1.95	1.98	1.69	1.67	1.74	1.30	1.55	1.43	1.39	1.41

Table A.31: WPESQ scores in street noise at 15 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.57	1.33	1.34	1.35	1.43	1.25	1.26	1.25	1.30	1.18
SS	1.76	1.44	1.47	1.44	1.52	1.37	1.40	1.36	1.43	1.27
SSV	1.81	1.59	1.56	1.50	1.55	1.59	1.57	1.53	1.60	1.36
MMSE	2.07	1.93	1.88	1.85	1.88	1.70	1.79	1.75	1.77	1.54
KFO	1.88	2.00	1.73	2.02	1.92	1.36	1.45	1.44	1.40	1.34
KFOM	1.80	2.00	1.66	1.99	1.95	1.32	1.44	1.37	1.36	1.27
KFN	1.60	1.44	1.41	1.38	1.46	1.31	1.31	1.30	1.34	1.20
KFNM	1.58	1.45	1.39	1.48	1.46	1.26	1.30	1.27	1.30	1.17
KFI	1.52	1.38	1.32	1.29	1.36	1.31	1.23	1.23	1.23	1.18
KFIM	1.53	1.41	1.35	1.40	1.40	1.27	1.27	1.23	1.27	1.17
KFM	1.54	1.66	1.44	1.48	1.53	1.32	1.34	1.32	1.34	1.28
KFMM	1.62	1.70	1.48	1.55	1.54	1.30	1.40	1.32	1.35	1.26

Table A.32: WPESQ scores in street noise at 10 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.29	1.14	1.13	1.19	1.21	1.09	1.11	1.10	1.11	1.07
SS	1.37	1.20	1.19	1.23	1.27	1.17	1.16	1.15	1.16	1.11
SSV	1.48	1.29	1.25	1.27	1.31	1.29	1.24	1.26	1.25	1.17
MMSE	1.56	1.52	1.47	1.49	1.52	1.50	1.50	1.51	1.53	1.29
KFO	1.59	1.68	1.44	1.71	1.71	1.21	1.27	1.21	1.21	1.18
KFOM	1.51	1.63	1.37	1.68	1.69	1.17	1.23	1.14	1.17	1.13
KFN	1.33	1.21	1.17	1.22	1.26	1.12	1.13	1.12	1.13	1.08
KFNM	1.30	1.23	1.17	1.28	1.26	1.09	1.13	1.10	1.10	1.06
KFI	1.32	1.20	1.16	1.19	1.24	1.13	1.12	1.12	1.12	1.08
KFIM	1.30	1.23	1.17	1.26	1.24	1.10	1.13	1.10	1.10	1.06
KFM	1.40	1.41	1.31	1.32	1.38	1.23	1.24	1.20	1.24	1.19
KFMM	1.43	1.44	1.33	1.39	1.40	1.19	1.26	1.19	1.23	1.18

Table A.33: WPESQ scores in street noise at 5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.13	1.06	1.05	1.09	1.10	1.04	1.05	1.04	1.04	1.03
SS	1.15	1.08	1.07	1.11	1.12	1.07	1.07	1.06	1.05	1.05
SSV	1.21	1.11	1.10	1.14	1.15	1.11	1.09	1.10	1.07	1.07
MMSE	1.29	1.23	1.18	1.24	1.32	1.21	1.21	1.21	1.20	1.14
KFO	1.39	1.41	1.24	1.51	1.48	1.12	1.13	1.09	1.10	1.09
KFOM	1.32	1.34	1.19	1.46	1.42	1.10	1.10	1.06	1.07	1.06
KFN	1.15	1.08	1.07	1.11	1.13	1.05	1.05	1.05	1.04	1.03
KFNM	1.12	1.08	1.07	1.13	1.13	1.04	1.05	1.04	1.04	1.03
KFI	1.16	1.10	1.07	1.11	1.13	1.05	1.05	1.05	1.04	1.03
KFIM	1.13	1.09	1.07	1.13	1.13	1.04	1.05	1.04	1.04	1.03
KFM	1.29	1.28	1.18	1.25	1.26	1.17	1.12	1.12	1.12	1.12
KFMM	1.28	1.29	1.18	1.29	1.28	1.13	1.13	1.10	1.11	1.10

Table A.34: WPESQ scores in street noise at 0 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.06	1.03	1.03	1.05	1.05	1.03	1.03	1.03	1.03	1.02
SS	1.06	1.03	1.03	1.05	1.05	1.04	1.04	1.03	1.03	1.03
SSV	1.07	1.04	1.04	1.05	1.06	1.04	1.05	1.04	1.03	1.03
MMSE	1.12	1.07	1.07	1.11	1.14	1.07	1.06	1.06	1.06	1.06
KFO	1.24	1.21	1.13	1.30	1.27	1.08	1.07	1.05	1.05	1.05
KFOM	1.18	1.17	1.10	1.23	1.23	1.06	1.06	1.04	1.04	1.04
KFN	1.07	1.04	1.04	1.07	1.06	1.03	1.03	1.03	1.03	1.02
KFNM	1.06	1.04	1.03	1.09	1.07	1.03	1.03	1.03	1.03	1.02
KFI	1.07	1.05	1.04	1.08	1.07	1.03	1.03	1.03	1.03	1.03
KFIM	1.06	1.05	1.04	1.08	1.07	1.03	1.03	1.03	1.03	1.02
KFM	1.13	1.16	1.10	1.13	1.16	1.08	1.06	1.06	1.06	1.06
KFMM	1.11	1.14	1.09	1.13	1.15	1.07	1.07	1.06	1.05	1.05

Table A.35: WPESQ scores in street noise at -5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	1.04	1.02	1.02	1.04	1.04	1.03	1.07	1.02	1.02	1.02
SS	1.04	1.02	1.02	1.12	1.03	1.03	1.03	1.02	1.02	1.02
SSV	1.04	1.02	1.03	1.03	1.03	1.03	1.03	1.03	1.02	1.02
MMSE	1.05	1.03	1.03	1.04	1.05	1.03	1.03	1.03	1.03	1.03
KFO	1.14	1.12	1.07	1.16	1.16	1.06	1.05	1.04	1.04	1.04
KFOM	1.11	1.10	1.06	1.13	1.15	1.05	1.05	1.04	1.03	1.04
KFN	1.04	1.03	1.03	1.05	1.05	1.03	1.03	1.03	1.02	1.02
KFNM	1.04	1.03	1.03	1.06	1.05	1.03	1.03	1.03	1.02	1.02
KFI	1.04	1.03	1.03	1.08	1.05	1.03	1.03	1.03	1.02	1.02
KFIM	1.04	1.03	1.03	1.06	1.05	1.03	1.03	1.03	1.02	1.02
KFM	1.05	1.06	1.04	1.07	1.08	1.04	1.03	1.03	1.03	1.03
KFMM	1.04	1.05	1.03	1.07	1.07	1.04	1.03	1.03	1.03	1.03

Table A.36: WPESQ scores in street noise at -10 dB

A.7 Segmental SNRs in Street Noise

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	11.49	12.76	12.43	10.21	10.30	9.66	10.71	11.85	12.49	10.85
SS	12.39	13.72	13.72	11.47	11.22	10.63	11.71	12.94	13.02	12.06
SSV	12.70	14.50	13.81	11.89	11.62	11.62	12.58	13.84	13.70	13.31
MMSE	1.90	1.21	1.74	-0.92	-0.04	-2.47	-0.31	0.73	2.67	-0.24
KFO	11.31	12.78	12.64	8.27	10.94	9.59	9.88	10.55	11.55	9.84
KFOM	7.47	8.73	7.45	4.62	7.53	6.47	3.70	6.89	7.26	7.19
KFN	9.73	11.60	11.27	6.43	8.90	7.34	8.02	9.25	9.88	8.28
KFNM	5.71	7.50	5.99	2.84	5.48	4.25	1.64	5.50	5.49	5.69
KFI	7.93	10.73	10.69	5.62	8.14	5.78	6.79	7.72	3.02	7.32
KFIM	5.71	7.27	5.90	2.98	5.34	4.08	1.60	5.19	5.42	5.67
KFM	5.97	2.20	2.55	2.05	2.00	-0.83	-1.91	1.22	3.19	0.99
KFMM	3.56	0.73	0.94	0.58	0.30	-3.01	-2.93	-0.09	2.12	0.49

Table A.37: Segmental SNRs in street noise at 20 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	7.45	8.32	8.06	5.93	6.25	5.81	6.57	7.46	8.33	6.38
SS	8.53	9.51	9.60	7.55	7.35	6.95	7.63	8.75	9.10	7.88
SSV	8.86	10.54	9.87	8.26	7.63	8.05	8.74	9.68	9.77	9.09
MMSE	5.74	4.45	3.98	1.51	0.98	-0.46	1.57	2.31	3.91	1.48
KFO	8.99	9.63	9.65	6.72	8.38	7.69	7.81	8.24	9.32	7.71
KFOM	6.22	6.99	6.13	3.72	6.04	5.20	2.70	5.88	6.02	5.50
KFN	7.02	8.10	7.73	3.96	5.84	4.82	5.32	6.47	7.29	5.35
KFNM	4.06	5.32	4.11	0.98	3.57	2.48	0.02	4.02	3.96	3.25
KFI	6.00	7.65	7.39	3.57	5.44	3.90	4.47	5.60	2.61	4.72
KFIM	4.19	5.20	4.07	1.15	3.51	2.47	0.01	3.84	3.41	3.34
KFM	6.29	5.68	5.33	3.97	4.10	-1.00	0.22	3.36	3.42	2.15
KFMM	4.70	3.38	1.77	2.05	3.63	-2.97	-2.63	1.55	2.51	2.08

Table A.38: Segmental SNRs in street noise at 15 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	3.62	4.13	3.80	1.84	2.40	2.15	2.74	3.50	4.38	2.32
SS	4.93	5.46	5.66	3.79	3.77	3.55	3.90	4.91	5.36	3.87
SSV	5.35	6.54	6.06	4.59	4.27	4.59	4.96	5.98	6.17	5.11
MMSE	5.27	6.20	6.03	3.51	3.54	0.61	3.56	5.94	4.25	1.23
KFO	6.55	6.80	6.86	4.98	5.92	5.64	5.80	5.93	6.86	5.52
KFOM	4.56	4.89	4.45	2.67	4.32	3.65	1.59	4.43	4.28	3.67
KFN	3.79	4.31	3.95	0.90	2.50	1.84	2.35	3.31	4.24	1.98
KFNM	1.55	2.37	1.49	-1.36	1.03	0.12	-1.83	1.81	1.74	0.34
KFI	3.54	4.18	3.83	0.85	2.44	1.43	1.80	1.26	3.73	1.69
KFIM	1.91	2.36	1.53	-1.38	1.05	0.23	-1.84	1.03	1.60	0.52
KFM	3.72	5.51	5.11	3.04	3.92	0.18	2.72	4.13	3.81	0.40
KFMM	3.15	4.25	3.42	0.91	2.91	-0.89	-0.07	3.07	2.49	1.09

Table A.39: Segmental SNRs in street noise at 10 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	0.17	0.28	-0.19	-1.93	-0.97	-1.23	-0.72	0.08	0.60	-1.27
SS	1.55	1.70	1.91	0.10	0.50	0.30	0.53	1.50	1.88	0.21
SSV	2.08	2.71	2.57	0.93	1.27	1.25	1.34	2.44	2.75	1.19
MMSE	2.75	4.12	3.94	2.49	2.17	2.47	2.02	3.92	3.76	0.72
KFO	4.19	4.25	4.33	3.22	3.68	3.63	3.85	3.83	4.44	3.49
KFOM	2.56	2.82	2.57	1.47	2.48	1.97	0.52	2.79	2.41	1.98
KFN	0.26	0.61	0.13	-2.34	-0.71	-1.29	-0.64	0.19	0.89	-1.48
KFNM	-1.51	-1.01	-1.58	-3.92	-1.80	-2.61	-3.81	-0.82	-1.14	-2.72
KFI	0.49	0.70	0.21	-2.27	-0.66	-1.40	-0.95	0.11	0.82	-1.56
KFIM	-1.30	-1.01	-1.48	-3.94	-1.73	-2.43	-3.39	-0.83	-1.08	-2.61
KFM	2.01	3.01	2.86	0.97	1.71	0.88	0.95	2.21	2.39	1.06
KFMM	1.43	2.31	1.82	-0.44	1.02	-0.09	-1.48	1.42	1.36	0.40

Table A.40: Segmental SNRs in street noise at 5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-2.72	-3.19	-3.36	-4.84	-3.84	-4.27	-3.60	-2.71	-2.54	-4.15
SS	-1.55	-1.75	-1.56	-3.16	-2.39	-2.81	-2.31	-1.41	-1.29	-2.92
SSV	-1.43	-0.98	-0.68	-2.59	-1.66	-1.90	-1.73	-0.71	-0.79	-2.46
MMSE	0.09	1.12	1.30	-0.33	-0.09	-0.05	-0.25	0.71	1.36	0.69
KFO	2.07	2.01	2.13	1.53	1.73	1.67	1.85	2.01	2.34	1.61
KFOM	0.89	0.71	0.71	0.16	0.61	0.35	-1.00	1.26	0.53	0.32
KFN	-2.87	-2.83	-3.11	-5.05	-3.57	-4.30	-3.53	-2.61	-2.23	-4.47
KFNM	-4.11	-4.26	-4.46	-6.09	-4.56	-5.06	-5.90	-3.31	-3.70	-5.27
KFI	-2.89	-2.64	-3.05	-4.95	-3.49	-4.29	-3.57	-2.62	-2.16	-4.49
KFIM	-4.15	-4.15	-4.39	-6.08	-4.48	-5.18	-5.75	-3.32	-3.65	-5.21
KFM	-0.16	0.65	0.53	-1.15	-0.26	-0.96	-0.90	0.17	0.34	-0.25
KFMM	-1.12	0.05	-0.13	-2.15	-0.84	-1.53	-2.35	-0.59	-0.29	-1.15

Table A.41: Segmental SNRs in street noise at 0 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-5.00	-5.86	-5.83	-6.98	-6.01	-6.55	-5.95	-4.77	-5.03	-6.10
SS	-4.16	-4.80	-4.18	-5.89	-4.76	-5.34	-4.73	-3.83	-4.07	-5.48
SSV	-4.42	-4.48	-3.53	-5.62	-4.55	-4.87	-4.73	-3.72	-3.55	-5.48
MMSE	-2.91	-2.12	-1.30	-3.33	-2.83	-2.66	-2.62	-2.19	-1.09	-2.37
KFO	0.28	0.67	0.02	0.79	-0.13	0.77	-0.14	0.51	0.91	-0.18
KFOM	-0.60	-1.45	-1.04	-1.49	-1.02	-1.60	-2.57	-0.10	-1.23	-1.34
KFN	-5.13	-5.59	-5.74	-7.06	-5.81	-6.46	-6.13	-4.78	-4.88	-6.46
KFNM	-6.04	-6.59	-6.34	-7.69	-6.15	-7.02	-7.42	-5.24	-5.70	-7.02
KFI	-5.18	-5.40	-5.65	-6.94	-5.75	-6.60	-6.04	-4.91	-4.81	-6.45
KFIM	-6.01	-6.50	-6.29	-7.64	-6.01	-7.21	-7.40	-5.31	-5.64	-6.98
KFM	-1.73	-1.02	-1.21	-2.57	-2.06	-2.44	-2.54	-1.30	-1.33	-1.77
KFMM	-2.92	-2.86	-2.16	-4.24	-3.17	-3.53	-5.24	-2.36	-2.12	-3.72

Table A.42: Segmental SNRs in street noise at -5 dB

Algos.	male1	male2	male3	male4	male5	female1	female2	female3	female4	female5
Noisy	-6.65	-7.22	-6.77	-8.15	-6.91	-7.95	-7.54	-5.96	-5.97	-7.22
SS	-5.90	-6.78	-5.85	-7.34	-6.17	-7.20	-6.75	-5.42	-5.65	-6.92
SSV	-5.84	-6.75	-5.40	-7.46	-6.08	-7.13	-6.77	-5.15	-5.38	-7.06
MMSE	-5.17	-5.65	-4.91	-6.57	-5.83	-5.54	-5.98	-4.43	-3.91	-5.35
KFO	-1.06	-0.12	-1.70	0.04	-1.41	0.04	-1.64	-0.70	0.19	-1.78
KFOM	-2.02	-3.20	-2.26	-2.69	-2.35	-3.32	-3.76	-1.21	-2.18	-2.76
KFN	-6.73	-7.39	-6.77	-8.29	-6.81	-7.96	-7.57	-6.03	-6.15	-7.28
KFNM	-6.92	-7.49	-6.88	-8.32	-6.84	-8.00	-7.61	-6.13	-6.56	-7.43
KFI	-6.71	-7.51	-6.75	-8.26	-6.70	-7.92	-7.06	-6.05	-6.39	-7.22
KFIM	-6.88	-7.52	-6.86	-8.29	-6.74	-7.99	-7.61	-6.13	-6.52	-7.42
KFM	-2.54	-1.78	-2.04	-3.49	-2.91	-3.24	-3.38	-1.93	-2.01	-2.46
KFMM	-4.35	-4.36	-3.86	-6.21	-4.90	-5.07	-5.63	-3.55	-3.24	-4.92

Table A.43: Segmental SNRs in street noise at -10 dB

COMPLETE SIMULATION RESULTS

Segmental SNRs in Street Noise

REFERENCES

[1] D. O'Shaughnessy and L. Deng, *Speech Processing A Dynamic and Optimization-Oriented Approach.* Marcel Dekker, Inc., 2003.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hall, 2002.

[3] N. Ma, M. Bouchard and R. Goubran, "Speech Enhancement Using a Perceptually Masking Threshold Constrained Kalman Filter and its Heuristic Implementations," *Special Issue on Statistical and Perceptual Audio Processing, IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 19-32, Jan. 2006

[4] N. Ma, M. Bouchard and R. Goubran, "A Kalman Filter with a Perceptual Post-filter to Enhance Speech Degraded by Colored Noise," *Journal of the Canadian Acoustical Association (Proceedings of Acoustics Week in Canada 2004)*, vol. 32, no. 3, pp. 138-139, Ottawa, Ontario, Oct. 2004

[5] N. Ma, M. Bouchard and R. Goubran, "Frequency and Time Domain Auditory Masking Threshold Constrained Kalman Filter for Speech Enhancement," *Proceedings of IEEE Seventh International Conference on Signal Processing (ICSP'04)*, vol. 3, pp. 2659-2662, Beijing, China, Aug. 31-Sept. 4 2004

[6] N. Ma, M. Bouchard and R. Goubran, "Perceptual Kalman filtering for speech enhancement in colored noise," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2004*, vol. 1, pp.717-720, Montreal, May 2004.

[7] N. Ma, M. Bouchard and R. Goubran, "Constrained Dual Unscented Kalman Filter for Enhancing Speech Degraded by Colored Noise," *International Transactions on Communication & Signal Processing*, vol. 7, no. 1, pp. 170-188, June 2006.

[8] *Wideband Speech Coding Standards and Applications*. AMR-WB White Paper, VoiceAge, 2005.

[9] *Wideband Coding of Speech at Around 16 kbits/s Using Adaptive Multi-Rate Wideband (AMR-WB)*, Geneva, ITU-T Recommendation G.722.2, 2002

[10] A. J. Heron, R. S. Turnbull and P. J. Hughes, "Communicating Naturally - The Opportunities of Wideband Coding," *BT Technology Journal*, vol. 24, no. 2, pp. 159-166, April 2006

[11] Institute of Hearing Research, "Intensity and Decibels," http://www.ihr.mrc.ac.uk/research/intensity_decibels.php, 2006.

[12] Hyperphysics, "Loudness," http://hyperphysics.phy-astr.gsu.edu/hbase/sound/loud.html, Georgia State University, 2005.

[13] The Physics Classroom, "Intensity and the decibel scale," http://www.glenbrook.k12.il.us/GBSSCI/PHYS/CLASS/sound/u11l2b.html

[14] J. Wolfe, "What is a decibel?," http://www.phys.unsw.edu.au/~jw/dB.html#log, University of New South Wales, 2005.

[15] A. Akmajian, R. A. Demers and R. M. Harnish, *Linguistics: An Introduction to Language and Communication*, The MIT Press, 2nd ed., 1984.

[16] T. Islam, *Interpolation of Linear Prediction Coefficients for Speech Coding*, Master thesis, McGill University, 2000.

[17] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing - Principles, Algorithms, and Applications*. Prentice Hall, 3rd ed., 1996.

[18] D. O'Shaughnessy, *Speech Communications - Human and Machine*. IEEE Press, 2nd ed., 2000.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 4th ed., 1997.

[20] E. Terhardt, G. Stoll and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, vol. 71, March 1982.

[21] Critical Bands - Wikipedia the free encyclopedia, http://en.wikipedia.org/wiki/Critical_bands, 2006

[22] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *Journal of the Acoustical Society of America*, vol. 68, pp. 1523-1525, Novermber 1980.

[23] J. Thiemann, *Acoustic Noise Suppression for Speech Signals using Auditory Masking Effects*, Master thesis, McGill University, 2001.

[24] R. P. Hellman, "Asymmetry of masking between noise and tone," *Perception & Psychophysics*, vol. 11, pp. 241-246, March 1972.

[25] M. R. Schroeder, B. S. Atal and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of Acoustical Society of America*, vol. 66, no. 6, pp. 1647-1651, December 1979.

[26] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314-323, February 1988.

[27] B. Scharf, *Foundations of Modern Auditory Theory*. New York: Academic, 1970.

[28] W. Jesteadt, S. Bacon and J. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *Journal of Acoustical Society of America*, vol. 71, no. 4, pp. 950-962, April 1982.

[29] Y. Huang and T. Chiueh, "A New Audio Coding Scheme Using a Forward Masking and Perceptually Weighted Vector Quantization," *IEEE Transactions Speech Audio Processing*, vol 10, no. 5, pp. 325-335, July 2002.

[30] E. Zwicker, "Procedure for calculating loudness of temporal variable sounds", *Journal of the Acoustical Society of America*, vol. 62, pp. 675-682, September 1977.

[31] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*. Springer-Verlag, 2nd ed., 1999.

[32] R. A. Garcia, *Digital watermarking of audio signals using a pshychoacoustic auditory model and spread spectrum theory*. University of Miami, 1999.

[33] D. O'Shaughnessy, "Enhancing speech degraded by additive noise of interfering speakers," *IEEE Communications Magazine*, vol. 27, no. 2, pp. 46-52, February 1989.

[34] *Methods for subjective determination of transmission quality*, Geneva, ITU-T Recommendation P.800, 1996

[35] *Subjective performance assessment of telephone-band and wideband digital codecs*, Geneva, ITU-T Recommendation P.830, 1996

[36] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Geneva, ITU-T Recommendation P.862, 2001

[37] *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, Geneva, ITU-T Recommendation P.862.2, 2005

[38] *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, Geneva, ITU-T Recommendation P.861 (withdrawn), 1998

[39] *Method for the subjective assessment of intermediate quality levels of coding systems*, Geneva, ITU-R Recommendation BS.1534, 2001

[40] M. Klein, *Signal Subspace Speech Enhancement With Perceptual Post-Filtering*, Master thesis, McGill University, 2002.

[41] J. H. L. Hansen and B. L. Pellom, *An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms*, Robust Speech Processing Laboratory, Duke University, 1998.

[42] S. R. Quackenbush, T. P. Barnwell and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, 1988.

[43] I. Potamitis, N. Fakotakis, N. Liolios, and G. Kokkinakis. "Speech Enhancement Using Mixtures of Gaussians for Speech and Noise," *International conference on text, speech and dialogue*, vol. 2448, pp. 337-340, 2002.

[44] T. Agarwal, *Pre-Processing of Noisy Speech for Voice Coders,* Master thesis, McGill University, 2002.

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

[46] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE, Speech, and Signal Processing*, vol. 67, no. 12, pp. 1586-1604, December 1979.

[47] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, March 1999.

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

[49] M. Kato, M. Serizawa, N Toki, U. Mori, Y Morishita, and K. Hayashi, "Noise suppression with high speech quality based on weighted noise estimation for 3G handsets," *NEC Research and Development*, vol. 44, no. 4, October 2003.

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

[51] O. Cappé, "Elimination of the musical noise phenomenon with Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345-349, April 1994.

[52] Y. Ephraim, D. Malah, and B. Juang, "On the application of hidden markov models for enhancing noisy speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1846-1856, December 1989.

[53] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526-1555, December 1992.

[54] Y. Ephraim, H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, July 1995.

[55] U. Mittal and N. Phamdo, "Signal/Noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 159-167, March 2000.

[56] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, vol. 82, pp. 35-45, 1960.

[57] G. Welch and G. Bishop. "An Introduction to the Kalman Filter," *SIGGRAPH 2001*, ACM, 2001.

[58] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 12, pp. 297-300, April 1987.

[59] D. C. Popescu and I. Zeljkovic, "Kalman filtering of colored noise for speech enhancement," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 997-1000, May 1998.

[60] M. Gabrea, "Adaptive Kalman filtering-based speech enhancement algorithm," *Canadian Conference on Electrical and Computer Engineering 2001*, vol. 1, pp. 521-526, May 2001.

[61] S. M. Bozic, *Digital and Kalman Filtering - An introduction to discrete-time filtering and optimum linear estimation*, Edward Arnold, 1979.

[62] C. E. Davila, "On the noise-compensated Yule-Walker equations," *IEEE Transactions on Signal Processing*, vol. 49, no. 6, pp 1119-1121, June 2001.

[63] C. E. Davila, "A subspace approach to estimation of autoregressive parameters from noisy measurements," *IEEE Transactions on Signal Processing*, vol. 46, no. 2, pp 531-534, February 1998.

[64] W. X. Zheng, "Autoregressive parameter estimation from noisy data," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp 71-75, January 2000.

[65] J. Timoney, T. Lysaght, and M. Schoenwiesner, "Implemeting loudness models in Matlab," *Procedures of the 7th International Conference on Digital Audio Effects*, October 2004.