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

Ning Ma

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Electrical Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Speech Enhancement Algorithms Using Kalman Filtering and Masking Properties of Human Auditory Systems

TITRE DE LA THÈSE / TITLE OF THESIS

Martin Bouchard

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

Rafik Goubran

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

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Richard Dansereau

Éric Dubois

Christian Giguère

Roch Lefebvre

Gary W. Slater

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

Speech Enhancement Algorithms Using Kalman Filtering and Masking Properties of Human Auditory Systems

Ning Ma

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in fulfillment of the requirements for the degree of

Doctor of Philosophy

in Electrical Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
School of Information Technology and Engineering

Faculty of Engineering
University of Ottawa

August 2005

Your file Votre référence
ISBN: 0-494-10988-2
Our file Notre référence
ISBN: 0-494-10988-2

ABSTRACT

Speech enhancement algorithms have been employed successfully in many areas such as VoIP, automatic speech recognition and speaker verification. Many approaches are presented in the literature. This thesis focuses on enhancing single channel speech degraded by white noise or colored noise. A Kalman filter algorithm combined with the masking properties of human auditory systems is proposed. The threshold computed from the masking properties is used as a constraint in the Kalman filter to theoretically derive a modified Kalman filter. The derivation gives a theoretical foundation for the feasibility of combining masking properties with a Kalman filter. Some heuristic methods are also proposed for an easier implementation. One algorithm proposes to use the frequency domain masking level as a hard threshold to reshape the Kalman filtered signal. Another algorithm is to use a post-filter concatenated with the Kalman filter, using a threshold where both time-domain and frequency domain masking properties are taken into account. The goal of the masking is to make the energy of the estimate state error smaller than the threshold. To further decrease the computational cost, a wavelet Kalman filter combined with masking thresholds is also introduced. In the above algorithms, the speech model is assumed to be linear. Nonlinear speech models are also considered in the thesis. To address the nonlinear model problem, dual Extended Kalman Filter (EKF) and dual Unscented Kalman Filter (UKF) algorithms are studied. In these cases, both time-domain and frequency domain masking properties are taken into account. The simulation results show that all the proposed methods combining Kalman filter and masking properties can produce promising results from the point of view of PESQ scores. The average PESQ score gains obtained by these proposed methods are from about 0.35 to 0.45. Some informal subjective tests also show that the performance of the proposed methods is promising. No voice activity detection is required in the proposed methods.

ACKNOWLEDGMENTS

I would like to thank my principal supervisor, Professor Martin Bouchard, for having always believed in my work and for providing insightful and prompt feedback during all the stages of the research described in this thesis. His vision was fundamental in shaping my research and I am very grateful for having had the opportunity to learn from him for his guidance. He has been the person that I could always go to with ideas and have informative technical discussions.

I would also like to express my appreciation to my co-supervisor Professor Rafik A. Goubran for his advice and encouragement. His suggestions helped me in forming the thesis structure and the content. As well, thanks to my colleagues Cathy Shao, Qiongfeng Pan, Jianping Deng and other students in the Barlo lab for their help in the testing of results and for sharing their knowledge.

Finally, I am grateful for the support of my husband, Hongqing Zeng and my parents during all my Ph. D study.

CONTENTS

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Thesis Motivation

With the development of communication techniques and the increased requirements on speech recognition and speaker verification systems, speech enhancement technology has been a hot research area for the past two decades. More and more hands-free systems and mobile communication devices used with more hostile noise conditions demand an efficient speech enhancement algorithm. In the real world, speech signals are often distorted by background noise such as car engine, traffic, wind or people chattering. Such degradation can lower the intelligibility or quality of speech. Thus the performance of the speech processing related system is deteriorated. For example, if a mobile telephony system encodes a noisy speech signal without any preprocessing, then the system performance is further degraded since most speech coding algorithms rely on clean signals and are not suitable for noisy signals.

Speech enhancement can be applied to these systems to increase the intelligibility and/or quality of speech. 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 [O'Sh89]. The objectives of speech enhancement algorithms vary with the application. In an automated speech recognition system, the major goal of speech enhancement is to increase the intelligibility of speech, while in communication systems, the target depends on the signal-to-noise ratio (SNR) of the distorted speech. In medium-to-high SNR (e.g. > 5dB) environments, the goal is to reduce the noise level

to produce a natural speech signal. For low SNRs, the objective could be to decrease the noise level while retaining or increasing the intelligibility and reducing the auditory fatigue caused by heavy noise. Hearing aid systems can also benefit from the speech enhancement technologies by reducing the environmental noise.

The number of microphones varies with different applications. In some applications such as a cell phone system, a single microphone is used. In other applications such as a video conference system, a microphone array is often used. The thesis addresses the problem of a single microphone system. The algorithms for speech enhancement in a single microphone system can be classified into two categories. The first category is frequency domain methods such as spectral subtraction algorithms [Bol79, Eph84, Vir99]. In frequency-domain methods, the spectral subtraction approach is commonly used as a standard reference. It estimates the Power Spectral Density (PSD) of a clean speech signal by subtracting the PSD of the noise from the PSD of the noisy speech signal [Bol79]. Each estimate of the PSD is performed within a short-time segment. The advantage of the spectral subtraction based approaches is that they are straightforward and easy to implement. However, the limitations are:

- Musical noise is unavoidable in the approach. The noise is mostly perceived in weak bands and weak speech frames.
- The performance of the methods depends on a good noise and SNR estimation.

The other category of speech enhancement algorithms is time domain methods such as Kalman filtering based methods [Pal87, Gib91, Gan98, Gab01, Pop98]. The Kalman filtering algorithm was first proposed to be applied to speech enhancement by Paliwal and Basu [Pal87], with the speech parameters obtained from the clean speech signal and the noise characteristics obtained from non-speech frames. In [Pop98], a Kalman filter based approach was proposed for colored noise cases. These methods have to detect non-speech frames for the noise covariance estimation. The Kalman filter can provide an Minimum Mean-Squared Error (MMSE) estimate of the clean

signal if the noise is a Gaussian process or it can give an Linear Minimum Mean-Squared Error (LMMSE) estimate if the noise is non-Gaussian. There is no musical noise produced in the results. With the approach in [Gab01], the estimation of the environmental noise process and the speech model noise process can be performed within the estimation steps by the calculated Kalman filtering parameters, with no need to detect the speech/non-speech frames in the procedure. Though the computation cost is higher than the spectral subtraction method, the performance of the Kalman filter is better in the point of view of the PESQ scores (Perceptual Evaluation of Speech Quality [Rix01, ITUT01], an ITU-T recommendation ITU-T P.862). PESQ is an objective speech quality assessment model. As an enhanced version of the perceptual speech quality measure (PSQM), it was developed for use across a wider range of network conditions, including analogue connections, codecs, packet loss and variable delay. Background noise and processing noise can be assessed by presenting PESQ with the clean, unprocessed original and the degraded signal. This thesis uses the PESQ score as the performance index. Because the conventional spectral subtraction has musical noise in the results, the PESQ scores are typically lower than that of the Kalman filter. Some methods have been proposed to suppress the musical noise such as in [Vir99], but the results are not fully satisfactory, especially under very low input noisy speech SNR (<5dB). The Kalman filter has a better performance in such cases. However, there is still room for significant improvements.

1.2 Goal of the Thesis

When a speech enhancement algorithm is designed, a tradeoff between speech distortion and the residual noise has to be achieved. When conventional spectral subtraction methods try to remove noise too aggressively, then the amount of musical noise produced by the method will increase. In [Vir99], the masking properties of human auditory systems were applied to a spectral subtractive

method to make the residual noise "perceptually white" and to keep the speech distortion small. Human auditory masking is a highly complicated process which is only partially understood. In order to be audible, sounds must achieve a minimum pressure. This threshold of hearing (audibility) is unique from person to person and furthermore changes with a person's age [Zwi99]. The threshold is called Absolute Threshold of Hearing (ATH). Masking effects can be classified as time-domain (temporal) masking and frequency-domain (simultaneous) masking. In frequency domain masking, the masking sound and the masked sound are present at the same time. Whereas temporal masking refers to the effect of masking with a small time offset. Temporal masking effects can be further classified as forward masking and backward masking. By forward masking, a sound is inaudible for a short time until the masker has been diminished. The time offset in forward masking can be from 5 ms to more than 150 ms. Whereas in backward masking, a weak signal is inaudible before the beginning of the masking signal. The time offset is usually below 5 ms [Thi00]. The frequency selectivity of the masking effects is described in terms of *Critical Bands* (CB). In general, a CB is the bandwidth around a center frequency which marks a (sudden) change in subjective response [Har97]. An absolute frequency scale based on the original CB measurement (as used by Zwicker in [Zwi99, Zwi80]) is in common use. This scale is called the Bark scale. Sound under the masking threshold is inaudible to people.

Building on top of the previous work in [Vir99] for the use of masking properties in speech enhancement, the applicability of using human auditory masking in Kalman filtering for speech enhancement is considered in this thesis. A standard Kalman filter can produce an enhanced speech signal which has a small distortion, however the residual noise is larger than the masking threshold. This thesis proposes a masking threshold constrained Kalman filter to achieve the goal of making the residual noise smaller than the masking threshold, so that it cannot be heard while keeping the speech distortion small. In the Kalman filtering algoritm, the estimate error covariance matrix is updated at each step. The objective of a standard Kalman algorithm is to

minimize the estimate error variance. When the masking properties of a human auditory system are taken into account, the problem becomes to minimize the estimate error variance under the constraint that the energy of the estimate error should be smaller than the masking threshold. Thus a new family of Kalman filtering algorithms is derived for speech enhancement, including truly constrained algorithms and heuristic approximates for real-time implementation.

1.3 Contributions of the Thesis

The contributions of the thesis are:

1. Proposed a truly perceptually constrained Kalman filtering algorithm for both white and colored noise. This contribution was presented in the following publications:
   (1) N. Ma, M. Bouchard and R. Goubran, "Speech Enhancement Using a Perceptually Masking Threshold Constrained Kalman Filter and its Heuristic Implementations", accepted for *IEEE Trans. on Speech and Audio Processing*.
   (2) 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
2. Proposed a heuristic method combining the adaptive Kalman filtering algorithm with masking properties of human auditory systems for white noise. This contribution was presented in the following publication:
   (1) N. Ma, M. Bouchard and R. Goubran, " A perceptual Kalman filtering-based approach for speech enhancement", *Proceedings of IEEE-EURASIP Seventh International Symposium on Signal Processing and its Applications (ISSPA) 2003*, vol. 1, pp. 373-376, Paris, France, July 2003.

3. Proposed a heuristic method combining the adaptive Kalman filtering algorithm with masking properties of human auditory systems for colored noise. This contribution was presented in the following publications:
   (1) 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, n. 3, pp. 138-139, Ottawa, Ontario, Oct. 2004
   (2) 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.
4. Proposed a method combining the wavelet Kalman filter with masking properties to reduce computational cost. This contribution was presented in the following publication:
   (1) N. Ma, M. Bouchard and R. Goubran, "A Wavelet Kalman Filter with Perceptual Masking for Speech Enhancement in Colored Noise", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2005*, vol. 1, pp. 149-152, Philadelphia, USA, Mar. 2005.
5. Proposed a perceptually constrained dual unscented Kalman filter, with a non-linear speech model. This contribution was presented in the following publications:
   (1) N. Ma, M. Bouchard and R. Goubran, "Constrained Dual Unscented Kalman Filter for Enhancing Speech Degraded by Colored Noise", submitted to *Speech Communication*, under review.
   (2) N. Ma, M. Bouchard and R. Goubran, "Dual Perceptually Constrained Unscented Kalman Filter for Enhancing Speech Degraded by Colored Noise", *Proceedings of IEEE Seventh International Conference on Signal Processing (ICSP'04)*, vol. 3, pp. 2522-2525,

Beijing, China, Aug. 31-Sept. 4 2004

6. Compared the complexity of all the above methods and tested the methods under different types of noise and for different applications.

1.4 Organization of the Thesis

In Chapter 2, a literature review and some background introduction of speech enhancement is presented. A theoretical derivation of frequency and time domain masking threshold constrained Kalman filters is discussed in Chapter 3. In Chapter 4, Kalman filtering algorithms heuristically combined with frequency or both frequency and time domain masking properties are proposed for enhancing speech degraded by white noise and colored noise. In Chapter 5, a wavelet domain Kalman filter and perceptual masking post-filter is proposed. In Chapter 6, a constrained dual unscented Kalman filter and its heuristic implementation, i.e. a dual unscented Kalman filter with a perceptual masking post-filter, are proposed. Chapter 7 will present some subjective test results to validate the simulation results of the previous chapters, and Chapter 8 will present conclusions and some future work.

Chapter 2 Literature Review and Background

2.1 Introduction

In the past four decades, a lot of research has been performed in the field of speech enhancement, e.g. [Eph84, Lim79, O'Sh89]. With the development of communications, the study of speech enhancement methods has been very intensive. The main purpose of speech enhancement is to improve perceptual aspects (e.g., quality and intelligibility) of a given degraded speech signal.

In most speech enhancement algorithms, it is assumed that the speech is degraded by additive noise which does not depend on the clean speech. Some other practical noises can be transformed into additive noise. For example, a multiplicative or convolutional noise degradation will be converted to an additive noise degradation by a homomorphic transformation [Opp89].

The quality of speech signals is a subjective measure which reflects the way the signal is perceived by listeners. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted by listeners from the given signal, whether the signal is clean or noisy. Until now, most methods can only improve one aspect at the cost of not improving the other one. In this review, some speech enhancement approaches studied in recent years will be introduced.

Some papers have proposed approaches to improve the intelligibility [Nie92, Wit94]. The main

philosophy is that the features relevant to speech intelligibility can be extracted from noise-corrupted speech, and used to process the corrupted speech to increase its intelligibility.

But most of speech enhancement research has focused on removing the corrupting noise, to improve the overall quality of the speech signal. The approaches can be classified into two major categories: frequency-domain methods and time-domain methods. In frequency-domain methods, the spectral subtraction approach is used almost as a standard. It estimates the Power Spectral Density (PSD) of the clean signal by subtracting the PSD of the noise from the PSD of the noisy signal [Bol79]. Each estimate of PSD is performed within a short-time segment.

The thesis focuses on the problem of single channel speech enhancement. Methods introduced in section 2.2.1 and section 2.2.2 are derived and modified from the basic spectral subtraction approach. Sections 2.2.3 to 2.2.5 describe the time-domain methods.

2.2 Speech Enhancement Techniques

2.2.1 Wiener Filtering Based Methods

Wiener filters are a class of linear optimum discrete-time filters [Hay96]. Given a number of observations, $y(n)$, it is desired to find a linear optimum estimate $\hat{s}(n)$ of signal $s(n)$. The observations are assumed to be the sum of the desired signal $s(n)$ plus noise $d(n)$ as follows:

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

The estimate $\hat{s}(n)$ is obtained by applying a linear filter to the observations:

$$\hat{s}(n) = \sum_{k=0}^{M-1} w_k y(n-k) \tag{2.2}$$

where the filter is assumed to be a FIR filter and $w_k\,(k=0,\cdots,M-1)$ are the filter coefficients. According to Wiener filter theory, the filter is to minimize the mean squared error:

$$E\{e(n)^2\} = E\{[\hat{s}(n)-s(n)]^2\} \tag{2.3}$$

To minimize $E\{e(n)^2\}$, the filter makes the error orthogonal to all observations:

$$E\{y(n-k)e^*(n)\}=0,\quad k=0,\cdots,M-1 \tag{2.4}$$

where the notation * stands for the conjugate operator. It can be shown [Hay96] that if $y(n)$ and $s(n)$ are zero-mean jointly wide-sense stationary processes, the filter coefficients must satisfy the following equations:

$$\sum_{i=0}^{M-1} w_i r(i-k) = p(-k) \qquad k=0,1,\cdots,M-1 \tag{2.5}$$

where $r(i-k)=E\{y(n-k)y^*(n-i)\}$, $p(-k)=E\{y(n-k)s^*(n)\}$.

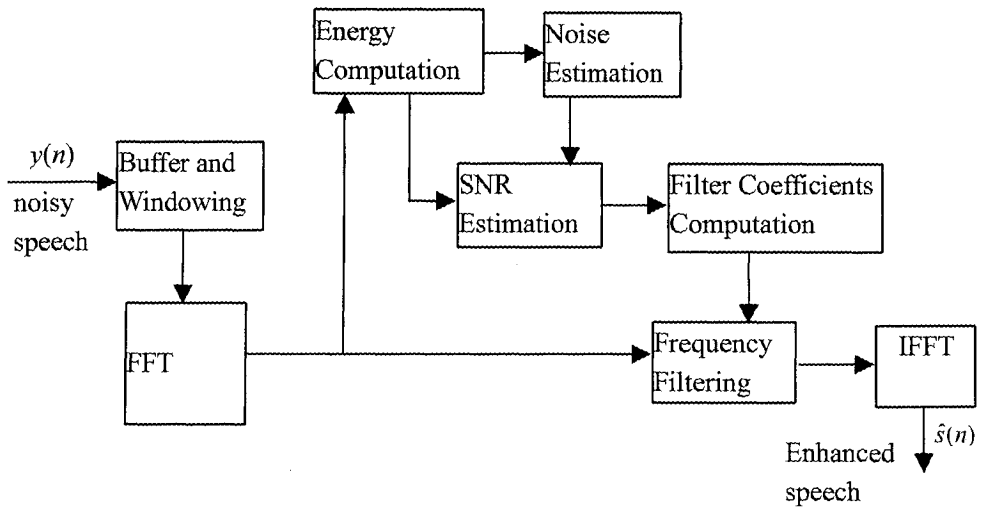


Fig. 2.1 Generalized Wiener filter for speech enhancement

By solving equation (2.5), the Wiener filter coefficients can be obtained. In the frequency domain,

the frequency response of the filter can be represented by:

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} = \frac{SNR(\omega)}{SNR(\omega)+1} \tag{2.6}$$

In the field of speech enhancement, $s(n)$ is the clean speech signal and $y(n)$ is the noisy speech signal. Hence, if the power spectrum of the signal and noise could be obtained or estimated, a speech enhancement system may be constructed. Fig 2.1 is a diagram of a general Wiener-based enhancement system [Nem99].

Power Subtraction Methods

From a Wiener filter, a power subtraction filter can be obtained. The response of the power subtraction filter is the square root of a Wiener filter. Hence,

$$|H(\omega)|_p = \sqrt{\frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)}} \tag{2.7}$$

The filter can suppress noise in the power spectrum sense.

Magnitude Subtraction Methods

Another kind of filter based on Wiener filter theory is magnitude subtraction filters. The response of such a kind of filters is as follows:

$$|H(\omega)|_M = 1 - \sqrt{\frac{P_d(\omega)}{P_s(\omega) + P_d(\omega)}} \tag{2.8}$$

The magnitude spectrum of the output of this filter is:

$$M_{out} = \sqrt{P_s(\omega) + P_d(\omega)} - |D(\omega)| \tag{2.9}$$

Thus the filter can suppress the noise in the magnitude spectrum sense.

Implementation of the Methods

In the implementation, power subtraction method can be represented as the following equations:

let $G(\omega) = P_y(\omega) - P_d(\omega)$ (2.10)

$$P_{\hat{s}}(\omega) = \begin{cases} G(\omega), & \text{if } G(\omega) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.11)$$

where $P_{\hat{s}}(\omega)$ is power spectrum of the enhanced speech signal, $P_y(\omega)$ is the spectrum of the noisy speech signal $y(n)$, and $P_d(\omega)$ is the smoothed estimate of the noise spectrum. The enhanced speech signal is obtained by an inverse Fourier transform from both $P_{\hat{s}}(\omega)$ and the phase of the noisy signal:

$$\hat{s}(n) = \mathbf{F}^{-1}(\sqrt{P_{\hat{s}}(\omega)} \cdot e^{j\theta_y(\omega)}) \quad (2.12)$$

where $\theta_y(\omega)$ is the phase spectrum of the noisy speech. The major problem of the above method is that "musical noise" is produced in the enhanced speech signal.

To reduce the effect of "musical noise", in [Ber79], a generalized spectral subtraction method is proposed:

let $G(\omega) = \tilde{G}[P_y^{\gamma}(\omega) - \alpha P_d^{\gamma}(\omega)]$ (2.13)

$$P_{\hat{s}}(\omega) = \begin{cases} G^{1/\gamma}(\omega), & \text{if } G^{1/\gamma}(\omega) > \beta P_d(\omega) \\ \beta P_d(\omega) & \text{otherwise} \end{cases} \quad (2.14)$$

with $\alpha \geq 1$, $0 < \beta << 1$

where α is the noise over-subtraction factor employed to compensate the imperfect noise estimation, which leads to a reduction of residual noise peaks but also to an increased audible distortion. β is the spectral floor factor, which leads to a reduction of residual noise but an increased level of background noise remains in the enhanced speech. The exponent γ determines the sharpness of the filter, and $\tilde{G}$ is the normalization factor. The choice for the value of γ is not as critical as that of α and β. If $\gamma = 1$, it is a power subtraction filter. If $\gamma = 0.5$, it is a magnitude subtraction method.

The advantage of the Wiener-based approach is that it is straightforward and easy to implement. However, the limitations are:

- Musical noise is not avoidable in the approach. The noise is mostly perceived in weak bands and weak speech frames. This is an unnatural sounding residual noise. It can be easily explained by the statistical nature of the noise. For example, consider some frequency of the instantaneous spectrum which does not contain any speech. On the one hand, the effect of too small a noise estimate is a remaining excitation at this frequency. On the other hand, if the noise is estimated to be higher than it actually is, the result will be zero due to the necessary bounding. The result is short sinusoids randomly distributed over time and frequency, which remain in the processed signal. Normally this kind of noise is called musical noise.
- The performance of the methods depends on a good noise and SNR estimation.

2.2.2 MMSE Spectral Estimation Based Methods

In 1984, Ephraim and Malah proposed a speech enhancement method based on minimum mean-squares error (MMSE) spectral estimation [Eph84]. An MMSE short-time spectral amplitude (STSA) estimator is derived by assuming *a priori* probability distribution of the speech and noise Fourier expansion coefficients. In the paper, the Fourier expansion coefficients of each process are modeled as statistically independent Gaussian random variables with zero mean.

If the noisy speech signal is represented by equation (2.1), the kth spectral component of the signals $s(n)$ and $y(n)$ can be expressed as follows:

$$S_k = A_k e^{j\alpha_k}, \tag{2.15}$$

$$Y_k = R_k e^{j\vartheta_k}, \quad k = 0, \pm 1, \pm 2, \cdots \tag{2.16}$$

D_k is the kth spectral component of the noise $d(n)$. Then the MMSE amplitude estimate $\hat{A}_k$ is:

$$\hat{A}_k = E\{A_k | Y_0, Y_1, \cdots\} = E\{A_k | Y_k\} \tag{2.17}$$

where $\{Y_0, Y_1, \cdots\}$ is the set of spectral observations over a few frames. The estimator can be expressed as a spectral gain with the following equation:

$$\begin{aligned} G_{MMSE}(\xi_k, \gamma_k) &\equiv \frac{\hat{A}_k}{R_k} \\ &= \frac{\sqrt{\pi}}{2} \frac{\sqrt{v_k}}{\gamma_k} \exp\left(-\frac{v_k}{2}\right)\left[(1+v_k) I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k \end{aligned} \tag{2.18}$$

with: $I_0(\cdot)$ and $I_1(\cdot)$ are modified Bessel functions of zero and first order respectively.

$$v_k \equiv \frac{\xi_k}{1+\xi_k} \gamma_k \tag{2.19}$$

where ξ_k and γ_k are defined as *a priori* and *a posteriori* signal-to-noise ratios (SNR), respectively.

If the uncertainty of speech presence in the noisy observations is taken into account, a modified gain function is obtained in [Eph84]:

$$G^D_{MMSE}(\eta_k, \gamma_k, q_k) \equiv \frac{\Lambda(\xi_k, \gamma_k, q_k)}{1+\Lambda(\xi_k, \gamma_k, q_k)} G_{MMSE}(\xi_k, \gamma_k)\Big|_{\xi_k = \eta_k/(1-q_k)} \tag{2.20}$$

with:

$\Lambda(Y_k, q_k) = \mu_k \frac{\exp(v_k)}{1+\xi_k}$, which is derived from the following generalized likelihood ratio definition:

$$\Lambda(Y_k, q_k) = \mu_k \frac{p(Y_k | H_k^1)}{p(Y_k | H_k^0)} \tag{2.21}$$

with $\mu_k \equiv (1-q_k)/q_k$, and q_k is the probability of signal absence in the kth spectral component. H_k^0 and H_k^1 denote the two hypotheses of signal absence and presence, respectively, in the kth spectral component. Assume that the spectral components are Gaussian random variables,

equation (2.20) can be obtained. In (2.20), v_k is defined as in (2.19), ξ_k is redefined by

$$\xi_k \equiv \frac{E\{A_k^2 | H_k^1\}}{\lambda_d(k)} \tag{2.22}$$

with $\lambda_d(k) = E\{|D_k|^2\}$.

The performance of the MMSE based method depends on the good estimate of *a priori* and *a posteriori* SNR which is still an open issue in the current research.

2.2.3 KLT Based Method

In [Eph95], the author proposed a signal subspace approach for speech enhancement. The underlying principle is to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace. The decomposition is performed by Karhunen-Loeve transform (KLT). Enhancement is performed by linear estimation of the clean signal from the noisy signal. The noise is assumed to be additive and white. The linear estimation is obtained by modifying the KLT components which represent the signal subspace by a gain function determined by the estimation criterion. The remaining KLT components are nulled. The enhanced signal is obtained from inverse KLT of the altered components. The estimation criterion is to keep the level of the residual noise below some threshold while minimizing the signal distortion. Two estimators are studied in [Eph95]. One is to maintain the energy of the residual noise in the entire frame below a given threshold. Another one is to keep each spectral component of the residual noise below a given threshold. The first one allows time domain constraint (TDC) on the residual noise, while the second one is designed for noise shaping using spectral domain constraints (SDC). In implementing the linear signal estimators, one must have good estimates of the covariance matrices of the vectors of the noisy signal and the noise process. Furthermore, a good estimate of

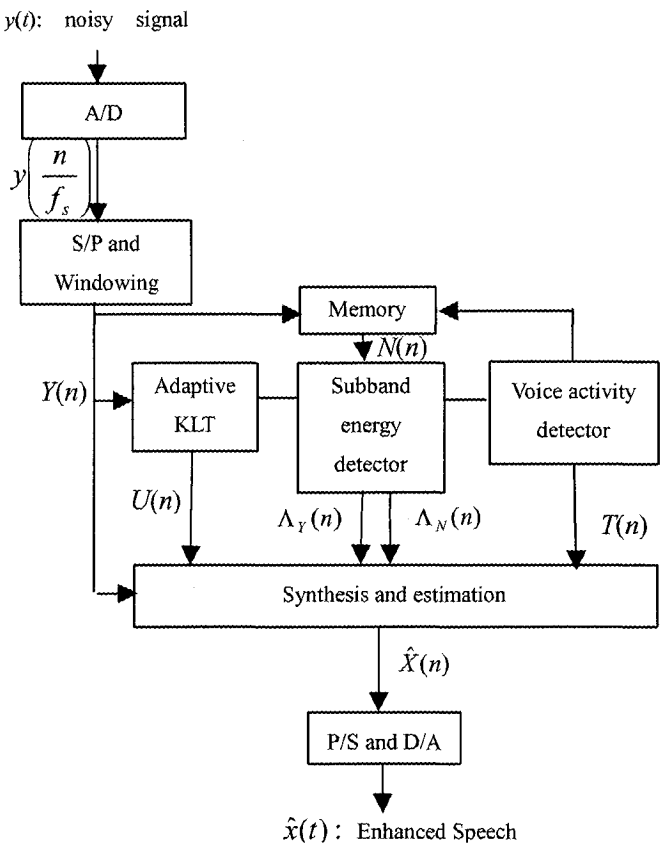


Fig. 2.2 Block diagram of the system in [Rez01]

the dimension of the signal subspace is required. The covariance matrices of the vectors of the noise process are estimated from vectors of the noisy signal during which speech is absent. If the noise is stationary, the estimation can be performed from an initial segment of the noisy signal which was recorded before speech was present. When the noise is not stationary, a speech/noise detector must be used, and the noise covariance is estimated and updated from non-speech frames of the noisy signal.

[Mit00] proposed a KLT based approach for enhancing speech degraded by colored noise. The additive noise is assumed stationary. Speech frames are classified into speech activity frames and

silent frames. The speech activity frames can further be divided into voiced speech and unvoiced speech frames. The energy of speech in the voiced speech frames is typically higher than in the unvoiced speech frames and in the silent frames. The noisy speech frames are classified into two categories, namely speech dominated frames and noise dominated frames, in this paper. Noise dominated frames are either non-speech activity frames or unvoiced speech frames. The classification is done through an empirical equation. The signal KLT matrix is used for speech dominated frames and a noise KLT matrix is used for noise dominated frames. The algorithm depends on the *a priori* knowledge of SNR. This SNR can be evaluated by long term averaging. [Rez01] proposed an adaptive KLT approach for the case of colored noise. The new type of KLT tracking algorithm was named projection approximation subspace tracking. In this method, Eigenvalue Decomposition (ED) is considered as a constrained optimization problem and an adaptive algorithm is used to find a very close approximation of the eigenvectors of the clean speech covariance matrix using noisy speech signal samples. This algorithm is fast and has a simple construction. It is assumed that the statistical characteristics of noise do not vary too much until the next noise interval arrives and noise samples are renewed. An efficient voice activity detector must be used. Fig.2.2 shows the block diagram of the proposed system.

Since the KLT and the DFT are related, the popular spectral subtraction method can be seen as an approximate signal subspace approach.

2.2.4 Kalman Filter Based Methods

The approaches based on Kalman filter are time-domain methods. When the measurement noise process and the signal are jointly Gaussian, the Kalman filter can yield an MMSE estimate of the clean signal. Otherwise, the Kalman filter can produce an LMMSE estimate of the signal. The Kalman filtering algorithm was first proposed to apply to speech enhancement by Paliwal and

Basu [Pal87]. A summary of the standard Kalman filtering algorithm is shown below [Chu99].

A linear system with state-space description is written as the following state-space model:

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{B}_k \xi_k + \mathbf{G}_k \mathbf{u}_k \\ \mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{D}_k \xi_k + \mathbf{v}_k \end{cases} \tag{2.23}$$

where $\mathbf{F}_k, \mathbf{B}_k, \mathbf{G}_k, \mathbf{H}_k, \mathbf{D}_k$ are $n \times n$, $n \times m$, $n \times p$, $q \times n$, $q \times m$ (known) constant matrices, respectively, with $1 \le m, p, q \le n$, $\{\xi_k\}$ is a known sequence of m-vectors (called a deterministic input sequence), and $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are, respectively, (unknown) system and observation noise sequences, with known statistical information such as mean, variance, and covariance.

When a Kalman filter is applied to speech enhancement, the deterministic input is zero. Thus the state-space model for speech enhancement is a linear (purely) stochastic system:

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{G}_k \mathbf{u}_k \\ \mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \end{cases} \tag{2.24}$$

The estimate of the state vector depends on the statistical information of the noise sequences.

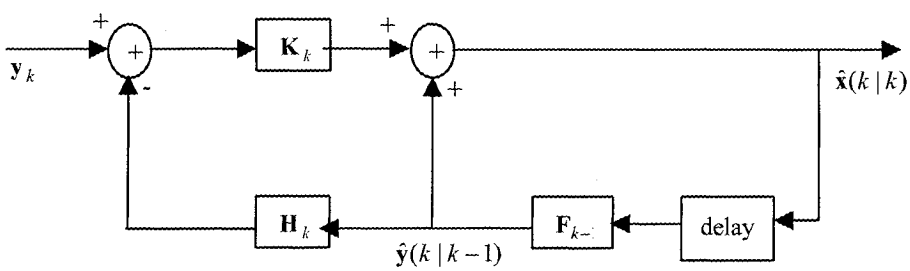


Fig. 2.3 Diagram of standard Kalman filter

The derivation of the algorithm can be found in [Chu99]. Assume that $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ are sequences of zero-mean Gaussian white noise such that $Var(\mathbf{u}_k) = \mathbf{Q}_k$ and $Var(\mathbf{v}_k) = \mathbf{R}_k$ are

positive definite matrices and $E(\mathbf{u}_k \mathbf{v}_l^T) = \mathbf{0}$ for all k and l. The initial state $\mathbf{x}_{-1}$ is also assumed to be independent of $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ in the sense that $E(\mathbf{x}_{-1}\mathbf{u}_k^T) = \mathbf{0}$ and $E(\mathbf{x}_{-1}\mathbf{v}_k^T) = \mathbf{0}$ for all k. The following Kalman filtering algorithm for the linear stochastic system (2.24) is given by:

$$\begin{cases} \mathbf{P}(-1|-1) = Var(\mathbf{x}_{-1}) \\ \mathbf{P}(k|k-1) = \mathbf{F}_k \mathbf{P}(k|k-1)\mathbf{F}_{k-1}^T + \mathbf{G}_{k-1}\mathbf{Q}_{k-1}\mathbf{G}_{k-1}^T \\ \mathbf{K}_k = \mathbf{P}(k|k-1)\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}(k|k-1)\mathbf{H}_k^T + \mathbf{R}_k)^{-1} \\ \mathbf{P}(k|k-1) = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}(k|k-1) \\ \hat{\mathbf{x}}(-1|-1) = E(\mathbf{x}_{-1}) \\ \hat{\mathbf{x}}(k|k-1) = \mathbf{F}_k\hat{\mathbf{x}}(k-1|k-1) \\ \hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathbf{K}_k[\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}(k|k-1)] \\ k = 0, 1, \cdots \end{cases} \tag{2.25}$$

where $\mathbf{P}(k|k-1) = E\{[\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)][\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)]^T\} = Var[\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)]$ is the predicted state error vector covariance matrix, $\mathbf{P}(k|k) = E\{[\mathbf{x}_k - \hat{\mathbf{x}}(k|k)][\mathbf{x}_k - \hat{\mathbf{x}}(k|k)]^T\} = Var[\mathbf{x}_k - \hat{\mathbf{x}}(k|k)]$ is the estimate state error vector covariance matrix. $\mathbf{K}_k$ is called the Kalman gain matrix. The algorithm is shown in Fig. 2.3.

In speech enhancement, the clean speech signal $s(n)$ is modeled as a p-th order Auto-Regressive (AR) process as follows:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{2.26}$$

The measured speech signal $y(n)$ is

$$y(n) = s(n) + v(n) \tag{2.27}$$

where a_i is the i-th AR coefficient, $u(n)$ is the model noise process and $v(n)$ is the background noise. $u(n)$ and $v(n)$ are uncorrelated Gaussian white noise processes with means $\bar{u}$ and $\bar{v}$ and variances σ_u^2 and σ_v^2.

The state space model, written as follows, can then be obtained.

$$\mathbf{x}(n) = \mathbf{F}\mathbf{x}(n-1) + \mathbf{G}u(n) \tag{2.28}$$

$$y(n) = \mathbf{H}\mathbf{x}(n) + v(n) \tag{2.29}$$

where

$$\mathbf{x}(n) = [s(n-p+1)\cdots s(n)]^T \tag{2.30}$$

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{2.31}$$

$$\mathbf{H} = \mathbf{G}^T = [0 \quad 0 \quad \cdots \quad 0 \quad 1] \tag{2.32}$$

where p is the order of the AR process, $\mathbf{F}$ is a $p \times p$ transition matrix, $\mathbf{G}$ and $\mathbf{H}$ are respectively the $p \times 1$ input vector and the $1 \times p$ observation row vector.

A Kalman filtering algorithm for speech enhancement is summarized in [Gab01] as follows:

$$e(n) = y(n) - \mathbf{H}\hat{\mathbf{x}}(n \mid n-1) - \bar{v} \tag{2.33}$$

$$\mathbf{K}(n) = \mathbf{P}(n \mid n-1)\mathbf{H}^T \times (\mathbf{H}\mathbf{P}(n \mid n-1)\mathbf{H}^T + \sigma_v^2)^{-1} \tag{2.34}$$

$$\hat{\mathbf{x}}(n \mid n) = \hat{\mathbf{x}}(n \mid n-1) + \mathbf{K}(n)e(n) \tag{2.35}$$

$$\mathbf{P}(n \mid n) = (\mathbf{I} - \mathbf{K}(n)\mathbf{H})\mathbf{P}(n \mid n-1) \tag{2.36}$$

$$\hat{\mathbf{x}}(n+1 \mid n) = \mathbf{F}\hat{\mathbf{x}}(n \mid n) + \mathbf{G}\bar{u} \tag{2.37}$$

$$\mathbf{P}(n+1 \mid n) = \mathbf{F}\mathbf{P}(n \mid n)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T\sigma_u^2 \tag{2.38}$$

In [Gab01], the sequential estimators are derived for sub-optimal adaptive estimation of the unknown *a priori* driving process and additive noise statistics simultaneously with the system state by reformulating and adapting the classical approach used for control applications. The algorithm provides improved state estimates at little computational expense. A distinct advantage of the proposed algorithm is that a speech/noise detector is not required.

In [Gan98], an iterative and sequential Kalman filter-based algorithm is proposed. The distinct advantage of the proposed algorithm compared to alternative algorithms is that it enhances the quality and SNR of the speech, while preserving its intelligibility and natural sound. This paper has proposed to incorporate higher-order statistics to improve the performance of the algorithm.

The noisy signal is expressed as:

$$z(t) = s(t) + v(t) \tag{2.39}$$

with:

$s(t)$: clean speech signal, which is modeled as a stochastic AR process:

$$s(t) = -\sum_{k=1}^{p} \alpha_k s(t-k) + \sqrt{g_s}\, u(t) \tag{2.40}$$

where $u(t)$ is a normalized zero mean unit variance white noise, and g_s is the spectral level.

The additive noise $v(t)$ is also assumed to be a realization from a zero-mean, possibly nonwhite stochastic AR process:

$$v(t) = -\sum_{k=1}^{q} \beta_k v(t-k) + \sqrt{g_v}\, w(t) \tag{2.41}$$

Then the estimate-maximum method is applied to the problem. The algorithm iterates between state and parameter estimation. The state estimation uses a forward Kalman filtering recursion, followed by a backward Kalman smoothing recurring.

For high SNR, we can use a standard LPC process to obtain the initial estimate of the speech parameters and an initial estimate of the noise parameters may be obtained by employing a voice activity detector. However, at low SNR, the initialization procedure breaks down. If the additive noise v(t) is assumed to be Gaussian, the higher-order statistics (HOS) may be incorporated in order to improve the initial estimate of the speech parameters.

The iterative EM algorithm requires the use of an analysis window over which the signal and noise statistics are assumed to be stationary. A sequential speech enhancement algorithm is proposed in this paper to avoid this assumption. This algorithm consists of a forward Kalman filter.

2.2.5 Hidden Markov Model Based Methods

In [Eph89], the author proposed to apply hidden Markov models (HMM) to enhance noisy speech. The approach is based upon statistical modeling of the clean speech and the noise process using long training sequences from the two processes. HMM's with mixtures of Gaussian autoregressive (AR) output probability distributions are used to model the clean speech signal. The model for the noise process depends on its nature. The enhancement of the noisy speech is done by means of re-estimation of the clean speech waveform using the EM algorithm. This method can increase SNR. For the case where the speech is degraded by statistically independent additive noise, the probability distribution functions of the clean speech signal and the noise process must be known. The algorithm starts from an initial HMM, and iteratively generates a sequence of HMM's with non-decreasing likelihood values by maximizing in each iteration the so-called auxiliary function.

In [Eph92], Gaussian subsources are used to represent clusters of spectrally similar acoustic signals. Hence, the power spectral densities of those subsources represent the different spectra of the speech signal or spectral prototypes of that signal. The Markovian states in the HMM provide a Markov model for the evolution of speech spectra or a model for the correlation between those spectra, since states are associated with signal spectra. The HMM can represent both the

intervector and the intravector correlation.

In the interpretation of speech production theory, each Gaussian subsource represents acoustic signals generated from a fixed configuration of the vocal tract. The vocal tract is considered an all-pole filter for the spectrally flat (Gaussian white noise for unvoiced signals and an impulse function for one pitch period of voiced signals) excitation signal. The set of all possible vocal tract shapes is represented by a finite number of configurations which equals the number of states in the HMM. This number should approximately be 1024, as observed in speech coding using vector quantization (VQ). The parameter set of the HMM for the clean signal (and similarly for the noise process) is normally estimated from training sequences using the ML estimation approach.

In [Sam98], an HMM-based strategy is proposed for enhancement of speech signals embedded in nonstationary noise. Compared to previous HMM-based speech enhancement approaches, the improvements to the systems are: 1) incorporation of mixture components in the HMM for noise in order to handle noise nonstationarity in a more flexible manner, 2) two efficient methods in the speech enhancement system design that make the system real-time implementable, and 3) an adaptation method to the noise type in order to accommodate a wide variety of noise expected under the enhancement system's operating environment.

The HMM's employed in this paper for clean speech and for noise are ergodic mixture autoregressive hidden Markov models (AR-HMM's). Three enhancement methods are studied in this paper. The first one is MAP enhancement method. The Expectation-Maximization (EM) algorithm is used in constructing the speech enhancement system. A new estimate of the clean speech is calculated by filtering the noisy speech through a weighted sum of the Wiener filters. This estimation is then used to find a probability sequence to supply a new sequence of Wiener

filters and another estimate of the clean speech. This iterative process continues until a preset convergence criterion is reached. In the first iteration, noisy speech is used as an estimate of the clean speech. The second one is approximate MAP method. This is a simplified approximation of the MAP method. Neither of these above two methods is capable of handling nonstationary noise due to the calculation of the filter weights being based on the clean signal information only and ignores variations. The third one is improved MMSE enhancement method. In this system, a multiple state and mixture noise model was employed to accommodate non-stationarity in noise. No iterations are necessary for the MMSE enhancement. For each frame, a large number of filter weights has to be calculated using a lot of computation. This makes the enhancement procedure very time consuming and far from being real-time implementable. This paper devised two methods to solve this problem.

In addition, the paper proposed a noise adaptation algorithm for diversified types of noise. The algorithm carries out noise-model selection and adaptation of the variances (LP gains) of the Gaussian AR processes associated with the noise HMM's. During intervals of non-speech activity, a Viterbi algorithm is performed on noise data using different noise models.

The HMM-based enhancement systems are inherently relying on the training data of different noise types. Only the noise types that have been trained can be handled by the system. Therefore data from various noise types should be used for training the noise HMM. This makes the noise HMM have a large model size and the computation cost grows drastically with the increasing number of the noise types.

2.2.6 The Application of Human Auditory Masking Properties to Speech Enhancement

2.2.6.1 Summary of Human Auditory Masking Properties

Human Auditory Masking is a highly complicated process which is only partially understood. In order to be audible, sounds must achieve a minimum pressure. This threshold of hearing (audibility) is unique from person to person and furthermore changes with a person's age [Zwi99]. The threshold is called Absolute Threshold of Hearing (ATH).

In [Ter82], the threshold is approximated by

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 (dB\ SPL) \qquad (2.42)$$

which is measured in dB SPL, or dB relative to 20 μ Pa [Har97]. The ATH is shown is Fig. 2.4. When applied to speech enhancement, $T_q(f)$ could be interpreted naively as a maximum

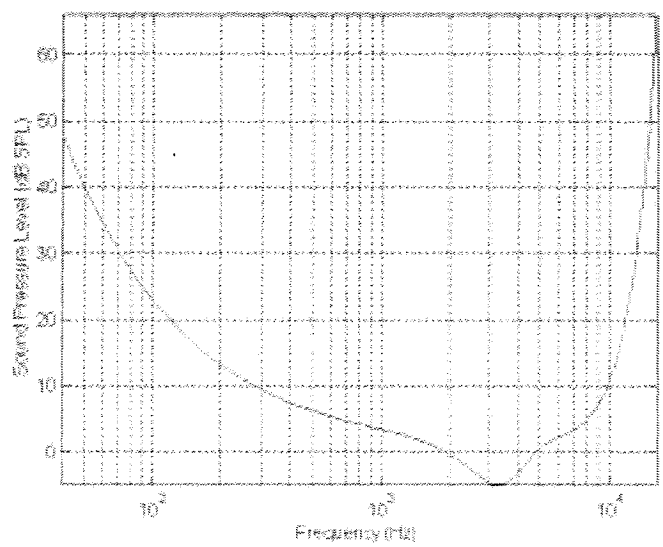


Fig. 2.4 The absolute threshold of hearing (ATH) in quiet.

allowable energy level for residual noise introduced in the frequency domain. The absolute threshold is of limited value in the speech enhancement context. The detection threshold for residual noise is determined by the stimuli present at any given time. Since stimuli are in general time-varying, the detection threshold is also a time-varying function of the input signal. In order to estimate this threshold, we must first understand how the ear performs spectral analysis. A frequency-to-place transformation takes place in the cochlea (inner ear), along the basilar membrane. Distinct regions in the cochlea, each with a set of neural receptors, are "tuned" to different frequency bands. In fact, the cochlea can be viewed as a bank of highly overlapping bandpass filters. The magnitude responses are asymmetric and non-linear (level-dependent). Moreover, the cochlear filter passbands are of non-uniform bandwidth, and the bandwidths increase with increasing frequency. The frequency selectivity of masking effects is described in terms of *Critical Bands* (CB). In general, a CB is the bandwidth around a center frequency which marks a (sudden) change in subjective response [Har97]. The actual width of Critical Bands is still in dispute. An absolute frequency scale based on the original (as used by Zwicker in [Zwi99, Zwi80]) CB measurement is in common use. This scale is called the Bark scale, and the common function to convert from Hz to Bark is (see [Zwi99, Zwi80]):

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right], \tag{2.43}$$

and the bandwidth (in Hz) of a CB at any frequency is given by

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69}. \tag{2.44}$$

The bandwidth in Bark of a CB at any frequency is (by definition) 1.

Alternatively, Moore in [Moo97] describes the CB as a measure of the 'effective bandwidth' of the auditory filters. Moore's measurement is called Effective Rectangular Band (ERB). It indicated narrower bandwidths than Zwicker's measurement. Table 2.1 and Fig. 2.5 show the mapping from FFT bins to Bark scale from Zwicker's model. In the thesis, the speech signals are

sampled at 8kHz, and we always use a FFT size of 256.

Table 2.1 Mapping from FFT Bins to Bark Scale (sampling rate: 8 kHz, FFT size: 256)

Critical band number k	FFT bins Intervals	Real frequency (Hz)
1	1-3	0-94
2	4-6	94-187
3	7-10	187-312
4	11-13	312-406
5	14-16	406-500
6	17-20	500-625
7	21-25	625-781
8	26-29	781-906
9	30-35	906-1094
10	36-41	1094-1281
11	42-47	1281-1469
12	48-55	1469-1719
13	56-64	1719-2000
14	65-74	2000-2312
15	75-86	2312-2687
16	87-100	2687-3125
17	101-118	3125-3687
18	119-128	3687-4000

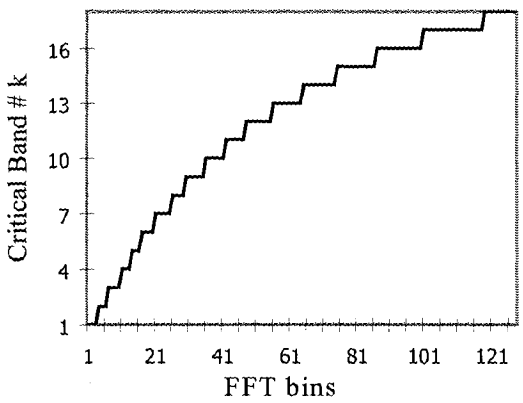


Fig. 2.5 Mapping from FFT bins to Bark scale (sampling rate: 8 kHz, FFT size: 256)

Masking effects can be classified as time-domain (temporal) masking and frequency-domain (simultaneous) masking. In frequency domain masking, the masking sound and the masked sound are present at the same time, whereas, temporal masking refers to the effect of masking with a small time offset. Temporal masking effects can be further classified as forward masking and backward masking, respectively. By forward masking, a sound is inaudible for a short time until the masker has been diminished. The time offset in forward masking can be from 5 ms to more than 150 ms, whereas in backward masking, a weak signal is inaudible before the beginning of the masking signal, and the time offset is usually below 5 ms [Thi00]. In the thesis, frequency-domain masking and time-domain forward masking are applied.

Simultaneous masking refers to a frequency-domain phenomenon that can be observed whenever two or more stimuli are simultaneously presented to the auditory system. Depending on the shape of the magnitude spectrum, the presence of certain spectral energy will mask the presence of other spectral energy. Although arbitrary audio spectra may contain complex simultaneous masking scenarios, for speech enhancement it is convenient to distinguish between only two types of simultaneous masking, namely *tone-masking-noise* and *noise-masking-tone*. In the first case, a

tone occurring at the center of a critical band masks noise of any subcritical bandwidth or shape, provided the noise spectrum is below a predictable threshold directly related to the strength of the masking tone. The second masking type follows the same pattern with the roles of masker and maskee reversed. A simplified explanation of the mechanism underlying both masking phenomena is as follows. The presence of a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane at the critical band location to block effectively detection of a weaker signal. Inter-band masking has also been observed, i.e., a masker centered within one critical band has some predictable effect on detection thresholds in other critical bands. This effect, also known as the spread of masking, is often modeled by an approximately triangular spreading function that has slopes of +25 and -10 dB per bark. A convenient analytical expression[Jab02] of the spread function, $SF(k)$, is given by :

$$SF(k) = 15.81 + 7.5(k + 0.474) - 17.5\sqrt{1 + (k + 0.474)^2} \tag{2.45}$$

where k is the critical band number. The spread function is also used in the the auditory model of PESQ. Thus, to comply with the model of PESQ, the same spread function is used in the thesis.

To calculate the frequency domain masking threshold, the following steps are used [Joh88, Jab02]:

(1) Compute an N-point Fast Fourier Transform (FFT) of a frame of speech, e.g. $\mathbf{x}(n)$ (x(n) refers to a frame of any speech signals where we want to calculate the simultaneous masking threshold). The transformed signal is represented by $\tilde{\mathbf{X}}(n)$.

(2) Map the frequency domain to a critical band scale, or Bark scale [Joh88]. Equation (2.43) gives the mapping formula from Hz to Bark. The mapping is performed by calculating the power spectrum of the speech samples in critical band $P(i)$. The squared magnitude of the Fourier Transform (FFT) of $\mathbf{x}(n)$ is computed with a normalization by the frame size

L, *i.e.* $P(i) = \sum_{band\ i} \left\{ \frac{\left|\tilde{X}(n)\right|^2}{L} \right\}$.

(3) To take into account masking between different critical bands, convolve $P(i)$ with the following spread function $SF(k)$ [Jab02] to get the spread critical band spectrum $C(k)$:

$$C(k) = SF(k) * P(k) \tag{2.46}$$

where k is the critical band number, and "$*$" denotes a convolution.

(4) Calculate the masking threshold. The tone-masking-noise threshlod is estimated as $14.5+k$ dB below $C(k)$, where this estimate is from [Sch70] via [Sch79]. The noise-masking-tone threshold is estimated as 5.5 dB below $C(k)$ uniformly across the critical band spectrum. The estimate for noise-masking-tone is due to [Hel72]. A threshold offset $O(k)$ is determined by this two types of thresholds and the final spread threshold estimate $T(k)$ is obtained by subtracting the threshold offset $O(k)$ from the spread critical band spectrum $C(k)$ [Jab02]:

$$T(k) = 10^{(\log_{10}[C(k)]-[(O(k)/10])} \tag{2.47}$$

The calculation of $O(k)$ depends on the noise-like or tone-like nature of the masker and the maskee. The offset $O(k)$ in decibels is calculated as:

$$O(k) = \alpha(14.5+k) + (1-\alpha)5.5 \tag{2.48}$$

To calculate $O(k)$, first the Spectral Flatness Measure (SFM) is defined as follows:

$$SFM_{dB} = 10\log_{10}\frac{G_m}{A_m} \tag{2.49}$$

where G_m is the geometric mean of the power spectrum, and A_m is the arithmetic mean of the power spectrum. Then a coefficient of tonality α is generated as:

$$\alpha = \min(\frac{SFM_{dB}}{SFM_{dB\max}}, 1) \tag{2.50}$$

where $SFM_{dB\max} = -60dB$. So if $SFM = -60dB$, the signal is estimated as entirely tone-like $(\alpha = 1)$ and if $SFM = 0dB$ the signal is completely noise-like ($\alpha = 0$).

(5) Map $T(k)$ back from the Bark scale to the linear frequency domain.

Notions of critical bandwidth and simultaneous masking are illustrated in Fig. 2.6, where we consider the case of a single masking tone occurring at the center of a critical band. All levels in the figure are given in terms of dB SPL. A hypothetical masking tone occurs at some masking level. This generates an excitation along the basilar membrane that is modeled by a spreading function and a corresponding *masking threshold*. For the band under consideration, the *minimum masking threshold* denotes the spreading function in-band minimum. *Signal-to-mask ratio* (SMR) and *noise-to-mask ratio* (NMR) denote the log distances from the minimum masking threshold to the masker and noise levels, respectively.

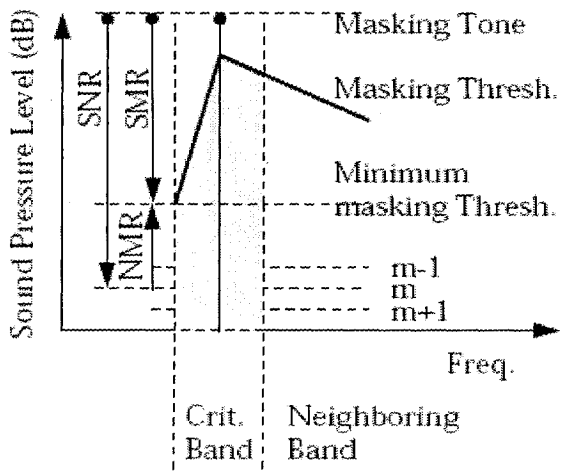


Fig. 2.6. Schematic representation of simultaneous masking

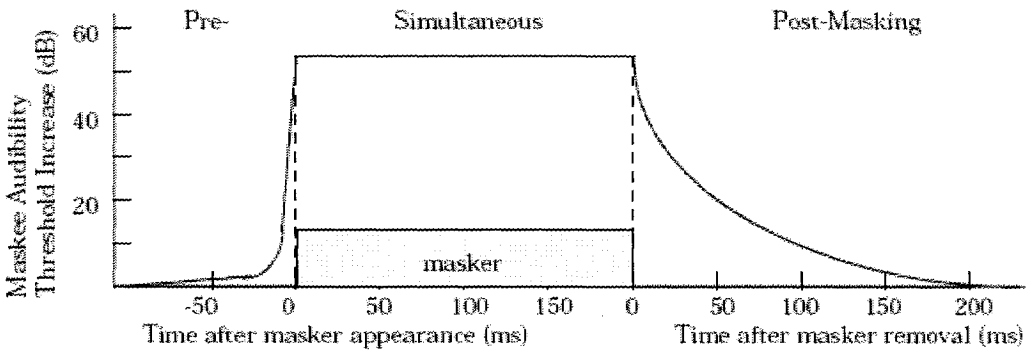


Fig. 2.7. Temporal masking properties of the human ear. Pre-masking occurs prior to masker onset and lasts only a few milliseconds; Post-masking may persist for more than 100 milliseconds after masker removal

Fig. 2.7 shows the temporal masking properties of the human ear. The time domain forward masking effect is often modeled as a psychoacoustic specific loudness versus critical-band rate and time. A non-linear circuit as shown in Fig. 2.8 was proposed in [Hua02] to estimate the output

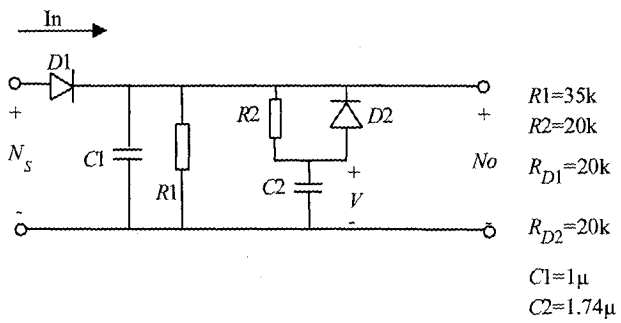


Fig. 2.8 Nonlinear circuit for loudness estimation

specific loudness, one circuit for each critical band. In each circuit, the input $N_S(i)$ is the i^{th} critical-band specific loudness of the current frame, which is computed from $E(i)$ by:

$$N_s(i) = 0.08(E_T(i)/E_0)^{0.23} \cdot \left[(0.5 + 0.5 \cdot E(i)/E_T(i))^{0.23} - 1\right] \tag{2.51}$$

where $E_T(i)$ is the excitation at the absolute threshold of hearing, E_0 is the excitation for 0dB SPL, and $E(i)$ is the convolution of the instantaneous power spectral density of the speech samples with a spread function in critical band, which is the same as $C(i)$ in (2.46). Then the circuit can simulate the nonlinear loudness processing of the neural system and generate the output specific loudness $N_o(i)$.

For numerical computations, the differential equations of the circuit shown in Fig. 2.8 are converted into the following difference equations:

$$I_n = \begin{cases} \dfrac{N_s - N_o^*}{R_{D1}}, & \text{if } N_s > N_o^* \\ 0, & \text{otherwise} \end{cases} \tag{2.52}$$

where R_{D1} is the "on" resistance of diode $D1$, N_s and N_o^* are respectively the specific loudness of the current frame and the output specific loudness of the previous frame. N_o^* and V^* are also respectively the voltages across capacitors $C1$ and $C2$ in the previous frame. If $N_o^* > V^*$, then:

$$N_o = \frac{I_n + C2 \cdot V^* / (\Delta t + C2 \cdot R2) + C1 \cdot N_o^* / \Delta t}{1/R1 + C2/(\Delta t + C2 \cdot R2) + C1/\Delta t} \tag{2.53}$$

$$V = N_o^* \cdot \Delta t / (\Delta t + R2 \cdot C2) + V^* \cdot C2 \cdot R2 / (\Delta t + C2 \cdot R2) \tag{2.54}$$

whereas if $N_o^* < V^*$, then:

$$V = N_o = \frac{I_n + (C1 + C2) \cdot V^* / \Delta t}{1/R1 + (C1 + C2)/\Delta t} \tag{2.55},$$

where Δt is the time difference between two frames in ms. If the masker duration is longer, N_o will be larger. The values of the resistors and capacitors are determined by matching the phenomena measured using maskers longer than 200ms [Zwi84, Kid82]. The total loudness Q, defined as the sum of the output specific loudness in all critical bands, is used as an estimate of the time domain forward masking level.

Taking both the time domain forward masking of this section and the simultaneous frequency domain masking into account, the total masking levels can be determined. The levels depend not only on the current frame but also on the previous frames. The total masking level of the i^{th} critical band at time t is the maximum of the current simultaneous masking level and the total masking level of the previous frame decayed by some constant:

$$M_t(i) = \max\{M_s(i), M_t(i)^* \cdot \exp^{-\Delta t/(\tau(i)\cdot Q)}\} \tag{2.56}$$

where $M_t(i)$ and $M_t(i)^*$ are the total masking levels of the current frame and the previous frame, respectively; $M_s(i)$ is the masking level computed from the simultaneous frequency domain masking model [Joh88], which is the same as $T(i)$ in (2.47), Δt is the time difference between two frames, $\tau(i)$ is the maximum decay time constant in each critical band [Jes82]; and N is the total loudness level which is now normalized by the total loudness of a 60dB uniform masking noise (UMN) [Hua02]. If Q is larger than one, then Q is set to one. So Q lies between zero and one. When Q is zero, there is no forward masking. When Q is one, the total masking level decays with a maximum time constant.

2.2.6.2 Application in Speech Enhancement

In [Vir99], Virag proposed a single channel speech enhancement system based on masking properties of the human auditory system. The paper addresses the problem of speech enhancement at low signal-to-noise ratios (SNR's) (<5dB). It exploits the masking properties of the human auditory system to overcome the limitations of one channel subtractive-type enhancement system in additive background noise at low SNR's. The proposed enhancement scheme is composed of the following main steps:

- spectral decomposition;

- speech/noise detection and estimation of noise during speech pauses;
- calculation of the noise masking threshold;
- adaptation in time and frequency of the subtraction parameters based on the noise masking threshold;
- calculation of the enhanced speech spectral magnitude via parametric subtraction with adapted parameters;
- Inverse transform.

The approach proposed in this paper works toward rendering the residual noise produced by subtractive-type enhancement systems "perceptually white" so the problem of "musical residual noise" introduced by conventional subtractive-type enhancement methods is solved. A parametric subtraction algorithm is applied. Two parameters, the oversubtraction factor and the spectral flooring factor are calculated from the auditory masking threshold.

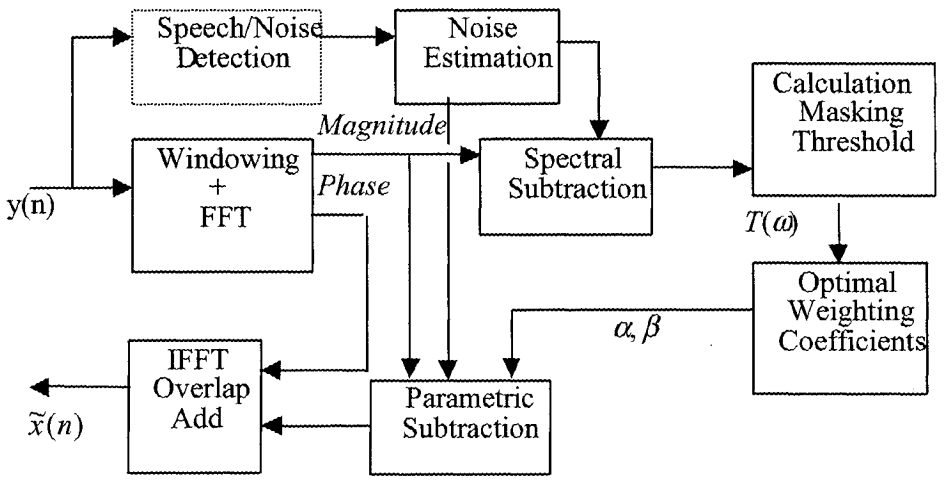


Fig. 2.9 The diagram of the system proposed in [Vir99]

Only simultaneous masking effects are taken into account in the paper. The masking model used was developed by Johnston for coding of audio sampled at 32 kHz in [Joh88]. The following steps are used to calculate the masking threshold:

- Critical band analysis of the signal

- Applying the spreading function to the critical band spectrum
- Calculating the spread masking threshold
- Relating the spread masking threshold to the critical band masking threshold
- Accounting for the absolute threshold of hearing.

Fig. 2.9 shows the diagram of the method proposed in [Vir99].

In the above figure, α is the oversubtraction factor and $\alpha > 1$. β is the spectral flooring factor and $0 \le \beta << 1$. The parametric subtraction block is performed by:

$$G(\omega) = \begin{cases} \left(1 - \alpha \cdot \left[\frac{|\hat{D}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2}, & if \left[\frac{|\hat{D}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1} < \frac{1}{\alpha + \beta} \\ \left(\beta \cdot \left[\frac{|\hat{D}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2}, & otherwise \end{cases} \tag{2.57}$$

where $|Y(\omega)|$ is the noisy speech short-time magnitude, $|\hat{D}(\omega)|$ is the noise spectral magnitude estimate, the exponent $\gamma_1 = 1/\gamma_2$ determines the sharpness of the transition from the $G(\omega) = 1$ (the spectral component is not modified) to the $G(\omega) = 0$ (the spectral component is suppressed).

2.2.7 Other Methods

2.2.7.1 Wavelet Based Methods

In [Bah01], the author proposed a method based on wavelet packet transform and the calculation of the Teager energy for speech enhancement. The approach used discriminative thresholds which

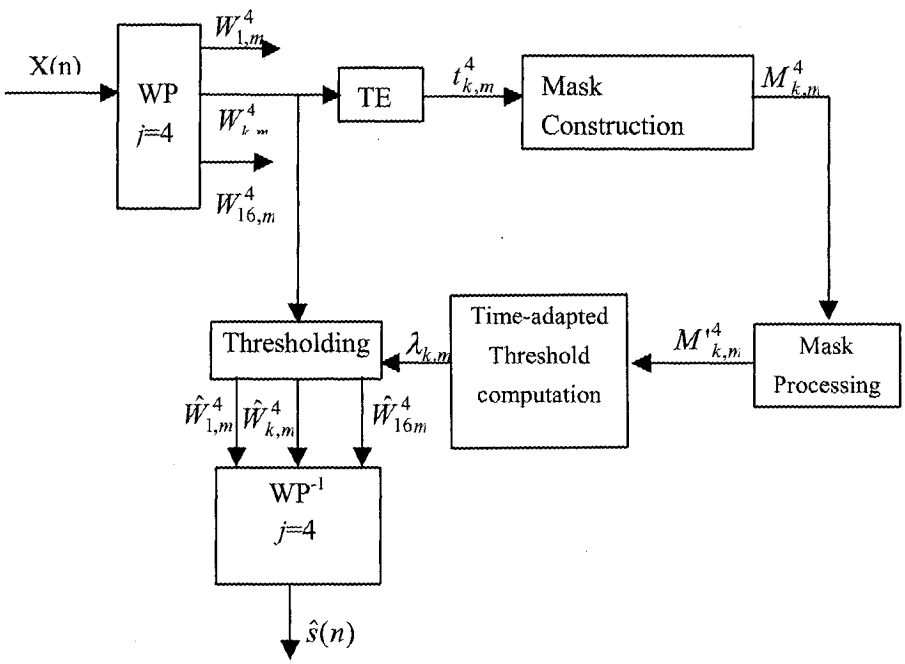


Fig. 2.10 Speech enhancement diagram using time-adapted thresholding in the wavelet packet domain

are time adapted in function of speech components in various scales. It does not require an explicit estimation of the noise level or the *a priori* knowledge of the SNR. The time dependence is introduced by approximating the Teager Energy of the wavelet coefficients.

Figure 2.10 shows the diagram of the proposed algorithm. In this figure, WP means wavelet

packets transform, the decomposition level j is set to be 4. $w_{k,m}^4$ defines the mth coefficient of the kth subband, where $m = 1,\cdots, N/2^4$ and $k = 1,\cdots,2^4$. N is the sample number of the speech signal. TE block calculates Teager energy approximation, $t_{k,m}^4 = \Psi_d[w_{k,m}^4]$, $k = 1,\cdots,2^4$, where $\Psi_d[x(n)] = [x(n)]^2 - x(n+1)x(n-1)$. This operation enhances the discriminability of speech coefficients among those of noise. In the mask construction block, for each subband coefficient, an initial mask is obtained by smoothing the TEO coefficients as $M_{k,m}^4 = t_{k,m}^4 * h_k(m)$ where h_k is an IIR low-pass filter (second order). After constructing masks, the masks have to be processed for the time-adapted threshold as follows:

$$M'^4_{k,m} = \left[\frac{M_{k,m}^4 - S_k^4}{\max(M_{k,m}^4 - S_k^4)}\right]^{1/8} \tag{2.58}$$

where S_k^4 is named *offset*, which estimates the valley's level of $M_{k,m}^4$. It is used to distinguish frames between speech presence and absence because the speech presence is interpreted by a significant contrast between peaks and valleys of $M_{k,m}^4$, while its absence is observed with a weaker contrast (smoother masks). S_k^4 is given by the abscissa of the maximum of the amplitude distribution H of the corresponding mask $M_{k,m}^4$ and is estimated over the analyzed frame

$$S_k^4 = abscissa[\max(H(M_{k,m}^4))]. \tag{2.59}$$

Then for a given subband k, the proposed approach defines the time adapted threshold as

$$\lambda_{k,m} = \lambda(1 - \alpha M'^4_{k,m}) \tag{2.60}$$

where λ is the standard threshold defined as $\lambda = \sigma\sqrt{2\log(N\log_2 N)}$, σ is the noise level. In the thresholding block the soft thresholding is applied to the wavelet packet coefficients

$$\hat{w}_{k,m}^4 = T_S(\lambda_a, w_{k,m}^4) \tag{2.61}$$

where λ_a is the threshold corresponding to the analyzed frame and is adapted only for speech frames and kept unchanged for non-speech ones.

$$\lambda_a = \begin{cases} \lambda_{k,m}, & if\ S_k^4 \le 0.35 \max(M_{k,m}^4) \\ \lambda, & if\ S_k^4 < 0.35 \max(M_{k,m}^4) \end{cases} \tag{2.62}$$

$T_S(,)$ is the soft thresholding function defined as

$$T_S(\lambda, w_k) = \begin{cases} \mathrm{sgn}(w_k)(|w_k| - \lambda), & if\ |w_k| \ge \lambda \\ 0, & if\ |w_k| < \lambda \end{cases} \tag{2.63}$$

The last step of the algorithm is the inverse transformation WP^{-1} of the resulting wavelet coefficients.

Unlike windowing transform methods, the whole speech frame is filtered once by the wavelet packets filter. However, a minimum length of the frame is required for the WPF. It must be sufficiently large to include simultaneously speech and silence.

2.2.7.2 Comb Filtering for Speech Enhancement

In [Ami91], a frequency-domain LMS comb filter is proposed for harmonic cancellation. Fig. 2.11 shows the diagram of the proposed system. It implements the least mean squares algorithm (LMS) in the frequency domain to simulate a fixed-notch filter that nulls sinusoids with harmonically related frequencies. The Fourier transform (FT) block length is set to represent the fundamental period of the undesired periodic signal.

Deterministic signals rather than observations are applied to the reference inputs. The concept behind this harmonic canceller is based on the fact that if the Fourier transform block length, *N*, is chosen equal to the fundamental period of the undesired periodic signal, then the harmonic signal transform at each frequency bin takes the same value irrespective of the time reference of the data

block. $d(n)$ is the noisy speech signal. The nonharmonic frequency components transform

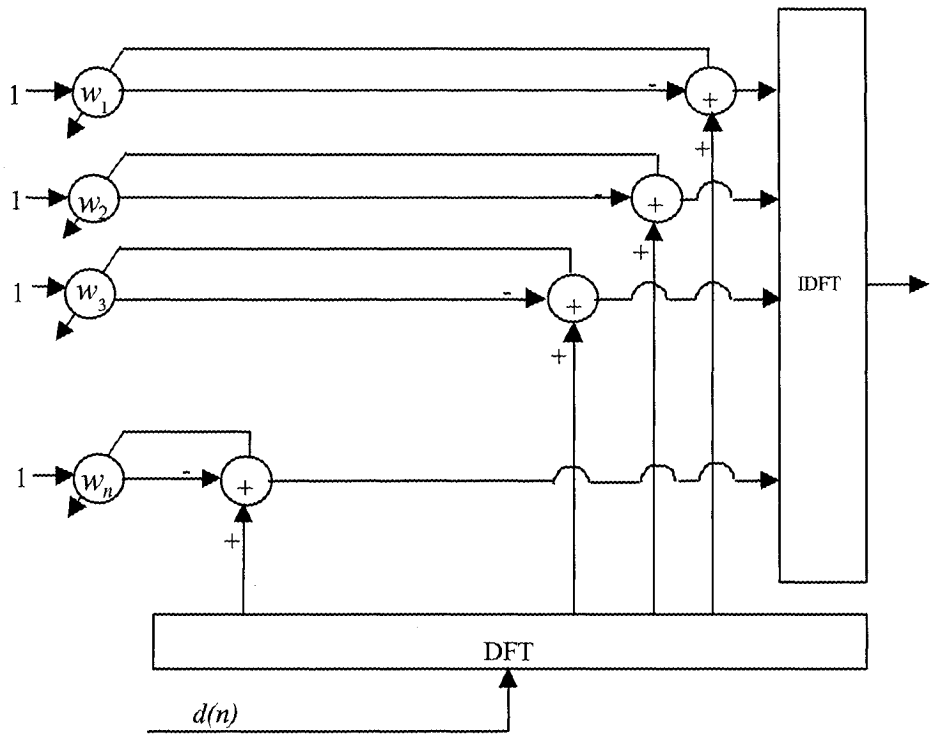


Fig. 2. 11 Diagram of the comb filtering system proposed in [Ami91]

exhibits a time-varying behavior. The LMS algorithm operates on canceling all harmonic frequency components leaving out other components in the signal to propagate to the error transform with minimum distortion.

This harmonic canceller could be used to reduce the noise from an interfering speaker, provided this speaker's fundamental frequency is known. Nulling time-varying fundamental and harmonic frequencies while maintaining a constant gain cannot be achieved using the proposed structure since it requires a variable transform block length.

The adaptive comb filter is a technique developed to enhance speech by exploiting the periodical nature of the voiced speech. However, this technique may also introduce temporal distortion whenever the speech signal exhibits larger variations in pitch and formant frequencies. [Gra93] proposed an approach to warp adjacent pitch periods into the underlying pitch period, so as to rectify pitch variations. But the work is done under the assumption that the pitch information is obtainable. In fact, similar results are attainable by adaptively altering the filter taps on a sample-by-sample basis. However this method would inevitably raise the amount of computation. In [HuH98], a compromise between the sample-wise adaptation and computational efficacy is made by keeping filter taps intact within a moderate interval and to smooth the transition of filter taps between adjacent intervals. A coarse estimate of the pitch period is performed for every frame using the autocorrelation method. A comb filter using the estimated pitch period, defined as p, is constituted as the following form:

$H(z) = \frac{1 + \alpha z^{p} + \beta z^{-p}}{1 + |\alpha| + |\beta|}$ where α and β are the filter taps.

The minimum sum-squared-error between the ensemble of the underlying subframe and those which are one period away from the current one is chosen as the criterion to determine the filter taps.

2.3 Summary

Table 2.2 Comparison between different methods

	Wiener	**MMSE**	**KLT**	**HMM**	**Kalman**
Advantage	Straight-forward, easy to implement				Obtain MMSE or LMMSE estimate
Disadvantage	Musical noise	Rely on the estimation	Poor at low SNR	Large model size, large comput. load	Some time consuming
Speech Characteristic Included	No	No	Yes	Yes	Yes
Masking Properties of Human Auditory Systems Included	No	No	No	No	No
Estimation of Noise Statistics or SNR used	Yes	Yes	Yes	Yes	Yes
VAD	Yes	Yes	Yes	No	No

Table 2.2 gives a comparison between some of the different methods mentioned in this chapter. From the table, it can be observed that very few of the recently published methods have included

perceptual models. Most of the methods required a VAD (Voice Activity Detector) and did not include some of the advantages of the Kalman based methods, i.e. no need to estimate the SNR, and no musical noise. In addition, the Kalman filtering algorithm can produce a MMSE or LMMSE estimate of the speech signal, so it can be a good candidate for the speech enhancement application. In the single microphone system, no *a priori* knowledge of the speech and noise can be obtained, so a method that can exploit most of the available resources is needed. Most methods only use the speech and noise statistics. However, since the speech quality is a kind of subjective issue, it depends much on the hearing properties of human auditory systems. Very few methods have used this property. This thesis studies the applicability of the masking properties to Kalman filtering algorithms for speech enhancement. A truly perceptually constrained Kalman filter is derived in Chapter 3. Both time domain and frequency domain masking properties are used. To reduce the computation cost, a heuristic method which combines the Kalman filter with a perceptual post-filter is proposed and tested in Chapter 4. The perceptual post-filter is constructed based on either frequency domain masking properties or both time domain and frequency domain masking properties of auditory systems. To further reduce the computational cost, a wavelet domain Kalman filter and post-filter is proposed in Chapter 5. By this method, the calculation of the Kalman filtering algorithm and the masking thresholds are performed in the same domain, *i.e.*, wavelet domain, thus the switching between the calculation of masking thresholds in frequency domain and the time domain Kalman filtering is avoided. The methods in Chapter 3, 4 and 5 and many other published methods [Vir99, Mit00, Pop98] only consider a linear speech model. However, the speech production process is nonlinear, and a nonlinear speech model is a more accurate representation. To test the effect of a nonlinear speech model on the speech enhancement algorithms, a dual unscented Kalman filter (UKF) is studied in Chapter 6. A truly perceptually constrained dual UKF algorithm and the heuristic alternative are both discussed and tested.

2.4 Performance Index of the Speech Enhancement Algorithms

To evaluate the speech enhancement algorithms, this thesis uses the Perceptual Evaluation of Speech Quality (PESQ) [Rix01, ITUT01], an ITU-T recommendation P.862, as the performance index. As a recently published standard (published in 2001), PESQ was only used very recently by researchers as the performance index for speech quality measurement [Che04, Coh04]. PESQ is an objective speech quality assessment model. As an enhanced version of the perceptual speech quality measure (PSQM), it was developed for use across a wider range of network conditions, including analogue connections, codecs, packet loss and variable delay. Background noise and processing noise can be assessed by presenting PESQ with the clean, unprocessed original and the degraded signal. PESQ uses a sensory model to compare the original, unprocessed signal with the degraded signal. The resulting quality score is analogous to the subjective "Mean Opinion Score" (MOS).

The structure of the PESQ model is shown in Fig. 2.12. The model begins by level aligning both signals to a standard listening level. They are filtered (using an FFT) with an input filter to model a standard telephone handset. The signals are aligned in time and then processed through an auditory transform similar to that of PSQM [Bee94]. The transformation also involves equalizing for linear filtering in the system and for gain variation. Two distortion parameters are extracted from the disturbance (the difference between the transforms of the signals), and are aggregated in frequency and time and mapped to a prediction of subjective mean opinion score (MOS). A linear combination of disturbance parameters was used as a predictor of subjective MOS. A combination of only two parameters, one symmetric disturbance (d_{sym}) and one asymmetric disturbance (d_{asym}), gave a good balance between accuracy of prediction and ability to generalize. The following equation gives the output mapping used in PESQ:

$$PESQMOS = 4.5 - 0.1d_{sym} - 0.0309d_{asym} \tag{2.64}$$

For normal subjective test material the values lie between 1.0 (bad) and 4.5 (no distortion). In extremely high distortion the PESQMOS may fall below 1.0.

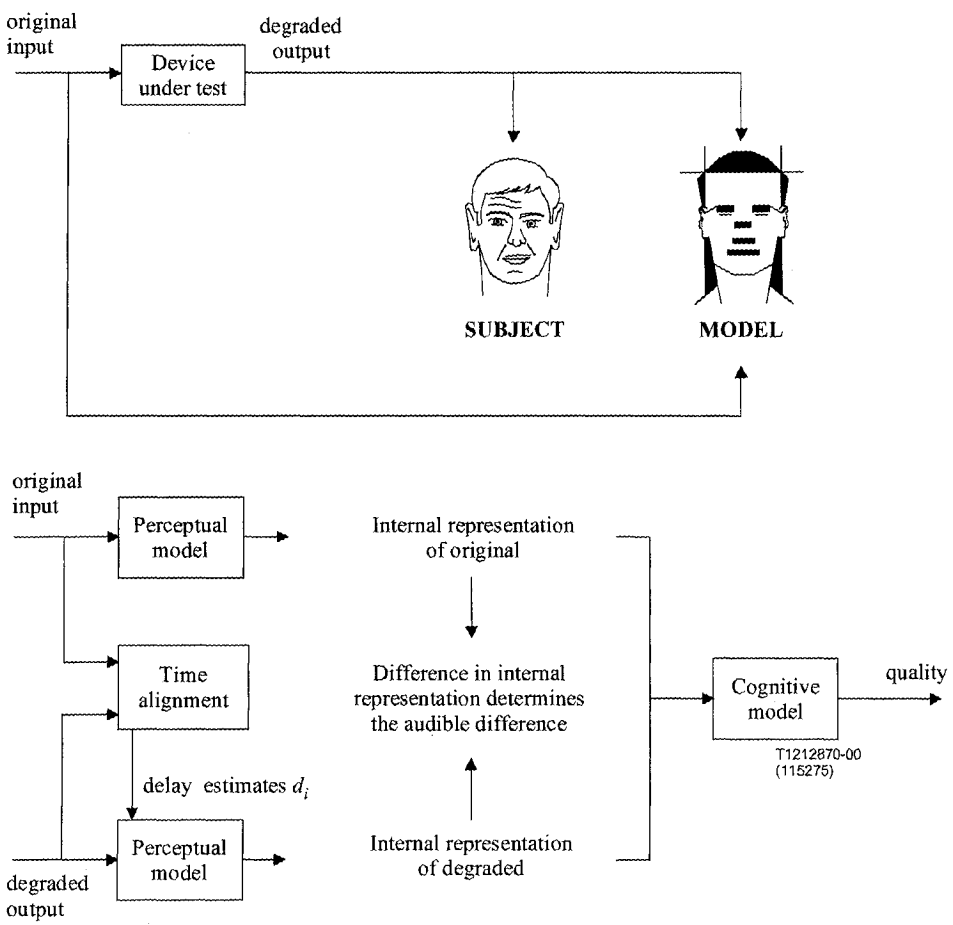


Fig. 2.12 The structure of PESQ model from [ITUT01]

Chapter 3 Frequency and Time Domain Auditory Masking Threshold Constrained Kalman Filters

3.1 Introduction

This chapter derives true masking threshold constrained Kalman filters where the masking threshold is computed from both time-domain and frequency-domain masking properties of human auditory systems. In the Kalman filtering algorithm, the estimate error covariance matrix is updated at each step. The goal of a standard Kalman algorithm is to minimize the estimate error variance. However, when the masking properties of a human auditory system are taken into account, the problem becomes to minimize the estimate error variance under the constraint that the energy of the estimate error should be smaller than the masking threshold. The derived constrained Kalman filtering algorithms are then applied to speech enhancement. The simulation results show that the proposed algorithms produce better ITU-T P.862 PESQ scores than the algorithms in [Gab01, Vir99].

3.2 Summary of Derivation of a Standard Kalman Filtering Algorithm for Speech Enhancement

3.2.1 Kalman Filtering for Speech Degraded by White Noise

In [Bro92] the derivation of a standard Kalman filter was proposed. The summary of the

derivation particular to applying the Kalman filter to speech enhancement for the white noise case was presented in section 2.2.4, and it is reproduced below with more details.

The clean speech signal $s(n)$ is modeled as a p-th order AR process as follows:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{3.1}$$

The measured speech signal $y(n)$ is

$$y(n) = s(n) + v(n) \tag{3.2}$$

where a_i is the i-th AR coefficient, $u(n)$ is the model noise process and $v(n)$ is the background noise. $u(n)$ and $v(n)$ are uncorrelated Gaussian white noise processes with means $\bar{u}$ and $\bar{v}$ and variances σ_u^2 and σ_v^2.

The state space model, written as follows, can then be obtained.

$$\mathbf{x}(n) = \mathbf{F}\mathbf{x}(n-1) + \mathbf{G}u(n) \tag{3.3}$$

$$y(n) = \mathbf{H}\mathbf{x}(n) + v(n) \tag{3.4}$$

where

$$\mathbf{x}(n) = [s(n-p+1) \cdots s(n)]^T \tag{3.5}$$

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{3.6}$$

$$\mathbf{H} = \mathbf{G}^T = [0 \quad 0 \quad \cdots \quad 0 \quad 1] \tag{3.7}$$

where p is the order of the AR process, $\mathbf{F}$ is a $p \times p$ transition matrix, $\mathbf{G}$ and $\mathbf{H}$ are respectively the $p \times 1$ input vector and the $1 \times p$ observation row vector.

The Kalman filtering algorithm for speech enhancement is as follows [Gab01]:

$$e(n) = y(n) - \mathbf{H}\hat{\mathbf{x}}(n \mid n-1) - \bar{v} \tag{3.8}$$

$$\mathbf{K}(n) = \mathbf{P}(n \mid n-1)\mathbf{H}^T \times (\mathbf{H}\mathbf{P}(n \mid n-1)\mathbf{H}^T + \sigma_v^2)^{-1} \tag{3.9}$$

$$\hat{\mathbf{x}}(n \mid n) = \hat{\mathbf{x}}(n \mid n-1) + \mathbf{K}(n)e(n) \tag{3.10}$$

$$\mathbf{P}(n \mid n) = (\mathbf{I} - \mathbf{K}(n)\mathbf{H})\mathbf{P}(n \mid n-1) \tag{3.11}$$

$$\hat{\mathbf{x}}(n+1 \mid n) = \mathbf{F}\hat{\mathbf{x}}(n \mid n) + \mathbf{G}\bar{u} \tag{3.12}$$

$$\mathbf{P}(n+1 \mid n) = \mathbf{F}\mathbf{P}(n \mid n)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T\sigma_u^2 \tag{3.13}$$

The derivation of the above equations (3.8) to (3.13) is shown below.

From the goal of the standard Kalman filter, an unbiased minimum variance estimate of the state $\mathbf{x}(n)$ is formed as

$$\hat{\mathbf{x}}(n \mid n) = \mathbf{K}'(n)\hat{\mathbf{x}}(n-1 \mid n-1) + \mathbf{K}(n)y(n) + C(n) \tag{3.14}$$

where $C(n)$ is a scalar at time n. Note that this estimator is both linear in $\hat{\mathbf{x}}(n-1 \mid n-1)$ and $y(n)$.

From the unbiased criteria, an expression for $\mathbf{K}'(n)$ can be obtained. Since $\hat{\mathbf{x}}(n \mid n)$ is unbiased, $E[\hat{\mathbf{x}}(n \mid n) - \mathbf{x}(n)] = 0$. From equations (3.3), (3.4) and (3.14), using substitutions, the terms in the brackets of $E[\hat{\mathbf{x}}(n \mid n) - \mathbf{x}(n)] = 0$ becomes

$$\begin{aligned}
\hat{\mathbf{x}}(n \mid n) - \mathbf{x}(n) &= \mathbf{K}'(n)\hat{\mathbf{x}}(n-1 \mid n-1) + \mathbf{K}(n)y(n) - \mathbf{x}(n) + C(n) \\
&= \mathbf{K}'(n)\hat{\mathbf{x}}(n-1 \mid n-1) + \mathbf{K}(n)[\mathbf{H}\mathbf{x}(n) + v(n)] - \mathbf{x}(n) \\
&\quad - \mathbf{K}'(n)\mathbf{x}(n-1) + \mathbf{K}'(n)\mathbf{x}(n-1) + C(n) \\
&= \mathbf{K}'(n)[\hat{\mathbf{x}}(n-1 \mid n-1) - \mathbf{x}(n-1)] + \mathbf{K}(n)\{\mathbf{H}[\mathbf{F}\mathbf{x}(n-1) + \mathbf{G}u(n)] + v(n)\} \\
&\quad - [\mathbf{F}\mathbf{x}(n-1) + \mathbf{G}u(n)] + \mathbf{K}'(n)\mathbf{x}(n-1) + C(n) \\
&= \mathbf{K}'(n)[\hat{\mathbf{x}}(n-1 \mid n-1) - \mathbf{x}(n-1)] + [\mathbf{K}(n)\mathbf{H}\mathbf{F} - \mathbf{F} + \mathbf{K}'(n)]\mathbf{x}(n-1) \\
&\quad + [\mathbf{K}(n)\mathbf{H}\mathbf{G} - \mathbf{G}]u(n) + \mathbf{K}(n)v(n) + C(n)
\end{aligned} \tag{3.15}$$

Take the expectation of equation (3.15) and make it equal to zero. A solution is found by forcing:

$$[\mathbf{K}(n)\mathbf{H}\mathbf{F} - \mathbf{F} + \mathbf{K}'(n)]E[\mathbf{x}(n-1)] = 0 \tag{3.16}$$

$$[\mathbf{K}(n)\mathbf{HG}-\mathbf{G}]\bar{u}+\mathbf{K}(n)\bar{v}+C(n)=0 \tag{3.17}$$

Equation (3.16) implies that

$$\mathbf{K}(n)\mathbf{HF}-\mathbf{F}+\mathbf{K}'(n)=0 \tag{3.18}$$

Thus

$$\mathbf{K}'(n)=(\mathbf{I}-\mathbf{K}(n)\mathbf{H})\mathbf{F} \tag{3.19}$$

From equation (3.17), the following can be obtained:

$$C(n)=[\mathbf{G}-\mathbf{K}(n)\mathbf{HG}]\bar{u}-\mathbf{K}(n)\bar{v} \tag{3.20}$$

Substitute equation (3.19) and (3.20) into equation (3.14) and use equation (3.8) and (3.12), then equation (3.10) can be obtained, which is the state update equation.

What remains is to find the value of $\mathbf{K}(n)$ which minimizes the variance of the estimate error.

The estimation error process is defined as

$$\tilde{\mathbf{x}}(n\mid n)=\hat{\mathbf{x}}(n\mid n)-\mathbf{x}(n) \tag{3.21}$$

$\mathbf{P}(n\mid n)$ is the covariance matrix of $\tilde{\mathbf{x}}(n\mid n)$. $\mathbf{K}(n)$ is called a Kalman gain which minimizes the trace of $\mathbf{P}(n\mid n)$.

From the definition,

$$\mathbf{P}(n\mid n)=E[\tilde{\mathbf{x}}(n\mid n)\tilde{\mathbf{x}}^T(n\mid n)] \tag{3.22}$$

To find $\mathbf{K}(n)$, the expression of the prediction error covariance matrix, $\mathbf{P}(n\mid n-1)$, is first found. The predicted state vector is defined by equation (3.12), thus the prediction error is as follows:

$$\tilde{\mathbf{x}}(n\mid n-1)=\hat{\mathbf{x}}(n\mid n-1)-\mathbf{x}(n)=\mathbf{F}\hat{\mathbf{x}}(n-1\mid n-1)-\mathbf{F}\mathbf{x}(n-1)+\mathbf{G}\bar{u}-\mathbf{G}u(n) \tag{3.23}$$

By expanding $E[\tilde{\mathbf{x}}(n\mid n-1)\tilde{\mathbf{x}}^T(n\mid n-1)]$, the following equation is obtained:

$$\begin{aligned}\mathbf{P}(n\mid n-1) &= E\left\{\tilde{\mathbf{x}}(n\mid n-1)\tilde{\mathbf{x}}^T(n\mid n-1)\right\}\\ &= \mathbf{F}E\left\{[\hat{\mathbf{x}}_k(n-1\mid n-1)-\mathbf{x}(n-1)][\hat{\mathbf{x}}(n-1\mid n-1)-\mathbf{x}(n-1)]^T\right\}\mathbf{F}^T+\mathbf{G}E\{[u(n)-\bar{u}]^2\}\mathbf{G}^T\\ &= \mathbf{F}\mathbf{P}(n-1\mid n-1)\mathbf{F}^T+\mathbf{G}\mathbf{G}^T\sigma_u^2\end{aligned} \tag{3.24}$$

Thus equation (3.13) is obtained.

Secondly, the expression of $\mathbf{P}(n\mid n)$ is to be derived.

From equation (3.8) and (3.10), the following equation can be obtained:

$$\begin{aligned}\hat{\mathbf{x}}(n\mid n) &= \hat{\mathbf{x}}(n\mid n-1)+\mathbf{K}(n)[y(n)-\mathbf{H}\hat{\mathbf{x}}(n\mid n-1)-\bar{v}]\\ &= [\mathbf{I}-\mathbf{K}(n)\mathbf{H}]\hat{\mathbf{x}}(n\mid n-1)+\mathbf{K}(n)y(n)-\mathbf{K}(n)\bar{v}\end{aligned} \tag{3.25}$$

Substitute equation (3.25) in equation (3.21), then

$$\begin{aligned}\tilde{\mathbf{x}}(n\mid n) &= [\mathbf{I}-\mathbf{K}(n)\mathbf{H}]\hat{\mathbf{x}}(n\mid n-1)+\mathbf{K}(n)y(n)-\mathbf{K}(n)\bar{v}-\mathbf{x}(n)\\ &= [\mathbf{I}-\mathbf{K}(n)\mathbf{H}]\hat{\mathbf{x}}(n\mid n-1)+\mathbf{K}(n)[\mathbf{H}\mathbf{x}(n)+v(n)]-\mathbf{x}(n)-\mathbf{K}(n)\bar{v}\\ &= \hat{\mathbf{x}}(n\mid n-1)-\mathbf{K}(n)\mathbf{H}\hat{\mathbf{x}}(n\mid n-1)+\mathbf{K}(n)\mathbf{H}\mathbf{x}(n)+\mathbf{K}(n)v(n)-\mathbf{x}(n)-\mathbf{K}(n)\bar{v}\\ &= [\hat{\mathbf{x}}(n\mid n-1)-\mathbf{x}(n)]-\mathbf{K}(n)\mathbf{H}[\hat{\mathbf{x}}(n\mid n-1)-\mathbf{x}(n)]+\mathbf{K}(n)v(n)-\mathbf{K}(n)\bar{v}\\ &= [\mathbf{I}-\mathbf{K}(n)\mathbf{H}][\hat{\mathbf{x}}(n\mid n-1)-\mathbf{x}(n)]+\mathbf{K}(n)v(n)-\mathbf{K}(n)\bar{v}\\ &= [\mathbf{I}-\mathbf{K}(n)\mathbf{H}]\tilde{\mathbf{x}}(n\mid n-1)+\mathbf{K}(n)v(n)-\mathbf{K}(n)\bar{v}\end{aligned} \tag{3.26}$$

So the covariance matrix $\mathbf{P}(n\mid n)$ can be obtained as follows:

$$\begin{aligned}\mathbf{P}(n\mid n) &= E[\tilde{\mathbf{x}}(n\mid n)\tilde{\mathbf{x}}^T(n\mid n)]\\ &= [\mathbf{I}-\mathbf{K}(n)\mathbf{H}]\mathbf{P}(n\mid n-1)[\mathbf{I}-\mathbf{K}(n)\mathbf{H}]^T+\mathbf{K}(n)\mathbf{K}^T(n)\sigma_v^2\end{aligned} \tag{3.27}$$

Note that $\mathbf{P}(n\mid n)$ is a function of $\mathbf{K}(n)$, $\mathbf{P}(n\mid n-1)$ and σ_v^2.

The final step is to find an expression for $\mathbf{K}(n)$ which minimizes the trace of $\mathbf{P}(n\mid n)$. For simplicity, we use $\mathbf{P}$ and $\mathbf{K}$ to represent $\mathbf{P}(n\mid n-1)$ and $\mathbf{K}(n)$.

$$\begin{aligned}\mathbf{P}(n\mid n) &= (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{P}(\mathbf{I}-\mathbf{K}\mathbf{H})^T+\mathbf{K}\mathbf{K}^T\sigma_v^2\\ &= \mathbf{P}-\mathbf{K}\mathbf{H}\mathbf{P}-\mathbf{P}\mathbf{H}^T\mathbf{K}^T+\mathbf{K}\mathbf{H}\mathbf{P}\mathbf{H}^T\mathbf{K}^T+\mathbf{K}\mathbf{K}^T\sigma_v^2\end{aligned} \tag{3.28}$$

$$\mathrm{Tr}\mathbf{P}(n\mid n) = \mathrm{Tr}\mathbf{P}-2\mathrm{Tr}\mathbf{K}\mathbf{H}\mathbf{P}+\mathrm{Tr}\mathbf{K}(\mathbf{H}\mathbf{P}\mathbf{H}^T)\mathbf{K}^T+\mathrm{Tr}\mathbf{K}\mathbf{K}^T\sigma_v^2 \tag{3.29}$$

where Tr represents the calculation of the trace of a matrix. $\mathbf{K}$ must minimize the value of $\mathrm{Tr}\mathbf{P}(n\mid n)$. Taking the partial derivative with respect to $\mathbf{K}$,

$$\frac{\partial \mathrm{Tr}\mathbf{P}(n|n)}{\partial \mathbf{K}} = -2\mathbf{P}\mathbf{H}^T + 2\mathbf{K}\mathbf{H}\mathbf{P}\mathbf{H}^T + 2\mathbf{K}\sigma_v^2 \tag{3.30}$$

Setting the right side of equation (3.30) equal to zero and solving for **K** gives

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \sigma_v^2)^{-1} \tag{3.31}$$

which is the Kalman gain.

Thus the Kalman gain for a standard Kalman filtering algorithm is derived.

Substituting equation (3.31) in equation (3.28), equation (3.11) is obtained.

3.2.2 Kalman Filtering for Speech Degraded with Colored Noise

In a colored noise environment, a vector Kalman filter is used [Chu99]. The clean speech signal $s(n)$ and the n-th sample of the noisy speech signal $y(n)$ are represented by (3.1) and (3.2), however the measurement noise $v(n)$ is a colored measurement noise process with covariance matrix **R**. Using a vector Kalman filter, the state-space model is expressed as

$$\mathbf{x}(n) = \mathbf{F}\mathbf{x}(n-1) + \mathbf{G}u(n) \tag{3.32}$$

$$\mathbf{y}(n) = \mathbf{H}\mathbf{x}(n) + \mathbf{v}(n) \tag{3.33}$$

where

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{3.34}$$

$$\mathbf{x}(n) = [s(n-p+1)\cdots s(n)]^T \tag{3.35}$$

$$\mathbf{y}(n) = [y(n-p+1)\cdots y(n)]^T \tag{3.36}$$

$$\mathbf{v}(n) = [v(n-p+1)\cdots v(n)]^T \tag{3.37}$$

$$\mathbf{G} = [0\ 0\cdots 1]^T \tag{3.38}$$

H equals a p-th order identity matrix.

The Kalman filtering algorithm for the speech enhancement in colored noise is as follows:

$$\mathbf{e}(n) = \mathbf{y}(n) - \hat{\mathbf{x}}(n \mid n-1) - \bar{\mathbf{v}} \tag{3.39}$$

$$\mathbf{K}(n) = \mathbf{P}(n \mid n-1) \times (\mathbf{P}(n \mid n-1) + \mathbf{R})^{-1} \tag{3.40}$$

$$\hat{\mathbf{x}}(n \mid n) = \hat{\mathbf{x}}(n \mid n-1) + \mathbf{K}(n) \times \mathbf{e}(n) \tag{3.41}$$

$$\mathbf{P}(n \mid n) = (\mathbf{I} - \mathbf{K}(n)) \times \mathbf{P}(n \mid n-1) \tag{3.42}$$

$$\hat{\mathbf{x}}(n+1 \mid n) = \mathbf{F}\hat{\mathbf{x}}(n \mid n) \tag{3.43}$$

$$\mathbf{P}(n+1 \mid n) = \mathbf{F}\mathbf{P}(n \mid n)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T\sigma_u^2 \tag{3.44}$$

where $\mathbf{e}(n)$ is the Innovation vector; $\bar{v}$ is the mean of $v(n)$; $\mathbf{R}$ is the covariance of $v(n)$, $\mathbf{K}(n)$ is the Kalman gain; $\hat{\mathbf{x}}(n \mid n)$ represents the filtered estimate of state vector $\mathbf{x}(n)$; $\hat{\mathbf{x}}(n \mid n-1)$ is the minimum mean-square estimate of the state vector $\mathbf{x}(n)$ given the past observations $y(1)$, …, y(n-1); $\mathbf{P}(n \mid n)$ is the filtered state error covariance matrix; and $\mathbf{P}(n \mid n-1)$ is the predicted state error covariance matrix.

To find an expression for $\mathbf{K}(n)$ for colored noise process, the error covariance matrix $\mathbf{P}(n \mid n)$ is first derived as a function of $\mathbf{K}(n)$, $\mathbf{P}(n \mid n-1)$ and **R**.

$$\mathbf{P}(n \mid n) = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}(\mathbf{I} - \mathbf{K}\mathbf{H})^T + \mathbf{K}\mathbf{R}\mathbf{K}^T = \mathbf{P} - \mathbf{K}\mathbf{H}\mathbf{P} - \mathbf{P}\mathbf{H}^T\mathbf{K}^T + \mathbf{K}\mathbf{H}\mathbf{P}\mathbf{H}^T\mathbf{K}^T + \mathbf{K}\mathbf{R}\mathbf{K}^T \tag{3.45}$$

The expression for $\mathbf{K}(n)$ can be obtained by minimizing the trace of $\mathbf{P}(n \mid n)$ if only the variance of the error is to be minimized. As in the previous section, **P** and **K** are used to represent $\mathbf{P}(n \mid n-1)$ and $\mathbf{K}(n)$.

$$\mathrm{Tr}\mathbf{P}(n \mid n) = \mathrm{Tr}\mathbf{P} - 2\mathrm{Tr}\mathbf{K}\mathbf{H}\mathbf{P} + \mathrm{Tr}\mathbf{K}(\mathbf{H}\mathbf{P}\mathbf{H}^T)\mathbf{K}^T + \mathrm{Tr}\mathbf{K}\mathbf{R}\mathbf{K}^T \tag{3.46}$$

where Tr represents the calculation of the trace of a matrix. **K** must minimize the value of $\mathrm{Tr}\mathbf{P}(n \mid n)$. Taking the partial derivative with respect to **K**,

$$\frac{\partial \mathrm{Tr}\mathbf{P}(n \mid n)}{\partial \mathbf{K}} = -2\mathbf{P}\mathbf{H}^T + 2\mathbf{K}\mathbf{H}\mathbf{P}\mathbf{H}^T + 2\mathbf{K}\mathbf{R} \tag{3.47}$$

Setting equation (3.30) equal to zero and solving for **K** gives

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1} \tag{3.48}$$

which is the Kalman gain.

Thus the Kalman gain for a standard Kalman filtering algorithm with colored background noise is derived.

3.3 Summary of Estimation of the Model and Measurement Noise Process

When the measurement noise process is assumed to be a white noise, the state space model for speech enhancement is a scalar state-space model as shown in (3.3) and (3.4). However, if the noise is assumed to be colored, the state space model is changed to be a vector state-space model as (3.32) and (3.33) show. For these two models, the estimation of the model and measurement noise process is different. In [Gab01], the author derived a recurrent estimation method to obtain the statistics of the model and measurement noise process when the measurement is assumed to be white. The method is based on the work of [Mye76]. We summarize the method in this section, and we also adapt the method for the case of colored measurement noise. The noise statistics computation is based on a covariance matching method. The measurement noise covariance matrix is found by using the theoretical and estimated covariance of the innovation process, while the model noise variance is found by using the theoretical and estimated variance of the residual process.

3.3.1 Estimation of the Measurement Additive Noise Statistics

Assume that the measurement noise $v(n)$ is white noise process and the mean and variance of the measurement noise $v(n)$ is constant over N samples, $v(n), v(n-1), \cdots, v(n-N+1)$. From equation (3.4), the samples of the additive noise are written as

$$v(n) = y(n) - \mathbf{H}\mathbf{x}(n) \tag{3.49}$$

Since the true state vector $\mathbf{x}(n)$ is unknown, only the approximation of $\mathbf{x}(n)$ can be used [Mye76]:

$$\alpha(n) = y(n) - \mathbf{H}\,\hat{\mathbf{x}}(n \mid n-1) = \mathbf{H}\,\tilde{\mathbf{x}}(n \mid n-1) + v(n) \tag{3.50}$$

The samples $\alpha(n)$ are assumed to be representative of $v(n)$ and can be considered independent and identically distributed [Mye76]. Based on the last N measurements $\alpha(n), \alpha(n-1), \cdots, \alpha(n-N+1)$, the mean value $\bar{\alpha}(n)$ and the variance $\sigma_\alpha^2(n)$ are estimated. An unbiased estimator for $\bar{\alpha}(n)$ is taken as the sample mean

$$\hat{\bar{\alpha}}(n) = \frac{1}{N}\sum_{i=0}^{N-1}\alpha(n-i) \tag{3.51}$$

and an unbiased estimator for $\sigma_\alpha^2(n)$ is obtained by

$$\hat{\sigma}_\alpha^2(n) = \frac{1}{N-1}\sum_{i=0}^{N-1}\left[\alpha(n-i) - \hat{\bar{\alpha}}(n)\right]^2 \tag{3.52}$$

The estimation of the additive noise mean is

$$\hat{\bar{v}}(n) = \hat{\bar{\alpha}}(n) \tag{3.53}$$

The expected value of $\sigma_\alpha^2(n)$ is

$$E\left\{\sigma_\alpha^2(n)\right\} = \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{H}\mathbf{P}(n-i \mid n-i-1)\mathbf{H}^T + \sigma_v^2 \tag{3.54}$$

Matching (3.52) and (3.54), an unbiased estimator of σ_v^2 is represented by

$$\hat{\sigma}_v^2 = \frac{1}{N-1}\left\{\sum_{i=0}^{N-1}\left[\alpha(n-i) - \hat{\bar{\alpha}}(n)\right]^2 - \frac{N-1}{N}\mathbf{H}\mathbf{P}(n-i \mid n-i-1)\mathbf{H}^T\right\} \tag{3.55}$$

If the measurement noise is a colored noise process, the estimation is derived under the assumption that the colored noise is wide sense stationary. From (3.33), the colored noise vector is represented by:

$$\mathbf{v}(n) = \mathbf{y}(n) - \mathbf{x}(n) \tag{3.56}$$

The true state vector $\mathbf{x}(n)$ is unknown, so $\mathbf{v}(n)$ cannot be determined. To estimate the mean vector $\bar{\mathbf{v}}$, the approximation of $\mathbf{x}(n)$ is used:

$$\boldsymbol{\alpha}(n) = \mathbf{y}(n) - \hat{\mathbf{x}}(n \mid n-1) \tag{3.57}$$

Based on the last N measurements $\boldsymbol{\alpha}(n), \boldsymbol{\alpha}(n-1), \cdots, \boldsymbol{\alpha}(n-N+1)$, the mean vector $\bar{\boldsymbol{\alpha}}$ is estimated by taking the sample mean as:

$$\hat{\bar{\boldsymbol{\alpha}}}(n) = \frac{1}{N}\sum_{i=0}^{N-1}\boldsymbol{\alpha}(n-i) \tag{3.58}$$

The estimation of the mean vector of the measurement noise $\mathbf{v}(n)$ is as follows:

$$\hat{\bar{\mathbf{v}}}(n) = \hat{\bar{\boldsymbol{\alpha}}}(n) \tag{3.59}$$

To estimate the covariance matrix, **R**, the innovation process $\mathbf{e}(n)$ can approximate the noise process as (3.39) shows. From (3.33) and (3.39), $\mathbf{e}(n)$ can be written as:

$$\mathbf{e}(n) = \mathbf{H}\mathbf{x}(n) - \hat{\mathbf{x}}(n \mid n-1) + \mathbf{v}(n) = \tilde{\mathbf{x}}(n \mid n-1) + \mathbf{v}(n) \tag{3.60}$$

where $\tilde{\mathbf{x}}(n \mid n-1)$ is the *a priori* error vector, **H** is the identity matrix for the colored noise case, and its covariance matrix is $\mathbf{P}(n \mid n-1)$.

The covariance matrix of $\mathbf{e}(n)$ is represented by:

$$\mathbf{C}_e(n) = \mathbf{P}(n \mid n-1) + \mathbf{R} \tag{3.61}$$

where $\mathbf{v}(n)$ and $\tilde{\mathbf{x}}(n \mid n-1)$ are assumed uncorrelated and **R** is the covariance matrix of the measured colored noise.

The current unbiased estimate of $\mathbf{C}_e(n)$ is computed as follows:

$$\hat{\mathbf{C}}_e(n) = \frac{n-1}{n}\hat{\mathbf{C}}_e(n-1) + \frac{1}{n}[\mathbf{e}(n) - \bar{e}(n)]\cdot[\mathbf{e}(n) - \bar{e}(n)]^T \tag{3.62}$$

where $\bar{e}(n)$ is the mean value of the innovation process calculated from all past values of $\mathbf{e}(n)$, $\mathbf{e}(n\text{-}1)$, ..., $\mathbf{e}(1)$. $\hat{\mathbf{C}}_e(n)$ is actually the mean value of $\mathbf{C}_e(n), \cdots, \mathbf{C}_e(1)$. From (3.61), the mean value of $\mathbf{C}_e(n)$ can also be computed by the following equation:

$$\bar{\mathbf{C}}_e(n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(i \mid i-1) + \bar{\mathbf{R}} \tag{3.63}$$

where $\bar{\mathbf{R}}$ is the mean of $\mathbf{R}$ (or the unbiased estimate of $\mathbf{R}$). Matching (3.62) and (3.63), the unbiased estimate of $\mathbf{R}$ is represented as:

$$\bar{\mathbf{R}} = \hat{\mathbf{C}}_e(n) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(i \mid i-1) \tag{3.64}$$

3.3.2 Estimation of Model Noise Process Statistics

Assume that the mean and variance of the model noise $u(n)$ are constant over N samples, $u(n), u(n-1), \cdots, u(n-N+1)$. Using the state propagation equation (3.3), the samples of the model noise or driving process are given by the following equation:

$$u(n) = \mathbf{H}[\mathbf{x}(n) - \mathbf{F}\mathbf{x}(n-1)] \tag{3.65}$$

The true state vectors $\mathbf{x}(n)$ and $\mathbf{x}(n-1)$ are unknown, so $u(n)$ cannot be determined, but the approximation,

$$\beta(n) = \mathbf{H}[\hat{\mathbf{x}}(n \mid n) - \hat{\mathbf{x}}(n \mid n-1)] \tag{3.66},$$

can be used [Mye76]. The samples $\beta(n)$ are assumed to be representative of $u(n)$ and can be considered independent and identically distributed [Mye76]. Based on the last N measurements the mean $\bar{\beta}(n)$ and the variance $\sigma_\beta^2(n)$ are estimated [Lea81]. An unbiased estimator for $\bar{\beta}(n)$ is taken as the sample mean

$$\hat{\bar{\beta}}(n) = \frac{1}{N}\sum_{i=0}^{N-1}\beta(n-i) \tag{3.67}$$

and an unbiased estimator for $\sigma_\beta^2(n)$ is obtained by

$$\hat{\sigma}_\beta^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}\left[\beta(n-i)-\hat{\bar{\beta}}(n)\right]^2 \tag{3.68}$$

The estimation of the driving process mean is

$$\hat{\bar{u}}(n) = \hat{\bar{\beta}}(n) \tag{3.69}$$

The expected value of $\sigma_\beta^2(n)$ is

$$E\left\{\sigma_\beta^2(n)\right\} = \frac{1}{N}\sum_{i=0}^{N-1}E\left\{\left[\beta(n-i)\right]^2\right\} - \bar{\beta}^2(n) \tag{3.70}$$

Then expand $E\left\{\left[\beta(n-i)\right]^2\right\}$ in terms of σ_u^2. $\beta(n)$ can be written in terms of the filtered state-error vectors as follows:

$$\beta(n) = -\mathbf{H}\tilde{\mathbf{x}}(n\,|\,n) + \mathbf{HF}\tilde{\mathbf{x}}(n-1\,|\,n-1) + u(n) - \bar{u} \tag{3.71}$$

Rearrange the above equation, we can obtain

$$\beta(n) + \mathbf{H}\tilde{\mathbf{x}}(n\,|\,n) = \mathbf{HF}\tilde{\mathbf{x}}(n-1\,|\,n-1) + u(n) - \bar{u} \tag{3.72}$$

The variance of the equation is

$$E\left\{\left[\beta(n) + \mathbf{H}\tilde{\mathbf{x}}(n\,|\,n)\right]^2\right\} - \bar{\beta}^2(n) = \mathbf{HFP}(n-1\,|\,n-1)\mathbf{F}^T\mathbf{H}^T + \sigma_u^2 \tag{3.73}$$

where the mean value of $\tilde{\mathbf{x}}(n\,|\,n)$ is assumed to be zero.

Now express $E\left\{\left[\beta(n) + \mathbf{H}\tilde{\mathbf{x}}(n\,|\,n)\right]^2\right\}$ in terms of $E\left\{\left[\beta(n)\right]^2\right\}$ and of other computed terms in the Kalman filter

$$E\left\{\left[\beta(n) + \mathbf{H}\tilde{\mathbf{x}}(n\,|\,n)\right]^2\right\} = E\left\{\left[\beta(n)\right]^2\right\} + 2E\left\{\beta(n)\tilde{\mathbf{x}}\ (n\,|\,n)\mathbf{H}^T\right\} + \mathbf{HP}(n\,|\,n)\mathbf{H}^T \tag{3.74}$$

The filtered state-error vector can be rewritten as

$$\tilde{\mathbf{x}}(n\,|\,n) = \left[\mathbf{I} - \mathbf{K}(n)\mathbf{H}\right]\tilde{\mathbf{x}}(n\,|\,n-1) - \mathbf{K}(n)\left[v(n) - \bar{v}\right] \tag{3.75}$$

and the second item in (3.74) is

$$E\left\{\beta(n)\mathbf{x}^T(n\,|\,n)\mathbf{H}^T\right\} = -\mathbf{HP}(n\,|\,n)\mathbf{H}^T + \mathbf{HP}(n\,|\,n-1)\times\left[\mathbf{I} - \mathbf{K}(n)\mathbf{H}\right]^T\mathbf{H}^T \tag{3.76}$$

Combining equation (3.73), (3.74) and (3.76), the resulting expression for $E\{[\beta(n)]^2\}$ is

$$E\{[\beta(n)]^2\} = \mathbf{HFP}(n-1|n-1)\mathbf{F}^T\mathbf{H}^T + \mathbf{HP}(n|n)\mathbf{H}^T - 2\mathbf{HP}(n|n-1)[\mathbf{I}-\mathbf{K}(n)\mathbf{H}]^T\mathbf{H}^T + \sigma_u^2 + \bar{\beta}^2(n) \tag{3.77}$$

Using equation (3.68), (3.70) and (3.77), an unbiased estimator of $\sigma_u^2(n)$ is given by

$$\hat{\sigma}_u^2(n) = \frac{1}{N-1}\left\{\sum_{i=0}^{N-1}\left[\beta(n-i)-\hat{\bar{\beta}}(n)\right]^2 - \frac{N-1}{N}\mathbf{HFP}(n-i-1|n-i-1)\mathbf{F}^T\mathbf{H}^T - \frac{N-1}{N}\mathbf{HP}(n-i|n-i)\mathbf{H}^T + 2\frac{N-1}{N}\mathbf{HP}(n-i|n-i-1)\times[\mathbf{I}-\mathbf{K}(n)\mathbf{H}]^T\mathbf{H}^T\right\} \tag{3.78}$$

When the measurement noise is a colored noise process, the state-space model is given as in (3.32) and (3.33). Then the estimation of the model noise process statistics is derived as follows. Assume that the model noise has a constant mean and variance over N samples of $u(n), u(n-1), \cdots, u(n-N+1)$ [Gab01]. From (3.32), the model noise process is given by the following equation:

$$u(n) = \mathbf{G}^T[\mathbf{x}(n) - \mathbf{F}\mathbf{x}(n-1)] \tag{3.79}$$

Although $\mathbf{x}(n)$ and $\mathbf{x}(n\text{-}1)$ are unknown, $u(n)$ can be represented by its approximation $\beta(n)$ as follows:

$$\beta(n) = \mathbf{G}^T[\mathbf{x}(n|n) - \mathbf{x}(n|n-1)] \tag{3.80}$$

An unbiased estimator for $\bar{\beta}(n)$ is taken as the sample mean

$$\hat{\bar{\beta}}(n) = \frac{1}{N}\sum_{i=0}^{N-1}\beta(n-i) \tag{3.81}$$

The estimation of the driving process mean is

$$\hat{\bar{u}}(n) = \hat{\bar{\beta}}(n) \tag{3.82}$$

Then the unbiased estimate of the variance of $u(n)$ is obtained by:

$$\hat{\sigma}_u^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}\left\{[\beta(n-i)-\hat{\bar{\beta}}(n)]^2 - \frac{N-1}{N}\mathbf{G}^T\mathbf{FP}(n-i-1|n-i-1)\mathbf{F}^T\mathbf{G} - \frac{N-1}{N}\mathbf{G}^T\mathbf{P}(n-i|n-i)\mathbf{G} + 2\frac{N-1}{N}\mathbf{G}^T\mathbf{P}(n-i|n-i-1)[\mathbf{I}-\mathbf{K}(n-i)\mathbf{G}^T]^T\mathbf{G}\right\} \tag{3.83}$$

where $\hat{\bar{\beta}}(n)$ is the mean of the last N measurements $\beta(n), \beta(n-1), \cdots, \beta(n-N+1)$. In the simulation, N is chosen to equal 80 samples.

3.4 Masking Threshold Constrained Kalman Filtering Algorithms

The computation of the masking threshold based on both time domain forward masking effects and frequency domain simultaneous masking properties was introduced in Chapter 2. The final obtained total masking level of the i^{th} critical band ($i = 1, 2, \cdots, 18$) at time t is rewritten below:

$$M_t(i) = \max\{M_s(i), M_t(i)^* \cdot \exp^{-\Delta t/(\tau(i)\cdot Q)}\} \tag{3.84}$$

where $M_t(i)$ and $M_t(i)^*$ are the total masking levels of the current frame and the previous frame, respectively; $M_s(i)$ is the masking level computed from the simultaneous frequency domain masking model [Vir99], Δt is the time difference between two frames, $\tau(i)$ is the maximum decay time constant in each critical band ; and Q is the total loudness level computed as in [Hua02].

Taking the masking threshold of human auditory systems into account, a new Kalman gain can be obtained by modifying the standard Kalman filter to become a perceptually constrained Kalman filter. The new Kalman gain makes the estimate error variance minimum under the constraint that the estimate error energy is smaller than the masking threshold. We derive the constrained Kalman filter as following.

The estimate error covariance matrix $\mathbf{P}(n \mid n)$ can be written as

$$\mathbf{P}(n \mid n) = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix} \tag{3.85}$$

where p is the order of the AR model of the speech signal.

Then the variance of the estimate error can be computed by

$$\Delta_0 = \frac{1}{p}\sum_{i=1}^{p} c_{ii} \tag{3.86}$$

The covariance between the ith and the $(i+j)$th estimate errors is calculated as

$$\Delta_j = \frac{1}{p-j}\sum_{i=1}^{p-j} c_{i,i+j} \quad (j = 1, 2, \cdots, p-1) \tag{3.87}$$

where the mean of the estimate error is assumed to be zero. In the above calculation, the estimate error is assumed to be wide sense stationary. Thus

$$\Delta_j = \Delta_{-j} \quad (j = 1, 2, \cdots, p-1) \tag{3.88}$$

Then a correlation vector of the estimate error is constructed as

$$\boldsymbol{\Delta} = [\Delta_{-p+1}, \Delta_{-p+2}, \cdots, \Delta_{-1}, \Delta_0, \Delta_1, \Delta_2, \cdots, \Delta_{p-1}] \tag{3.89}$$

Thus the power spectrum density of the estimate error can be calculated by the Fourier transform of the above correlation vector. If a 256-point Fourier transform is used, the frequency components of the power spectrum density are written as

$$f_m = \sum_{i=-p+1}^{p-1} \Delta_i e^{-j\frac{2\pi}{N} i \cdot m} = \Delta_0 + 2\sum_{i=1}^{p-1} \Delta_i \cos(\frac{2\pi}{N} \cdot i \cdot m) \quad (m = 0, 1, \cdots, 255, \ N = 256) \tag{3.90}$$

where $f_m \geq 0$.

If the masking thresholds of different frequency components are noted as $T(m)$, to make the error energy smaller than the masking threshold, we can have the following constraints:

$$f_m \leq T(m), \quad m = 0, 1, \cdots, 255 \tag{3.91}$$

The goal is to find a Kalman gain $\mathbf{K}(n)$ which can satisfy the constraints and make the estimate error variance least.

From the derivation of the standard Kalman filtering algorithm, the following equations can be obtained:

(1) For white noise cases,

$$\Delta_0 = \frac{1}{p}\left[\mathrm{Tr}\mathbf{P} - 2*\mathrm{Tr}\mathbf{KHP} + \mathrm{Tr}\mathbf{K}(\mathbf{HPH}^T)\mathbf{K}^T + \mathrm{Tr}\mathbf{K}\mathbf{K}^T\sigma_v^2\right] \tag{3.92}$$

$$\Delta_j = \frac{1}{p-j}\left[\sum_{i=1}^{p-j}\gamma_{i,i+j} - (\mathbf{KHP})_{i,i+j} - (\mathbf{PH}^T\mathbf{K}^T)_{i,i+j} + (\mathbf{KHPH}^T\mathbf{K}^T)_{i,i+j} + (\mathbf{KK}^T\sigma_v^2)_{i,i+j}\right] \tag{3.93}$$

(2) For colored noise cases,

$$\Delta_0 = \frac{1}{p}\left[\mathrm{Tr}\mathbf{P} - 2*\mathrm{Tr}\mathbf{KHP} + \mathrm{Tr}\mathbf{K}(\mathbf{HPH}^T)\mathbf{K}^T + \mathrm{Tr}\mathbf{KRK}^T\right] \tag{3.94}$$

$$\Delta_j = \frac{1}{p-j}\left[\sum_{i=1}^{p-j}\gamma_{i,i+j} - (\mathbf{KHP})_{i,i+j} - (\mathbf{PH}^T\mathbf{K}^T)_{i,i+j} + (\mathbf{KHPH}^T\mathbf{K}^T)_{i,i+j} + (\mathbf{KRK}^T)_{i,i+j}\right] \tag{3.95}$$

where Tr is the trace of a matrix, $\mathbf{P}$ and $\mathbf{K}$ represent $\mathbf{P}(n|n-1)$ and $\mathbf{K}(n)$ respectively, $\gamma_{i,i+j}$ is the element of $\mathbf{P}$ which lies on the ith row and the $(i+j)$th column, $(\mathbf{KHP})_{i,i+j}$, $(\mathbf{PH}^T\mathbf{K}^T)_{i,i+j}$, $(\mathbf{KHPH}^T\mathbf{K}^T)_{i,i+j}$ and $(\mathbf{KK}^T\sigma_v^2)_{i,i+j}$ or $(\mathbf{KRK}^T)_{i,i+j}$ are the elements of the corresponding matrix which are on the ith row and the $(i+j)$th column respectively. In the colored noise case, $\mathbf{H} = \mathbf{I}$, and in the white noise case, $\mathbf{H} = [0,\cdots,0,1]$, a row vector.

Because $\Delta_i\,(i = 0,\cdots, p-1)$ are functions of Kalman gain $\mathbf{K}(n)$, the following optimization problem is constructed:

$$\begin{aligned}&\min \quad \Delta_0 \\ &\text{subject to} \\ &\begin{cases} 0 \le \Delta_0 + 2\sum_{i=1}^{p-1} \Delta_i \cos(\frac{2\pi}{N} \cdot i \cdot m) \le T(m), \quad m = 0,1,\cdots,128 \\ |\Delta_0| \ge |\Delta_j| \\ \Delta_0 > 0 \end{cases}\end{aligned} \tag{3.96}$$

The constraints are established from the masking threshold and the characteristics of power spectrum density and correlation vectors. Now that the Kalman gain is calculated by solving the nonlinear optimization problem (3.96), it cannot be the same as the one expressed in equation (3.9) (white noise cases) or (3.40) (colored noise cases). Thus instead of using equation (3.11) (white noise cases) or (3.42) (colored noise cases), equation (3.28) (white noise cases) or (3.45) (colored noise cases) must be used as the update equation of the estimate error covariance matrix $\mathbf{P}(n \mid n)$.

Thus the constrained Kalman filtering algorithm is shown as follows:

(1) For white noise cases:

$$\begin{aligned}&\mathbf{P}(-1 \mid -1) = Var(\mathbf{x}_{-1}) \\ &\mathbf{P}(n \mid n-1) = \mathbf{F}\mathbf{P}(n-1 \mid n-1)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T\sigma_u^2 \\ &\mathbf{K}(n): \text{ obtained by solving equation (3.96)} \\ &\mathbf{P}(n \mid n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{H}]\mathbf{P}(n \mid n-1)[\mathbf{I} - \mathbf{K}(n)\mathbf{H}]^T + \mathbf{K}(n)\mathbf{K}^T(n)\sigma_v^2 \\ &\hat{\mathbf{x}}(-1 \mid -1) = E(\mathbf{x}_{-1}) \\ &\hat{\mathbf{x}}(n \mid n-1) = \mathbf{F}\hat{\mathbf{x}}(n \mid n) + \mathbf{G}\bar{u} \\ &e(n) = y(n) - \mathbf{H}\hat{\mathbf{x}}(n \mid n-1) - \bar{v} \\ &\hat{\mathbf{x}}(n \mid n) = \hat{\mathbf{x}}(n \mid n-1) + \mathbf{K}(n)e(n)\end{aligned} \tag{3.97}$$

where $\mathbf{x}_{-1}$ is the initial value of the state vector. In our simulation, it is assumed to be a zero vector.

(2) For colored noise cases:

$$
\begin{aligned}
&\mathbf{P}(-1\,|\,-1) = Var(\mathbf{x}_{-1}) \\
&\mathbf{P}(n\,|\,n-1) = \mathbf{F}\mathbf{P}(n-1\,|\,n-1)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T\sigma_u^2 \\
&\mathbf{K}(n)\text{: obtained by solving equation (3.96)} \\
&\mathbf{P}(n\,|\,n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{H}]\mathbf{P}(n\,|\,n-1)[\mathbf{I} - \mathbf{K}(n)\mathbf{H}]^T + \mathbf{K}(n)\mathbf{R}\mathbf{K}^T(n) \\
&\hat{\mathbf{x}}(-1\,|\,-1) = E(\mathbf{x}_{-1}) \\
&\hat{\mathbf{x}}(n\,|\,n-1) = \mathbf{F}\hat{\mathbf{x}}(n\,|\,n) \\
&e(n) = y(n) - \mathbf{H}\hat{\mathbf{x}}(n\,|\,n-1) - \bar{\mathbf{v}} \\
&\hat{\mathbf{x}}(n\,|\,n) = \hat{\mathbf{x}}(n\,|\,n-1) + \mathbf{K}(n)e(n)
\end{aligned}
\tag{3.98}
$$

where $\mathbf{H} = \mathbf{I}$.

Thus the estimate signal is a constrained MMSE estimate of the clean signal.

3.5 Simulation Results

To compare the performance of the proposed algorithms with previously published algorithms, 6 different speech sentences of 5 seconds each spoken by 3 females and 3 males were used in our simulations. The sampling rate was 8 kHz. The noise signals used were white noise, car noise, babble noise, street noise and a simulated colored noise. The speech signals and the white, car, babble and street noise signals were taken from the ITU-T Supplement P.23 Speech Database. The simulated colored noise was obtained by running a white noise record through a multi-bandpass filter. The response of the filter is represented by:

$$
\begin{aligned}
z(n) = &-0.0851z(n-1) + 0.19126z(n-2) + 0.0458z(n-3) + 0.0229z(n-4) \\
&+ 0.1097z(n-5) + 0.1553z(n-6) - 0.132z(n-7) - 0.76z(n-8) + w(n)
\end{aligned}
\tag{3.99}
$$

where $w(n)$ is the white noise with zero mean and variance one.

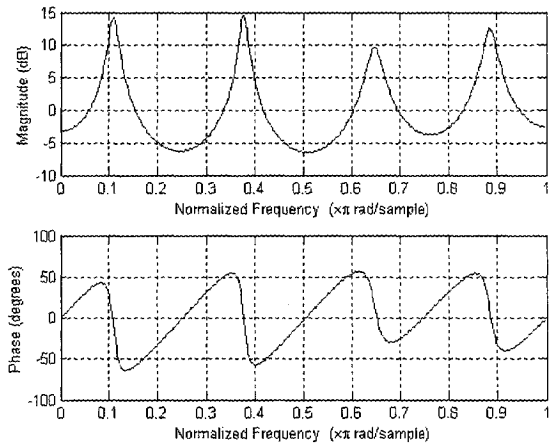


Fig. 3.1 Frequency response of the filter in (3.99)

Figure 3.1 shows the frequency response of the filter. The car noise was similar to pink noise and the street noise was more like a non-stationary pink noise. From common definition, the power density of pink noise decreases 3dB per octave with increasing frequency (density proportional to 1/f) over a finite frequency range. Each octave contains the same amount of power.

The input signals were normalized so that the amplitude is in the interval [-1, 1]. As mentioned earlier the frame size was 80 samples, i.e. 10 ms frames. The AR prediction order p was set to be 10 where it can provide an accurate enough representation with a limited cost of computation, and the AR coefficients were updated every frame by using the Levinson-Durbin algorithm. The data length used to compute the AR parameters (i.e. to compute correlation values) was 160 samples, which includes the current noisy frame and one previous enhanced frame. This frame length was chosen empirically, where it was not too long to represent the non-stationarity of speech signal or too short to represent the short-time stationarity of speech signal. Table 3.1 shows the results of average PESQ scores obtained by different methods for white noise and Table 3.2 shows average PESQ scores for colored noise. The detailed PESQ scores obtained by different methods for different types of noise are shown in

Appendix A with Table A.1 and Table A.2. The method proposed in this chapter was named as Constrained Kalman Filter with Simultaneous and Forward Masking (CKFSFM). In the simulation, the iteration number used for the KLT [Mit00] was fixed to 10 for simplicity. We used Matlab to run the simulation and the CPU was an Intel® Pentium® 4 2.0 GHz. To process a whole 5-second sentence, the average consumed time was 20 minutes for the CKFSFM, 1 minute for the method in [Vir99], 40 sec. for the conventional spectral subtraction method, 4 minutes for the method in [Gab01], 7 minutes for the method in [Pop98] and 6 minutes for the method in [Mit00].

In Appendix A, Fig A.1 shows the average PESQ scores obtained for white noise signals by different methods and different speakers (females and males). Fig. A.2 shows the average PESQ scores obtained for different colored noise signals by different methods and different speakers (females and males).

3.6 Conclusions

Since the Kalman gain makes the error variance minimum and the error energy smaller than the masking threshold, PESQ scores are improved with the proposed method compared to other methods [Gab01, Vir99, Pop98, Mit00]. From the simulation results, it can be observed that speech signals degraded by white noise or colored noise can be enhanced by the proposed method. However the algorithm requires a large amount of computations not only because the matrix inverse is calculated at each time step, but also a nonlinear optimization problem has to be solved in each step. A fast algorithm can be studied in the future for real time application. In addition, whether the solution of the nonlinear optimal problem always exists or not is not proved. In the simulations, at most time steps, a solution could be found. However, there are a few cases where a feasible solution could not be obtained. Further studies about the existence of the solution and the

algorithm for solving a nonlinear optimization problem could be made. The algorithms proposed in this chapter (CKFSFM) provided a theoretical solution of combining the masking threshold with a Kalman filter for speech enhancement, and have shown the great potential of this approach.

Table 3.1 Average PESQ Scores Obtained by Different Methods for White Noise

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM
-5	1.17	0.66	1.27	0.70	1.66
0	1.40	1.02	1.54	1.07	1.96
5	1.64	1.44	1.81	1.47	2.24
10	1.94	1.76	2.16	1.80	2.52
15	2.27	1.97	2.49	2.00	2.79

Table 3.2 Average PESQ Scores Obtained by Different Methods for Colored Noise

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Pop98	Mit00	CKFSFM
-5	1.25	0.79	1.36	0.83	0.83	1.71
0	1.56	1.15	1.68	1.20	1.20	1.97
5	1.84	1.53	1.98	1.57	1.57	2.25
10	2.15	1.81	2.27	1.85	1.85	2.49
15	2.47	2.02	2.57	2.06	2.06	2.75

Chapter 4 Kalman Filter Heuristically Combined with Frequency and Time Domain Masking Properties

4.1 Introduction

The computational cost of the algorithms proposed in Chapter 3 (CKFSFM) was large. The computation complexity for one recursion of the Kalman iteration is $O(6p^3)$ [Gor97], where p is the LPC order of the speech model (colored noise case is considered here). The complexity of the nonlinear optimization problem is exponential to the problem size. So the truly perceptually constrained Kalman filter has a very huge computation cost. To overcome the problem of the large computational load, this chapter proposes heuristic methods combining Kalman filters with frequency and time domain masking properties for speech enhancement in white and colored noise. The heuristic methods can only achieve a sub-optimal result compared with CKFSFM, but the performance is still quite acceptable. We propose first a method that only uses the frequency domain masking property and addresses the problem of white noise. A second heuristic approach using a post-filter which exploits both forward masking properties and simultaneous masking properties is then proposed. For this second approach, both the white noise case and the colored noise case are considered.

4.2 Kalman Filter Using Frequency Domain Masking Properties

In this section, an approach for single channel speech enhancement based on the Kalman filter and masking properties of the human auditory system is introduced [MaN03], which is named as KFSM (Kalman Filter with Simultaneous Masking). A standard time-varying Kalman filtering method is extended by combining the calculation of noise masking thresholds during the process of parameter updating. Simulation results of the traditional spectral subtraction method, an extended spectral subtraction with masking properties, a standard Kalman filtering based method, and the new proposed approach are computed and compared. In the simulations, the speech signals are degraded by white noise The proposed approach has no delay and better Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862), with no Voice Activity Detection (VAD) required. The PESQ score improvement obtained by the proposed method is about 0.3 compared with the original noisy signal. Fig. 4.1 shows the block diagram of the system.

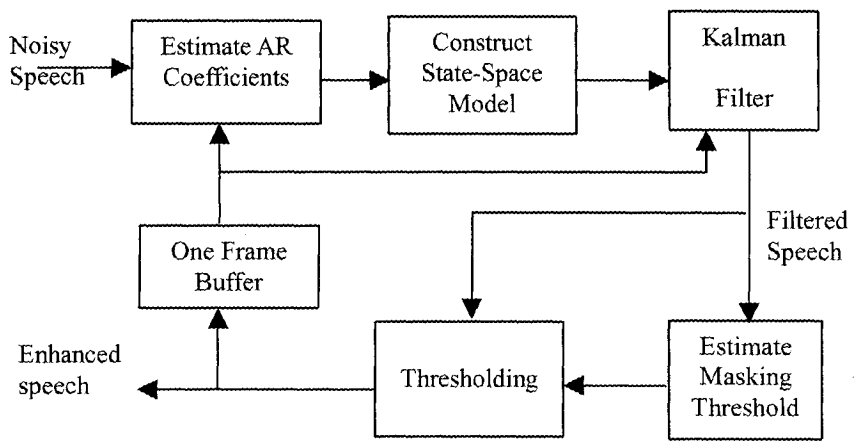


Fig. 4.1 Diagram of Kalman filter with a frequency domain only masking threshold

In Section 2.2.4, the state space model of a Kalman filter for speech enhancement was described. In

this chapter, the same state space model is used. Reference [Gab01] describes the procedure for updating the state-vector. This paper combines this Kalman filter with the masking properties of human auditory systems as follows:

$$e(n) = y(n) - \mathbf{H}\hat{\mathbf{x}}(n \mid n-1) - \bar{v} \tag{4.1}$$

$$\mathbf{K}(n) = \mathbf{P}(n \mid n-1)\mathbf{H}^T \times (\mathbf{H}\mathbf{P}(n \mid n-1)\mathbf{H}^T + \sigma_v^2)^{-1} \tag{4.2}$$

$$\hat{\mathbf{x}}(n \mid n) = \hat{\mathbf{x}}(n \mid n-1) + \mathbf{K}(n)e(n) \tag{4.3}$$

$$\tilde{\hat{\mathbf{x}}}(n \mid n) = \text{masking}(\hat{\mathbf{x}}(n \mid n)) \tag{4.4}$$

$$\mathbf{P}(n \mid n) = (\mathbf{I} - \mathbf{K}(n)\mathbf{H})\mathbf{P}(n \mid n-1) \tag{4.5}$$

$$\hat{\mathbf{x}}(n+1 \mid n) = \mathbf{F}\tilde{\hat{\mathbf{x}}}(n \mid n) + \mathbf{G}\bar{u} \tag{4.6}$$

$$\mathbf{P}(n+1 \mid n) = \mathbf{F}\mathbf{P}(n \mid n)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T\sigma_u^2 \tag{4.7}$$

with:

$e(n)$ innovation sequence;

$\mathbf{K}(n)$ Kalman gain;

$\hat{\mathbf{x}}(n \mid n)$ filtered estimate of the state vector $\mathbf{x}(n)$;

$\tilde{\hat{\mathbf{x}}}(n \mid n)$ masked filtered estimate of the state vector;

$\mathbf{P}(n \mid n) = E[\tilde{\mathbf{x}}(n \mid n)\tilde{\mathbf{x}}^T(n \mid n)]$ filtered state-error correlation matrix, where $\tilde{\mathbf{x}}(n \mid n) = \mathbf{x}(n) - \hat{\mathbf{x}}(n \mid n)$ is the filtered state-error vector;

$\hat{\mathbf{x}}(n \mid n-1)$ minimum mean-square estimate of the state vector $\mathbf{x}(n)$ given the past observations $y(1), \cdots, y(n)$ (i.e. predicted state vector);

$\mathbf{P}(n \mid n-1) = E[\tilde{\mathbf{x}}(n \mid n-1)\tilde{\mathbf{x}}^T(n \mid n-1)]$ predicted state-error correlation matrix, where $\tilde{\mathbf{x}}(n \mid n-1) = \mathbf{x}(n) - \hat{\mathbf{x}}(n \mid n-1)$ is the predicted state-error vector.

The difference between the proposed approach and the standard Kalman filter method in [Gab01] is that the masking properties are applied to $\hat{\mathbf{x}}(n \mid n)$, the filtered estimate of the state vector $\mathbf{x}(n)$, as shown in (4.4). This section only considers frequency domain masking, or simultaneous masking: a weak signal is made inaudible by a stronger signal occurring simultaneously. This property is modeled via a noise masking threshold, below which all components are inaudible. To have a better PESQ score, the filtered state-error vector signal should be inaudible, i.e. its spectrum should be below the masking threshold. This cannot directly be achieved by the method, but by masking some inaudible noise components in the filtered state estimate, the state estimate becomes closer to the real state of the clean speech. As a result, the innovation sequence contains more useful new information of the current clean speech and less model noise, leading to a better tracking of the real state. The procedure for the masking of $\hat{\mathbf{x}}(n \mid n)$ is described in detail in Section 4.4.

The estimated speech signal can be retrieved from the state-vector estimator:

$$\hat{s}(n) = \mathbf{H}\tilde{\hat{\mathbf{x}}}(n \mid n) \tag{4.8}$$

The noise variances σ_u^2 and σ_v^2 are estimated using the same approach as in [Gab01]. Both of these estimations are done at every time step, from previous filtered state-vectors.

The method to estimate the model and measurement noise statistics was described in Chapter 3 and is used in this chapter. The computation of the frequency domain masking threshold of human auditory system is also the same as introduced in Chapter 2. The spread threshold estimate $T(k)$ (where k is the number of critical band) is obtained, and mapped back from the Bark scale to the linear frequency domain. Then a hard decision is used: the FFT components whose energy is greater than the threshold are retained, otherwise they are set to zero. After the hard thresholding, the Kalman speech spectrum is modified. To obtain $\tilde{\hat{\mathbf{x}}}(n \mid n)$, an inverse Discrete Fourier

Transform (DFT) is then computed for p values where p is the AR model order in the **F** matrix (i.e. this is equivalent to computing a full inverse FFT and keeping only the last p values of the 256 time values).

4.2.1 Simulation Results

In this section, the standard Levinson-Durbin equations to estimate the AR coefficients were used, and the mean and the variance of $u(n)$ and $v(n)$ were estimated with the method from [Gab01].

Six different speech sentences of 4 seconds spoken by 3 females and 3 males were used in our simulations. In this section, only white noise is considered. All of the speech files and noise files were taken from the ITU-T Supplement P.23 speech database. The sampling frequency used was 8000 Hz, and the input signals were normalized so that the amplitude is in the interval [-1, 1]. As mentioned earlier the frame size was 80 samples, i.e. 10 ms frames. The AR prediction order p was set to be 10, and the AR coefficients were updated every frame. The data length used to compute the AR parameters (i.e. to compute correlation values) was 160 samples, which includes the current noisy frame and one previous enhanced frames. A Hanning window is used in the calculation of the AR coefficients. The average PESQ scores obtained by different methods are shown by Table 4.1. The detailed PESQ scores are shown in Appendix B with Table B.1

In Appendix B, Fig. B.1, the curves of the average PESQ scores of different methods computed from the data in Table B.1 are shown. Fig. B.1 (a) shows the PESQ scores of female speech signals and Fig. B.1 (b) shows those of male speech signals. As the results show, the proposed approach always produced the best PESQ scores in the simulations, except when compared to the algorithm previously introduced in Chapter 3 (which has a significantly higher complexity).

Table 4.1 Average PESQ scores obtained by different methods with white noise signal

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	1.17	0.66	1.27	0.70	1.66	1.48
0	1.40	1.02	1.54	1.07	1.96	1.77
5	1.64	1.44	1.81	1.47	2.24	2.07
10	1.94	1.76	2.16	1.80	2.52	2.34
15	2.27	1.97	2.49	2.00	2.79	2.60

4.3 Post-Filters Based on Frequency Domain and Time Domain Masking Properties

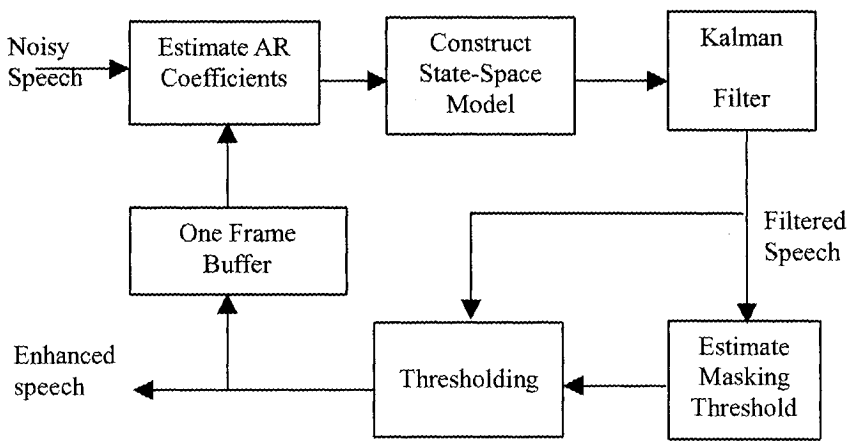


Fig. 4.2 Diagram of the Kalman filter with the perceptual post-filter

The simulation results obtained in the previous section show that the heuristic combination of a Kalman filter and masking properties is promising and feasible. Since the masking properties include both frequency domain and time domain, this section proposes two new perceptual post-filters based on these two domains and runs on a frame by frame basis. Fig. 4.2 shows the

diagram of this method.

As introduced in Chapter 2, if both forward time domain and frequency domain masking properties are considered, the total masking levels are determined by integrating the frequency domain simultaneous masking effects and the time domain forward masking effect of the previous section. The level depends not only on the current frame but also on the previous frames.

The total masking level of the i^{th} critical band at time t is the maximum of the current simultaneous masking level and the total masking level of the previous frame decayed by some constant:

$$M_t(i) = \max\{M_s(i), M_t(i)^* \cdot \exp^{-\Delta t/(\tau(i)\cdot Q)}\} \tag{4.9}$$

where $M_t(i)$ and $M_t(i)^*$ are the total masking levels of the current frame and the previous frame, respectively; $M_s(i)$ is the masking level computed from the simultaneous frequency domain masking model [Joh88], Δt is the time difference between two frames, $\tau(i)$ is the maximum decay time constant in each critical band [Jes82]; and Q is the total loudness level (as computed in Section 2.2.6.1) which is now normalized by the total loudness of a 60dB uniform masking noise (UMN) [Hua02]. If Q is larger than one, then Q is set to one. So Q lies between zero and one. When Q is zero, there is no forward masking. When Q is one, the total masking level decays with a maximum time constant.

4.3.1 First Approach to Combine a Kalman Filter with a Perceptual Post-Filter

Assume that the clean speech signal $s(n)$ is modeled as a p-th order AR process as follows:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{4.10}$$

The measured speech signal $y(n)$ is

$$y(n) = s(n) + v(n) \tag{4.11}$$

where a_i is the i-th AR coefficient. If the residual signal $u(n)$ and the measurement noise $v(n)$ are uncorrelated Gaussian white noise processes with means $\bar{u}$ and $\bar{v}$ and variances σ_u^2 and σ_v^2, then the state space model was described in Section 2.2.4, by equations (2.28) to (2.32). If $\bar{u}$, $\bar{v}$, σ_u^2, σ_v^2 and the transition matrix **F** in equation (2.28) are known, the standard Kalman filter can be used as shown by equations (2.33) to (2.38). The Kalman filtered speech signal can be retrieved from the state-vector estimator:

$$\hat{s}(n) = \mathbf{H}\hat{\mathbf{x}}(n \mid n) \tag{4.12}$$

In the post-filter, the Kalman filtered speech signals are processed on frame-by-frame basis. After each noisy speech frame is passed through the Kalman filter, the whole frame of the filtered signal $\hat{\mathbf{s}}(n) = [\hat{s}(n-L+1) \cdots \hat{s}(n)]$ is input into the post-filter, where L is the frame length. Both time domain forward masking effects and frequency domain simultaneous masking properties are considered in the proposed system. The total masking level is defined in equation (4.9).

The post-filter performs thresholding on the input signal based on the computed total masking level $M_t(i)$. To perform the thresholding, the following procedure is used:

(1) Mapping the total masking level $M_t(i)$ in each critical band to frequency domain (FFT bins) to obtain $T(\omega_i)$ ($i = 0,1, \cdots 255$).

(2) Performing the thresholding on the speech spectrum:

$$\left|\tilde{\tilde{S}}(\omega_i)\right| = \begin{cases} \left|\hat{S}(\omega_i)\right| \times \alpha^{c(i)/9} & \text{if } \dfrac{\left|\hat{S}(\omega_i)\right|^2}{L} < T(\omega_i) \\ \left|\hat{S}(\omega_i)\right| \times \alpha^{(c(i)-1)/9} & \text{otherwise} \end{cases} \tag{4.13}$$

where $i = 0,2,\cdots 255$, $c(i)$ is the critical band number corresponding to the frequency bin i, and α is the tonality coefficient computed with the simultaneous masking threshold.

(3) Doing an IFFT using $\left|\tilde{\tilde{S}}(\omega_i)\right|$ and the phase of $\hat{S}(\omega_i)$, keeping the last 80 values of the size-256 IFFT outputs to obtain the improved frame of speech signal $\tilde{\mathbf{s}}(n) = [\tilde{s}(n-79), \tilde{s}(n-78), \cdots, \tilde{s}(n)]$.

The reason to choose Eq. (4.13) as the thresholding procedure instead of a hard threshold is as follows. Since the filtered signal is only a MMSE estimate of the clean speech signal, the calculated masking threshold is only an approximation of the real masking threshold. To decrease the noise effect, the above masking procedure is used, where the characteristics of the speech frame are taken into account. If the frame is noise-like, most of the frame will be masked. If it is not noise-like, the masked result will keep most of the energy of the original frame. In addition, in the new masking procedure the power of the tonality coefficient is increased as the frequency increases. Thus the high frequency components are masked much more than the low frequency components. This is because high frequency components are typically more noise-like.

4.3.2 Second Approach to Combine a Kalman Filter with a Perceptual Post-Filter

This section will present a second approach to combine the Kalman filter with a post-filter. It is presented for the colored noise case, although it can obviously be simplified to the white noise case. In a colored noise environment, a vector Kalman filter is used. The state space model and

the Kalman filtering algorithm for this case is described in Chapter 3. The last element of $\hat{\mathbf{x}}(n \mid n)$, $\hat{s}(n)$, is the output of the Kalman filter.

The post-filter performs thresholding on the input signal based on the computed total masking level $M_t(i)$. To perform the thresholding, the following procedure is used:

(1) Mapping the total masking level $M_t(i)$ in each critical band to frequency domain (FFT bins) to obtain $T(\omega_j)$ ($j = 0, 1, \cdots, 255$).

(2) From the last 256 computed filtered state error covariance matrix, $\mathbf{P}(n-i \mid n-i)$ ($i = 0, \cdots, N-1$), estimate the power spectrum density (PSD) of the filtered state error. At first, obtain an average covariance matrix $\bar{\mathbf{P}}$ from $\mathbf{P}(n-i \mid n-i)$ ($i = 0, \cdots, N-1$) as follows:

$$\bar{\mathbf{P}} = \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{P}(n-i \mid n-i) \tag{4.14}$$

If the element of $\bar{\mathbf{P}}$ is represented by $\bar{p}_{k,l}$ ($k,l = 1,2,\cdots,p$), where p is the LPC order.

Then the variance of the state error, σ^2, is estimated by

$$\sigma^2 = \frac{1}{p} \sum_{k=1}^{p} \bar{p}_{k,k} \tag{4.15},$$

and the covariance of the state error, c_j (for $j = 1,2,\cdots,p-1$), is estimated by

$$c_j = \frac{1}{p-j} \sum_{m=1}^{p-j} \bar{p}_{m,m+j} \tag{4.16}.$$

Then compute the FFT of the covariance vector of the filtered state error. The PSD, $P_e(\omega_i)$ is estimated by using 256 frequency points, i.e., $\omega_i = \frac{2\pi}{256} \cdot i$, ($i = 0, \cdots, 255$).

(3) Then performing the thresholding on the speech spectrum:

$$\left|\tilde{\hat{S}}(\omega_i)\right| = \begin{cases} \left|\hat{S}(\omega_i)\right| \times (\alpha^{P_e(\omega_i)/T(\omega_i)} + \overline{SNR}^2) & \text{if } P_e(\omega_i) < T(\omega_i) \\ \left|\hat{S}(\omega_i)\right| \times (\alpha \times (1 + \alpha^{P_e(\omega_i)/T(\omega_i)}) + \overline{SNR}^2) & \text{otherwise} \end{cases} \tag{4.17}$$

where $i = 0,1,\cdots 255$, and α is the tonality coefficient.

(4) Doing an IFFT using $\left|\tilde{\hat{S}}(\omega_i)\right|$ and the phase of $\hat{S}(\omega_i)$, keeping the last 80 values of the size-256 IFFT outputs to obtain the improved frame of speech signal $\tilde{\mathbf{s}}(n)=[\tilde{s}(n-79),\tilde{s}(n-78),\cdots,\tilde{s}(n)]$.

When the post-filter is designed, a tradeoff is made between signal distortion and residual noise. Thus the estimated signal to noise ratio of each frame and each critical band is used. Both the masking properties and the characteristics of the speech frame are taken into account in the thresholding procedure (4.17). Consider the first item in the parentheses in (4.17). The smaller the α is, *i.e.*, the speech frame is more like a noise, the more the Kalman filtered signal power is reduced. When the error power density $P_e(\omega_i)$ is smaller than the masking threshold $T(\omega_i)$, most of the power of the Kalman filtered signal can be kept to reduce distortion. The smaller the $P_e(\omega_i)$ is, the more the Kalman filtered signal energy is kept. On the other hand, if $P_e(\omega_i)$ is larger than $T(\omega_i)$, whether to increase the power of the Kalman filtered signal or to reduce it depends on the value of α and $P_e(\omega_i)$. From (4.17), if $P_e(\omega_i)>T(\omega_i)$ and $\alpha(1+\alpha^{P_e(\omega_i)/T(\omega_i)})>1$, then the following two solutions can be derived, $\alpha>\frac{\sqrt{5}-1}{2}$ and $P_e(\omega_i)<\frac{T(\omega_i)\ln(1/\alpha-1)}{\ln\alpha}$. This case represents a speech frame which is like a tone and the estimate error is not too large. To mask the estimate of the error, the power of the Kalman filtered signal is increased. In other cases, the power of the Kalman filtered signal is reduced. If $\alpha>\frac{\sqrt{5}-1}{2}$ and $P_e(\omega_i)>\frac{T\ln(1/\alpha-1)}{\ln\alpha}$, the estimate error is too large. If the signal power is increased, it may incur large distortion. Thus in this case, the signal power is reduced. Then the larger the $P_e(\omega_i)$ is, the more the signal power is reduced. If $\alpha<\frac{\sqrt{5}-1}{2}$, the speech frame is assumed to be more like a noise frame. The larger the $P_e(\omega_i)$ is, the more the signal power is reduced. Furthermore, if $\overline{SNR}$ of a frame is high, *i.e.*, the signal to noise ratio of the frame is high, the Kalman filtered signals of that frame can be said to have more clean speech signal. Then

the signal power is increased more.

The model and measurement noise statistics computation is based on a covariance matching method and was introduced in Chapter 3.

4.4 Simulation Results and Conclusions

As in Chapter 3, six different speech sentences were used in the simulations. In this chapter, the noise signals used were white noise, car noise, babble noise, street noise, and a colored noise generated by running a white noise signal $w(n)$ through a 8th order AR filter as in (3.99). All of the speech files and noise files were taken from the ITU-T Supplement P.23 speech database. The sampling frequency used was 8 kHz. The AR coefficients were updated every frame. The data length used to compute the AR parameters was 160 samples, which includes the current noisy frame and one previous enhanced frames. Table 4.2 shows the results of average PESQ scores obtained by different methods for white noise and Table 4.3 shows average PESQ scores for colored noise. The detailed PESQ scores obtained by the different methods for different types of noise are shown in Appendix B with Table B.2 and Table B.3. KFSFMW (Kalman Filter with Simultaneous and Forward Masking for White noise) was used to represent the first approach proposed in Section 4.3.1 and KFSFMC (Kalman Filter with Simultaneous and Forward Masking for Colored noise) to represent the method proposed in Section 4.3.2. We used Matlab to run the simulation and the CPU was an Intel® Pentium® 4 2.0 GHz. To process a whole 5-second sentence, the average consumed time was approximately 15 minutes for the KFSM, 10 minutes for the KFSFMC, 9 minutes for the KFSFMW.

In Appendix B, Fig B.2 shows the average PESQ scores obtained for white noise signals by

different methods and different speakers (females and males). Fig. B.3 shows the average PESQ scores obtained for different colored noise signals by different methods and different speakers (females and males).

The algorithms tested in this section are the truly perceptually constrained Kalman filter proposed in Chapter 3 (CKFSFM), the Kalman filter with perceptual post-filter proposed in this section, the conventional spectral subtraction method, the modified spectral subtraction method with masking properties [Vir99], the Kalman filter for colored noise [Pop98], and the KLT approach for colored noise [Mit00]. The input SNR of the noisy signal is from -5dB to 15dB. It can be observed that the proposed truly constrained Kalman filtering method has again produced the highest PESQ improvement. The average PESQ scores gain obtained by this method is 0.43. The results also show that the proposed heuristic (the time domain Kalman filter with a masking threshold post-filter) also produces a good improvement of the noisy signals. The average PESQ scores gain obtained by the method is 0.37. The performance of the heuristic method is thus fairly close to that of the truly constrained Kalman filter method. The average PESQ score gain obtained by the perceptual spectral subtraction method [Vir99] is 0.15, while the average PESQ score gain produced by the conventional spectral subtraction method is –0.38. As for the methods in [Pop98] and [Mit00], the PESQ gain is 0.08 in both cases. In general, the improvement obtained by the proposed perceptual Kalman filter methods over previously published methods decreased with an input SNR increase. The proposed Kalman filter with post-filter method was also applied to codec systems as a preprocessing or a post-processing step. In Appendix B, Table B.4 shows the average PESQ scores gains obtained by that method when it is used as a preprocessor or a post-processor. It can be observed that the gains produced for coded speech are similar to those for the original speech (Table B.3). Table B.5 shows the detailed results obtained by the perceptual spectral subtraction method [Vir99] and the Kalman filter with a perceptual post-filter, for different types of noise. It can be observed that the proposed algorithm provides a better performance for all the noise types considered, that the difference in performance becomes smaller as the input SNR increases, and that the case of white noise is the case where the proposed algorithm can produce the largest PESQ gain. Table B.6 then shows the results obtained under a reverberation environment for the same two algorithms. The environment simulates a medium auditorium and the reverberation parameters were produced by Cool Edit Pro^{TM}.

Comparing Table B.4 to Table B.6, it can be observed that the considered speech enhancement methods perform much better for additive noise than for reverberation conditions, which is normal considering that the assumption of additive noise is found in the development of those algorithms. Yet the proposed Kalman filter with perceptual post-filter can slightly outperform the algorithm in [Vir99] under this reverberation condition. The simulation results thus have shown that the proposed methods are promising and can be applied to different speech communication systems.

Table 4.2 Average PESQ scores obtained by different methods with white noise signal

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	1.17	0.66	1.27	0.70	1.66	1.48	1.61
0	1.40	1.02	1.54	1.07	1.96	1.77	1.93
5	1.64	1.44	1.81	1.47	2.24	2.07	2.21
10	1.94	1.76	2.16	1.80	2.52	2.34	2.48
15	2.27	1.97	2.49	2.00	2.79	2.60	2.74

Table 4.3 Average PESQ scores obtained by different methods with colored noise signal

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.25	0.79	1.36	0.83	0.83	1.71	1.66
0	1.56	1.15	1.68	1.20	1.20	1.97	1.94
5	1.84	1.53	1.98	1.57	1.57	2.25	2.20
10	2.15	1.81	2.27	1.85	1.85	2.49	2.45
15	2.47	2.02	2.57	2.06	2.06	2.75	2.71

Chapter 5 A Wavelet Kalman Filter with Perceptual Masking for Speech Enhancement in Colored Noise

5.1 Introduction

In Chapter 4, a Kalman filter approach concatenated with a perceptual post-filter for speech enhancement was introduced as a simpler and heuristic alternative of the truly constrained Kalman filter proposed in Chapter 3 (CKFSFM). For real-time application, the computational cost is an important issue. Since the perceptual post-filter is based on human auditory masking properties and works in the frequency domain, whereas the Kalman filter is run in the time domain, the method proposed in Chapter 4 (KFSFMC) requires the switching between the two domains and thus the computation load is increased. To solve the problem, this chapter presents a wavelet domain Kalman filter for speech enhancement and maps the perceptual post-filter also into the wavelet domain. The new algorithm is named as WKFSFM (Wavelet Kalman Filter with Simultaneous and Forward Masking). Thus both the Kalman filtering procedure and the post-filter procedure can be done in the same domain. Then the computational load is less than that for the time domain Kalman filter of Chapter 4 (KFSFMC). The computation complexity for one recursion of the Kalman iteration is $O(6p^3)$, where p is the LPC order of the speech order. The complexity of the nonlinear optimization problem is exponential to the problem size. So the truly perceptually constrained Kalman filter of Chapter 3 (CKFSFM) has a high computation cost. For the perceptual post-filter introduced in Chapter 4 (KFSFMC), the complexity of the computation cost of the masking threshold is about $O(2N)+2O(N\log N)$ (N is the FFT size). However for the method proposed in this chapter (WKFSFM), the complexity of the whole method is $O(6p_b^3)+(5/8)*O(8p_b)+(1/8)*O(16p_b)$, where p_b is the size of a block data. If we set $p=10$, $N=256$, $p_b=8$, then the computation cost for KFSFMC is $O(8300)$, whereas for the method in

this chapter (WKFSFM) it is $O(3128)$, *i.e.*, the saving of the computation load is significant. In the rest of this chapter, Section 5.2 briefly introduces the wavelet Kalman filter for speech enhancement, Section 5.3 describes the perceptual post-filter, Section 5.4 shows the simulation results and Section 5.5 presents a brief conclusion.

5.2 Wavelet Kalman Filter for Speech Enhancement

For the problem of enhancing speech degraded by colored noise, a wavelet packet transform is used to construct the state-space model in the wavelet domain. The wavelet Kalman filter can get the MMSE or the LMMSE estimate of the speech spectrum in the wavelet domain.

5.2.1 Wavelet Packet Transform and Filter Banks

For a given sequence of signals $x(i,n)$ at time n and for a fixed resolution level i, the lower resolution signal $x_L(i-1,n)$ is obtained by downsampling the output of a halfband lowpass filter by two, where the impulse response of the filter is $h(n)$:

$$x_L(i-1,n) = \sum_{k=-\infty}^{\infty} h(2n-k)x(i,k) \tag{5.1}$$

The wavelet coefficients, as a complement of $x_L(i-1,n)$, are denoted by $x_H(i-1,n)$, and can be computed by first using a highpass filter with an impulse response $g(n)$ and then by downsampling the output of the highpass filter by two. This yields:

$$x_H(i-1,n) = \sum_{k=-\infty}^{\infty} g(2n-k)x(i,k) \tag{5.2}$$

with $g(n) = (-1)^n h(L-1-n)$, where L is the length of the filter and must be even. Equations (5.1) and (5.2) define the discrete wavelet transform. In this chapter, the filter impulse responses form an orthonormal set. Thus the reconstruction of the original signal $x(i,n)$ is computed by

$$x(i,n) = \sum_{k=-\infty}^{\infty} h(2k-n)x_L(i-1,k) + \sum_{k=-\infty}^{\infty} g(2k-n)x_H(i-1,k) \tag{5.3}$$

The discrete wavelet transform can be implemented by an octave-band filter bank [Chu99]. Fig. 5.1 shows the pyramid structure of the discrete wavelet transform.

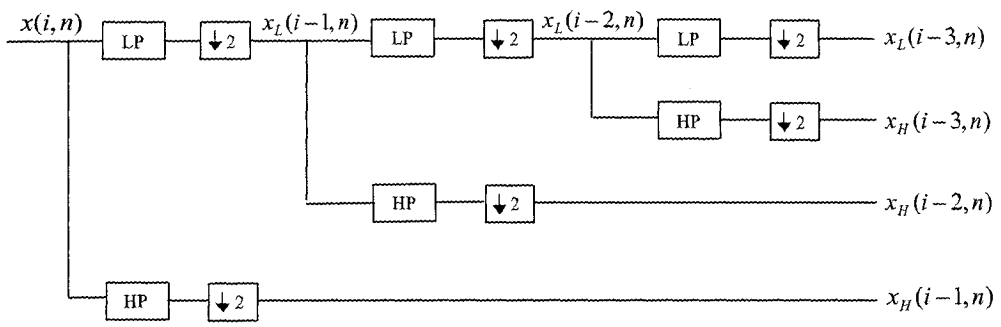


Fig. 5.1 Pyramid structure of the Wavelet Transform

However, the wavelet transform only decomposes the output of the lowpass filter at each resolution level. From Fig. 2.5, the frequency response of human auditory system has a multi-resolution property. Also from that figure, we can find that it is difficult to map the frequency response of the human auditory system to the wavelet transform domain because of the lack of the mapping in lower frequency by wavelet transform, whereas, by using a wavelet packet transform the mapping problem can be solved. By using a wavelet packet transform, both the output of the lowpass filter and the highpass filter can be further processed. A wavelet packet tree can be used to represent the operation [Chu99]. For this chapter, it is more convenient to describe the wavelet packet transform in an operator form. For simplicity, the wavelet packet transform is denoted by a matrix **T**. An orthogonal wavelet basis is used, so the inverse of matrix **T** equals its transpose, *i.e.*, $\mathbf{T}^{-1} = \mathbf{T}'$. In the thesis, a full binary wavelet packet decomposition is used.

5.2.2 State-Space Model

The speech signal is processed on a block by block basis, and the length of each (small) block equals the LPC order p. Since the length of the data to be transformed by wavelet packets usually has the form 2^n, for simplicity in the simulations the LPC order p is set to be 8 ($=2^3$) and the frame length F is set to be 128 ($=2^7$). The signal block index is denoted by N, and the frame index is denoted by w. A clean speech signal $s(n)$ can be described as:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{5.4}$$

where p is the LPC order, $u(n)$ is a zero mean, white Gaussian process with variance σ_u^2, and a_i is the i-th AR parameter. The n-th sample of the noisy speech signal $y(n)$ is described as:

$$y(n) = s(n) + v(n) \tag{5.5}$$

where $v(n)$ is a colored measurement noise process. The *N-th* block of data is written as:

$$\mathbf{S}_N = [s((N-1)\cdot p+1), s((N-1)\cdot p+2), \cdots, s((N-1)\cdot p+p)]' \tag{5.6},$$

where notation $'$ represents matrix transposition. From (5.5) and (5.6), the relationship of $\mathbf{S}_N$ and $\mathbf{S}_{N+1}$ is as follows:

$$\mathbf{S}_{N+1} = \mathbf{F}\mathbf{S}_N + \mathbf{G}\mathbf{U}_{N+1} \tag{5.7}$$

where

$$\mathbf{U}_N = [u((N-1)\cdot p+1), u((N-1)\cdot p+2), \cdots, u((N-1)\cdot p+p)]'$$

and the transition matrix $\mathbf{F}$ is computed by the following steps. First, a matrix $\mathbf{F}_1$ is written as:

$$\mathbf{F}_1 = \begin{bmatrix} a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \\ 0 & a_p & a_{p-1} & \cdots & a_2 \\ 0 & 0 & a_p & \cdots & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_p \end{bmatrix} \tag{5.8}$$

The first row of the matrix $\mathbf{F}$ equals the first row of the matrix $\mathbf{F}_1$. From the second row to the p-th row, the elements of the i-th row of matrix $\mathbf{F}$ are computed by:

$$(\mathbf{F})_i = \sum_{k=0}^{i-1} (\mathbf{F}_1)_{i-k} \cdot a_k \ , \tag{5.9}$$

where $a_0 = 1$, $(\mathbf{F})_i$ and $(\mathbf{F}_1)_{i-k}$ represent the i-th row of matrix $\mathbf{F}$ and the (i-k)-th row of matrix $\mathbf{F}_1$, respectively. The matrix $\mathbf{G}$ is computed by applying the same scheme as in (5.9) to an identity matrix of size $p \times p$.

Based on the above block state space model, the following frame by frame state space model is obtained. Each frame has 16 blocks in our simulation. The adjacent frames have an overlap of 15 blocks. At the output of the Kalman filter, the overlapped frames are added to obtain a smoothed version of the results. Let $\mathbf{X}_w$ denote the data of frame w:

$$\mathbf{X}_w = [\mathbf{S}'_{(w-1)+1}, \mathbf{S}'_{(w-1)+2}, \cdots \mathbf{S}'_{(w-1)+p}]' \tag{5.10}$$

$$\mathbf{X}_{w+1} = \mathbf{A}\mathbf{X}_w + \mathbf{\Gamma}\vec{\mathbf{U}}_{w+1} \tag{5.11}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{F} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{F} \end{bmatrix}, \mathbf{\Gamma} = \begin{bmatrix} \mathbf{G} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{G} \end{bmatrix}, \vec{\mathbf{U}}_w = [\mathbf{U}'_{(w-1)+1}, \mathbf{U}'_{(w-1)+2}, \cdots \mathbf{U}'_{(w-1)+p}]'.$$

The covariance matrix associated with $\vec{\mathbf{U}}_w$ is $\mathbf{Q}$. The observation data is represented by

$$\vec{\mathbf{Y}}_w = \mathbf{X}_w + \vec{\mathbf{V}}_w \tag{5.12}$$

where

$\vec{\mathbf{Y}}_w = [\mathbf{Y}'_{(w-1)+1}, \mathbf{Y}'_{(w-1)+2}, \cdots \mathbf{Y}'_{(w-1)+p}]'$, $\vec{\mathbf{V}}_w = [\mathbf{V}'_{(w-1)+1}, \mathbf{V}'_{(w-1)+2}, \cdots \mathbf{V}'_{(w-1)+p}]'$,

$\mathbf{Y}_N = [y((N-1)\cdot p+1), y((N-1)\cdot p+2), \cdots, y((N-1)\cdot p+p)]'$

and $\mathbf{V}_N = [v((N-1)\cdot p+1), v((N-1)\cdot p+2), \cdots, v((N-1)\cdot p+p)]'$.

The covariance matrix associated with $\vec{\mathbf{V}}_w$ is $\mathbf{R}$.

Equations (5.11) and (5.12) form the state space model in the time domain. To transform it into the wavelet domain, the wavelet packet transform is applied on both sides of (5.11):

$$\mathbf{T}\mathbf{X}_{w+1} = \mathbf{T}\mathbf{A}\mathbf{X}_w + \mathbf{T}\mathbf{\Gamma}\vec{\mathbf{U}}_{w+1} = \mathbf{T}\mathbf{A}\mathbf{T}'\mathbf{T}\mathbf{X}_w + \mathbf{T}\mathbf{\Gamma}\vec{\mathbf{U}}_{w+1} \tag{5.13}.$$

Let $\vec{\mathbf{X}}_w = \mathbf{T}\mathbf{X}_w$ which is the wavelet packet transform of the data, $\underline{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}'$ and $\underline{\mathbf{\Gamma}} = \mathbf{T}\mathbf{\Gamma}$, then in the wavelet domain, the state space model is represented by:

$$\vec{\mathbf{X}}_{w+1} = \underline{\mathbf{A}}\vec{\mathbf{X}}_w + \underline{\mathbf{\Gamma}}\vec{\mathbf{U}}_{w+1} \tag{5.14}$$

$$\vec{\mathbf{Y}}_w = \mathbf{T}'\vec{\mathbf{X}}_w + \vec{\mathbf{V}}_w \tag{5.15}.$$

The following Kalman filtering algorithm can be used for this model:

$$\mathbf{e}(w) = \vec{\mathbf{Y}}_w - \mathbf{T}'\vec{\mathbf{X}}_{w|w-1} \tag{5.16}$$

$$\mathbf{K}(w) = \mathbf{P}(w \mid w-1)\times\mathbf{T}(\mathbf{T}'\mathbf{P}(w \mid w-1)\mathbf{T}+\mathbf{R})^{-1} \tag{5.17}$$

$$\vec{\hat{\mathbf{X}}}_{w|w} = \vec{\hat{\mathbf{X}}}_{w|w-1} + \mathbf{K}(w)\times\mathbf{e}(w) \tag{5.18}$$

$$\mathbf{P}(w \mid w) = (\mathbf{I}-\mathbf{K}(w)\mathbf{T}')\times\mathbf{P}(w \mid w-1) \tag{5.19}$$

$$\vec{\hat{\mathbf{X}}}_{w+1|w} = \underline{\mathbf{A}}\vec{\hat{\mathbf{X}}}_{w|w} \tag{5.20}$$

$$\mathbf{P}(w+1 \mid w) = \underline{\mathbf{A}}\mathbf{P}(w \mid w)\underline{\mathbf{A}}' + \underline{\Gamma}\mathbf{Q}\underline{\Gamma}' \tag{5.21}$$

where $\mathbf{e}(w)$ is the innovation vector, $\mathbf{K}(w)$ is the Kalman gain, $\vec{\hat{\mathbf{X}}}_{w|w}$ represents the filtered estimate of state vector $\vec{\mathbf{X}}_w$, $\vec{\hat{\mathbf{X}}}_{w|w-1}$ is the minimum mean-square estimate of the state vector $\vec{\mathbf{X}}_w$, $\mathbf{P}(w \mid w)$ is the filtered state error covariance matrix, and $\mathbf{P}(w \mid w-1)$ is the *a priori* error covariance matrix. The estimation of the covariance matrices **Q** and **R** can be derived in the time domain [Ma04a].

5.3 Post-Filter Based on Masking Properties of Auditory Systems

Fig. 5.2 shows the diagram of the proposed system.

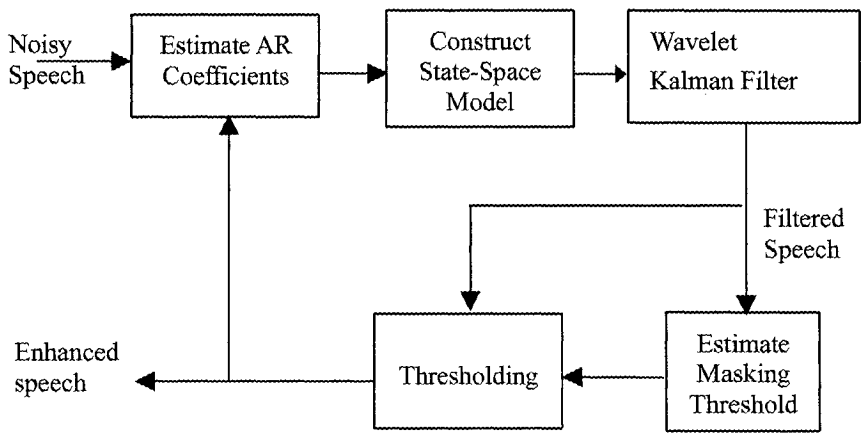


Fig. 5.2 Diagram of the proposed wavelet Kalman filter with perceptual post-filter system.

To perform the post-filter thresholding, the following procedure is used:

(1) Mapping the total masking level M_t in each critical band computed as in [Ma04a] to the wavelet packet domain, to obtain $T(\omega_j)$ ($j = 0, \cdots, 127$). In the simulation, a full wavelet packet tree was used to perform the decomposition, and each wavelet packet coefficient corresponds to a frequency bandwidth, which is analogous to 31.25 Hz in frequency domain (with a sampling frequency of 8 kHz). This step thus decomposes each critical band masking level M_t into the corresponding wavelet packet bandwidths.

(2) Estimating the power spectrum density (PSD) of the filtered state error in the wavelet packet domain, from the last 20 filtered state error covariance matrices $\mathbf{T}'\mathbf{P}(w-i\,|\,w-i)\mathbf{T}$ ($i=0,\cdots,N-1$). To do this, an average covariance matrix $\bar{\mathbf{P}}$ is obtained from $\mathbf{T}'\mathbf{P}(w-i\,|\,w-i)\mathbf{T}$. Then a covariance vector is constructed from matrix $\bar{\mathbf{P}}$, by averaging the elements of $\bar{\mathbf{P}}$ in each of its diagonals. The PSD in wavelet domain $P_e(\omega_i)$ is estimated by extending the covariance vector to the length of 128 data and using a wavelet packet transform on it. $P_e(\omega_i)$ is the absolute value of the transform coefficient. Then a thresholding is performed on the input wavelet packet transformed speech spectrum:

$$\left|\tilde{\hat{X}}(\omega_i)\right| = \begin{cases} \left|\hat{X}(\omega_i)\right| \times \alpha^{P_e(\omega_i)/T(\omega_i)} & P_e(\omega_i) < T(\omega_i) \\ \left|\hat{X}(\omega_i)\right| \times \alpha \times (1+\alpha^{P_e(\omega_i)/T(\omega_i)}) & \text{otherwise} \end{cases} \tag{5.22}$$

where $i=0,1,\cdots 127$, and α is a tonality coefficient [Ma04a] ($0 \le \alpha \le 1$, 0 is for a purely white noise signal and 1 is for a purely tonal signal).

(3) Using $\left|\tilde{\hat{X}}(\omega_i)\right|$ and the sign of $\hat{X}(\omega_i)$, keeping the last block of data and using the overlap and add method to obtain a smoothed wavelet spectrum. A frame (with length of 128) of enhanced wavelet domain data is thus obtained, and doing an inverse wavelet packet transform produces the improved frame of speech.

When the post-filter is designed, a tradeoff between signal distortion and residual noise is needed. Both the masking properties and the characteristics of the speech frame are taken into account in the thresholding procedure (5.22). The smaller the α is, i.e. the speech frame is more like a white noise, the more the Kalman filtered signal power is reduced. When the error power density $P_e(\omega_i)$ is smaller than the masking threshold $T(\omega_i)$, most of the power of the Kalman filtered signal can be kept to reduce distortion. The smaller the $P_e(\omega_i)$ is, the more the Kalman filtered signal energy is kept. On the other hand, if $P_e(\omega_i)$ is larger than $T(\omega_i)$, whether to increase the power of the Kalman filtered signal or to reduce it depends on the value of α and $P_e(\omega_i)$. From (5.22), if $P_e(\omega_i) > T(\omega_i)$ and $\alpha(1+\alpha^{P_e(\omega_i)/T(\omega_i)}) > 1$, then the following two solutions can be derived: $\alpha > \frac{\sqrt{5}-1}{2}$ and $P_e(\omega_i) < \frac{T\ln(1/\alpha-1)}{\ln\alpha}$. This case represents a speech frame which is more tone-like and the error estimate may not be too large. To mask the error estimate, the power of the

Kalman filtered signal is increased. In other cases, the power of the Kalman filtered signal is reduced. If $\alpha > \frac{\sqrt{5}-1}{2}$ and $P_e(\omega_i) > \frac{T\ln(1/\alpha - 1)}{\ln \alpha}$, the error estimate may be too large. If the signal power is increased, it may incur a large distortion. Thus in this case, the signal power is reduced. The larger the $P_e(\omega_i)$ is, the more the signal power is reduced. If $\alpha < \frac{\sqrt{5}-1}{2}$, the speech frame is assumed to be more like a noise frame. The larger the $P_e(\omega_i)$ is, the more the signal power is reduced.

5.4 Simulation Results

Six different speech sentences of 4 seconds each spoken by three females and three males were used in our simulations. The noise signals used were babble noise and street noise. All of the speech and noise files were taken from the ITU-T Supplement P.23 speech database. The sampling frequency used was 8000 Hz, and the input signals were normalized so that the amplitude is in the interval [-1, 1]. The AR prediction order p was set to be 8, and the AR coefficients were updated for every frame. The data length used to compute the AR parameters (i.e. to compute correlation values) was 256 samples, which includes the current noisy frame and one previous enhanced frame. DB4 wavelets were used and the decomposition level of the data is 7. In addition, no FFT is done on the input data of the post-filter of the proposed approach. Thus the computation cost is reduced by the new method. The performance index used was ITU-T P.862 PESQ scores, in order to have a close match with subjective speech quality scores. High scores stand for good speech quality. The average PESQ scores are shown in Table 5.1, and the detailed PESQ scores are shown in Appendix C with Table C.1, obtained by comparing classical and recently introduced speech enhancement algorithms for colored noise with the proposed method, under various input signal-to-noise ratios (SNR).

Table 5.1. Average PESQ scores obtained by different methods

input speech SNR → algorithm ↓	-5 dB	0 dB	5 dB	10 dB	15 dB
original noisy speech	1.25	1.56	1.84	2.15	2.47
spectral subtract. (SS)	0.79	1.15	1.53	1.81	2.02
perceptual SS [Vir99]	1.36	1.68	1.98	2.27	2.57
Kalman filtering [Pop98]	0.83	1.20	1.57	1.85	2.06
KLT approach [Mit00]	0.83	1.20	1.57	1.85	2.06
KFSFMC (Section 4.3.2)	1.66	1.94	2.20	2.45	2.71
WKFSFM	1.67	1.93	2.21	2.46	2.70

Compared with the time-domain Kalman filter method proposed in Section 4.3.2 (KFSFMC), the performance of the wavelet domain method proposed in this chapter (WKFSFM) produces very similar PESQ scores (but with a lower complexity). Both of these methods produce the best performance under any input SNR value.

5.5 Conclusions

This chapter proposed a wavelet Kalman filter concatenated with a perceptual post-filter for speech enhancement. The multi-resolution property of human auditory systems can be mapped very well to the wavelet packet transform domain. By constructing the state-space model in the wavelet domain, the output of the wavelet Kalman filtering algorithm is directly used by the post-filter without any transform, reducing the complexity compared to a pure time domain Kalman filter approach. The proposed method produced a better PESQ score performance than other published speech enhancement algorithms for colored noise, with a performance comparable to the time domain method in Section 4.3.2 (KFSFMC).

Chapter 6 Dual Constrained Unscented Kalman Filter for Enhancing Speech Degraded by Colored Noise

6.1 Nonlinear Speech Model

In the previous chapters, Kalman filtering algorithms including some perceptual hearing factors were introduced, using an AR model as the speech model. This is a linear prediction model for the vocal tract where the speech formant resonances are identified with the poles of the vocal tract transfer function. However there is strong theoretical and experimental evidence [MaP02] for the existence of important nonlinear aerodynamic phenomena during the speech production that cannot be accounted for by the linear model.

In [Joh03], the authors proposed a nonlinear speech enhancement approach using phase space or state space reconstruction. As shown in [Ban99], the speech is viewed as a nonlinear dynamical system. The time series data provide a scalar observation of a system's underlying dynamics. Using the method proposed by Takens [Tak81], the time series is embedded into an m-dimensional vector. The dimensions are time-lagged versions of the original time series [Joh03]. Since the clean signal will have a characteristic attractor structure but the additive noise will be distributed throughout the space, [Joh03] proposed to use a singular value decomposition (SVD) to project the data to a lower dimension, the final embedding dimension, where the true dynamics of the system reside. There is a relationship between this method and enhancement algorithms based on sub-space decomposition, for the special case where the time lag equals 1.

In [MaP02], the authors summarized a nonlinear speech model using the modulation, fractal and chaotic structure of speech signals. They proposed to model each speech resonance with an AM-FM signal as follows:

$$x(t) = a(t)\cos[\phi(t)] = a(t)\cos\left[2\pi\int f(\tau)d\tau\right] \tag{6.1}$$

and the total speech signal as a superposition of such AM-FM signals, $\sum_k a_k(t)\cos[\phi_k(t)]$, one for each formant. Here $a(t)$ is the instantaneous amplitude signal and $f(t)$ is the instantaneous cyclic frequency representing the time-varying formant signal.

However, the nonlinearity of speech is still an open problem, and a more flexible and self-adjusting approach that can accommodate various types of nonlinear behaviors representing a broad class of nonlinear mapping is required. Neural networks (NNs) are suitable for such a mapping and approximation. It has been shown that an NN with one hidden layer can approximate any continuous mapping defined on compact sets in $\mathbf{R}^n$ [Fun89, Hor91]. In this chapter a neural network is used to approximate the nonlinear speech model as in [Nel97].

6.2 A Neural State-Space Model for Speech Enhancement

In this chapter, the noisy speech signal $y(k)$ is modeled as a nonlinear autoregression with both process and additive observation noise as follows:

$$x(k) = f(x(k-1),\cdots,x(k-M),\mathbf{w}(k)) + v(k) \tag{6.2}$$

$$y(k) = x(k) + n(k) \tag{6.3}$$

where $x(k)$ represents the true underlying speech signal driven by process noise $v(k)$, f is a nonlinear function of past values of $x(k)$ parameterized by $\mathbf{w}$, and $\mathbf{w}$ is the weight vector. The speech is assumed to be stationary over short segments, with each segment having a different model. Only $y(k)$ is the available observation which contains additive noise, $n(k)$. The optimal

estimation given the noisy estimations $\mathbf{y}(k) = \{y(k), y(k-1), \cdots y(0)\}$ is $E\{x(k) \mid \mathbf{y}(k)\}$.

If a set of clean speech signal $x(k)$ is available, we can use the true $x(k)$ as the target to train a neural network. Then the approximation of $E\{x(k) \mid \mathbf{y}(k)\}$ can be obtained. However, it is assumed that clean speech signal is never available, $x(k)$ can only be estimated from the noisy observations alone. To solve this problem, as shown in [Wan98], function $f(\cdot)$ is assumed to be in the class of feedforward neural network models, and we compute the dual estimation of both states $\hat{x}(k)$ and weights $\hat{\mathbf{w}}$ based on a Kalman filtering approach.

To address such a nonlinear problem, an extended Kalman filter (EKF) is a candidate approach. However, as shown in [Wan00], there are some flaws in using the EKF. In the EKF, the state distribution is approximated by a Gaussian random variable (GRV), which is then propagated analytically through the first-order linearization of the nonlinear system. This can introduce large errors in the true posterior mean and covariance of the transformed GRV, which may lead to sub-optimal performance and sometimes divergence of the filter [Wan00]. In [Jul97], Julier and Uhlmann proposed a new extension of the Kalman filter to the nonlinear systems, the Unscented Kalman Filter (UKF). In the UKF, the state distribution is again approximated by a GRV, but is now represented using a minimal set of carefully chosen sample points. These sample points completely capture the true mean and covariance of the GRV, and when propagated through the nonlinear system, capture the posterior mean and covariance accurately to the 3rd order for any nonlinearity. In contrast, the EKF only achieves first-order accuracy. The computational complexity of the UKF is the same order as that of the EKF [Wan00]. We further extend the UKF to a perceptually constrained UKF by combining the masking properties of human auditory systems in this chapter. In addition, it is assumed that the speech is corrupted by colored noise.

As we did in Chapter 3, when colored noise is taken into account, a mixed state-space model of

equation (6.1) and (6.3) is formulated:

$$\begin{cases} \mathbf{x}(k) = \mathbf{F}[\mathbf{x}(k-1), \mathbf{w}] + Bv(k) \\ \mathbf{y}(k) = \mathbf{x}(k) + \mathbf{n}(k) \end{cases} \tag{6.4}$$

with

$$\mathbf{x}(k) = \begin{bmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-M+1) \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{F}[\mathbf{x}(k), \mathbf{w}] = \begin{bmatrix} f(x(k), \cdots, x(k-M+1), \mathbf{w}) \\ x(k) \\ \vdots \\ x(k-M+2) \end{bmatrix}, \quad \mathbf{y}(k) = \begin{bmatrix} y(k) \\ y(k-1) \\ \vdots \\ y(k-M+1) \end{bmatrix}$$

and $\mathbf{n}(k) = \begin{bmatrix} n(k) \\ n(k-1) \\ \vdots \\ n(k-M+1) \end{bmatrix}$,

where $\mathbf{x}(k)$ is the state vector. The process noise $v(k)$ is assumed to be zero mean, white Gaussian noise with variance σ_v^2. The measurement noise $n(k)$ is assumed to be colored noise with mean value $\bar{n}$ and covariance matrix, $\mathbf{R}$.

Because the weight vector $\mathbf{w}$ is unknown, a separate state-space formulation for the underlying weights is given by:

$$\begin{cases} \mathbf{w}(k) = \mathbf{w}(k-1) \\ \mathbf{y}(k) = \mathbf{F}[\mathbf{x}(k-1), \mathbf{w}(k)] + Bv(k-1) + \mathbf{n}(k) \end{cases} \tag{6.5}$$

where the neural network $\mathbf{F}(\cdot)$ plays the role of a time-varying nonlinear observation on $\mathbf{w}$. Although $\mathbf{w}(k) = \mathbf{w}(k-1)$ represents a stationary system, the estimate provided by the equation, $\mathbf{y}(k) = \mathbf{F}[\mathbf{x}(k-1), \mathbf{w}(k)] + Bv(k-1) + \mathbf{n}(k)$, can also be used for time varying signals such as speech.

6.3 Perceptually Constrained Dual Unscented Kalman Filter

6.3.1 Summary of EKF Algorithm and the Flaws

To solve the nonlinear problems shown in equation (6.4) and (6.5), we can use the extended

Kalman filter to effectively approximate the nonlinear function with a time-varying linear one. The EKF algorithm for (6.4) is given by:

$$\begin{cases} \hat{\mathbf{x}}(k \mid k-1) = \mathbf{F}[\hat{\mathbf{x}}(k-1 \mid k-1), \hat{\mathbf{w}}(k-1 \mid k-1)] \\ \mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k-1) = \mathbf{A}\mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k)\mathbf{A}^T + \mathbf{B}\sigma_v^2\mathbf{B}^T \\ \text{where} \qquad \mathbf{A} \equiv \left.\dfrac{\partial \mathbf{F}[\hat{\mathbf{x}}, \hat{\mathbf{w}}]}{\partial \hat{\mathbf{x}}}\right|_{\hat{\mathbf{x}}(k-1 \mid k-1)} \\ \mathbf{K}(k) = \mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k-1) \times (\mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k-1) + \mathbf{R})^{-1} \\ \mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k) = (\mathbf{I} - \mathbf{K}(k)) \times \mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k-1) \\ \hat{\mathbf{x}}(k \mid k) = \hat{\mathbf{x}}(k \mid k-1) + \mathbf{K}(k)(\mathbf{y}(k) - \hat{\mathbf{x}}(k \mid k-1)) \end{cases} \tag{6.6}$$

where $\mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k)$ is the filtered state error covariance matrix; and $\mathbf{P}_{\tilde{\mathbf{x}}}(k \mid k-1)$ is the predicted error covariance matrix; $\mathbf{K}(k)$ is the Kalman gain.

For equation (6.5), the EKF algorithm to estimate $\mathbf{w}(k)$ is as follows:

$$\begin{cases} \hat{\mathbf{w}}(k \mid k-1) = \hat{\mathbf{w}}(k-1 \mid k-1) \\ \mathbf{P}_{\tilde{\mathbf{w}}}(k \mid k-1) = \mathbf{P}_{\hat{\mathbf{w}}}(k-1 \mid k-1) \\ \mathbf{K}_{\hat{\mathbf{w}}}(k) = \mathbf{P}_{\tilde{\mathbf{w}}}(k \mid k-1)\mathbf{H}^T[\mathbf{H}\mathbf{P}_{\tilde{\mathbf{w}}}(k \mid k-1)\mathbf{H}^T + \mathbf{R} + \mathbf{B}\sigma_v^2\mathbf{B}^T]^{-1} \\ \text{where} \quad \mathbf{H} \equiv \left.\dfrac{\partial \mathbf{F}[\hat{\mathbf{x}}, \hat{\mathbf{w}}]}{\partial \hat{\mathbf{w}}}\right|_{\hat{\mathbf{w}}(k-1 \mid k-1)} \\ \mathbf{P}_{\tilde{\mathbf{w}}}(k \mid k) = (\mathbf{I} - \mathbf{K}_{\hat{\mathbf{w}}}(k)\mathbf{H})\mathbf{P}_{\tilde{\mathbf{w}}}(k \mid k-1) \\ \hat{\mathbf{w}}(k \mid k) = \hat{\mathbf{w}}(k \mid k-1) + \mathbf{K}_{\tilde{\mathbf{w}}}(k)\{\mathbf{y}(k) - \mathbf{F}[\hat{\mathbf{x}}(k-1 \mid k-1), \hat{\mathbf{w}}(k \mid k-1)]\} \end{cases} \tag{6.7}$$

As described in section 6.2, EKF has well-known drawbacks summarized below:

- Sensitivity to the disturbance due to the linearization process.
- Heavy computation load due to the calculation of Jacobian matrices.

To replace the EKF in nonlinear filtering problems, the UKF was proposed by Julier and Uhlmann [Jul97]. The following section describes the principle of the UKF in summary and the perceptually constrained dual UKF.

6.3.2 Perceptually Constrained Dual UKF

6.3.2.1 Unscented Transformation

The unscented Kalman filter is a straightforward extension of the unscented transformation (UT) to the recursive estimation. The UT is a method for calculating the statistics of a random variable which undergoes a nonlinear transformation [Jul97]. It is described below in summary.

Given an L-dimensional random variable $\mathbf{x}$ with mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{P}_x$, the statistics of a random variable $\mathbf{y} = g(\mathbf{x})$ can be calculated by the following procedure:
Firstly, form a matrix χ of $2L+1$ *sigma* vectors χ_i with corresponding weights W_i, according to the following rules:

$$\begin{cases} \chi_0 = \bar{\mathbf{x}} \\ \chi_i = \bar{\mathbf{x}} + \left(\sqrt{(L+\lambda)\mathbf{P}_x}\right)_i & i = 1,\cdots,L \\ \chi_i = \bar{\mathbf{x}} - \left(\sqrt{(L+\lambda)\mathbf{P}_x}\right)_{i-L} & i = L+1,\cdots,2L \\ W_0^{(m)} = \lambda/(L+\lambda) \\ W_0^{(c)} = \lambda/(L+\lambda) + (1-\alpha^2+\beta) \\ W_i^{(m)} = W_i^{(c)} = 1/\{2(L+\lambda)\} & i = 1,\cdots.2L \end{cases} \tag{6.8}$$

where $\lambda = \alpha^2(L+\gamma) - L$ is a scaling parameter, α determines the spread of the sigma points around $\bar{\mathbf{x}}$ and is usually set to a small positive value. γ is a secondary scaling parameter which is usually set to 0, and β is used to incorporate prior knowledge of the distribution of $\mathbf{x}$ (for Gaussian distributions, $\beta = 2$ is optimal). $\left(\sqrt{(L+\lambda)\mathbf{P}_x}\right)_i$ is the ith row of the matrix square root. These sigma vectors are propagated through the nonlinear function,

$$\mathrm{Y}_i = g(\chi_i) \qquad i = 0,\cdots,2L \tag{6.9}$$

and the mean and covariance for $\mathbf{y}$ are approximated by:

$$\bar{\mathbf{y}} \approx \sum_{i=0}^{2L} W_i^{(m)} \mathrm{Y}_i \tag{6.10}$$

$$\mathbf{P}_{\mathrm{y}} \approx \sum_{i=0}^{2L} W_i^{(c)} \{\mathrm{Y}_i - \bar{\mathbf{y}}\}\{\mathrm{Y}_i - \bar{\mathbf{y}}\}^T \tag{6.11}$$

For Gaussian inputs, the UT results can achieve the accuracy of the third order. For non-Gaussian inputs, approximations are accurate to at least the second-order. The accuracy of third and higher order moments is determined by the choice of α and β [Jul97].

6.3.2.2 Perceptually Constrained UKF Algorithms

By applying the UT sigma point selection scheme, as shown in equation (6.8), to the state vector to calculate the corresponding sigma matrix, the UKF algorithm for equation (6.6) is as shown in Table 6.1. Comparing Table 6.1 with equation (6.6), it can be found that only the equations for computing the predicted state-vector and predicted state-error covariance matrix are different. No derivative is required.

If the masking properties of human auditory systems are taken into account, the algorithm introduced in Table 6.1 can be modified referring to the method proposed in Section 5.3. The new algorithm is named as PCDUKF (Perceptually Constrained Dual UKF). To obtain the Kalman gain, the following optimization problem has to be solved:

$$\begin{aligned} &\min \quad \Delta_0 \\ &\text{subject to} \\ &\begin{cases} 0 \le \Delta_0 + 2\sum_{i=1}^{p-1} \Delta_i \cos(\frac{2\pi}{N} \cdot i \cdot m) \le T(m), \quad m = 0,1,\cdots,128 \\ |\Delta_0| \ge |\Delta_j| \\ \Delta_0 > 0 \end{cases} \end{aligned} \tag{6.12}$$

with:

$$\Delta_0 = \frac{1}{p}\left[\mathrm{Tr}\mathbf{P} - 2 * \mathrm{Tr}\mathbf{KHP} + \mathrm{Tr}\mathbf{K}(\mathbf{HPH}^T)\mathbf{K}^T + \mathrm{Tr}\mathbf{KRK}^T\right] \tag{6.13}$$

$$\Delta_j = \frac{1}{p-j}\left[\sum_{i=1}^{p-j} \gamma_{i,i+j} - (\mathbf{KHP})_{i,i+j} - (\mathbf{PH}^T\mathbf{K}^T)_{i,i+j} + (\mathbf{KHPH}^T\mathbf{K}^T)_{i,i+j} + (\mathbf{KRK}^T)_{i,i+j}\right] \tag{6.14}$$

where Tr is the trace of a matrix, $\mathbf{P}$ and $\mathbf{K}$ represent $\mathbf{P}_{\tilde{\mathbf{x}}}(k|k-1)$ and $\mathbf{K}(k)$ respectively, $\gamma_{i,i+j}$ is the element of $\mathbf{P}$ which lies on the ith row and the $(i+j)$th column, $(\mathbf{KHP})_{i,i+j}$, $(\mathbf{PH}^T\mathbf{K}^T)_{i,i+j}$, $(\mathbf{KHPH}^T\mathbf{K}^T)_{i,i+j}$ and $(\mathbf{KRK}^T)_{i,i+j}$ are the elements of the corresponding matrix which is on the ith row and the $(i+j)$th column respectively, $\mathbf{H}=\mathbf{I}$.

Table 6.1 The UKF equations for equation (6.6)

Initialize with:

$$\hat{\mathbf{x}}(0|0)=E[\mathbf{x}(0)]$$

$$\mathbf{P}_{\tilde{\mathbf{x}}}(0|0)=E\{[\mathbf{x}(0)-\hat{\mathbf{x}}(0|0)][\mathbf{x}(0)-\hat{\mathbf{x}}(0|0)]^T\}$$

For $k=1,\cdots,\infty$, calculate the sigma matrix:

$$\begin{cases}\chi_0(k-1|k-1)=\hat{\mathbf{x}}(k-1|k-1)\\ \chi_i(k-1|k-1)=\hat{\mathbf{x}}(k-1|k-1)+\left(\sqrt{(L+\lambda)\mathbf{P}_{\tilde{\mathbf{x}}}(k-1|k-1)}\right)_i\\ \chi_{i+L}(k-1|k-1)=\hat{\mathbf{x}}(k-1|k-1)-\left(\sqrt{(L+\lambda)\mathbf{P}_{\tilde{\mathbf{x}}}(k-1|k-1)}\right)_i\end{cases},\qquad i=1,\cdots,L$$

Time update:

$$\chi_i(k|k-1)=\mathbf{F}[\chi_i(k-1|k-1),\hat{\mathbf{w}}(k-1|k-1)],\qquad i=0,\cdots,2L$$

Predicted state-vector:

$$\hat{\mathbf{x}}(k|k-1)=\sum_{i=0}^{2L}W_i^{(m)}\chi_i(k|k-1)$$

Predicted state-error covariance matrix:

$$\mathbf{P}_{\tilde{\mathbf{x}}}(k|k-1)=\sum_{i=0}^{2L}W_i^{(c)}[\chi_i(k|k-1)-\hat{\mathbf{x}}(k|k-1)][\chi_i(k|k-1)-\hat{\mathbf{x}}(k|k-1)]^T+\mathbf{B}\sigma_v^2\mathbf{B}^T$$

Kalman gain:

$$\mathbf{K}(k)=\mathbf{P}_{\tilde{\mathbf{x}}}(k|k-1)\times(\mathbf{P}_{\tilde{\mathbf{x}}}(k|k-1)+\mathbf{R})^{-1}$$

Filtered estimate of the state vector:

$$\hat{\mathbf{x}}(k|k)=\hat{\mathbf{x}}(k|k-1)+\mathbf{K}(k)(\mathbf{y}(k)-\hat{\mathbf{x}}(k|k-1))$$

Filtered state-error covariance matrix:

$$\mathbf{P}_{\tilde{\mathbf{x}}}(k|k)=(\mathbf{I}-\mathbf{K}(k))\times\mathbf{P}_{\tilde{\mathbf{x}}}(k|k-1)$$

The perceptually constrained UKF equations are shown as in Table 6.2:

Table 6.2 The Perceptually Constrained UKF equations for equation (6.6)

Initialize with:

$$\hat{\mathbf{x}}(0\,|\,0) = E[\mathbf{x}(0)]$$

$$\mathbf{P}_{\tilde{\mathbf{x}}}(0\,|\,0) = E\{[\mathbf{x}(0) - \hat{\mathbf{x}}(0\,|\,0)][\mathbf{x}(0) - \hat{\mathbf{x}}(0\,|\,0)]^T\}$$

For $k = 1, \cdots, \infty$, calculate the sigma matrix:

$$\begin{cases} \chi_0(k-1\,|\,k-1) = \hat{\mathbf{x}}(k-1\,|\,k-1) \\ \chi_i(k-1\,|\,k-1) = \hat{\mathbf{x}}(k-1\,|\,k-1) + \left(\sqrt{(L+\lambda)\mathbf{P}_{\hat{\mathbf{x}}}(k-1\,|\,k-1)}\right)_i \\ \chi_{i+L}(k-1\,|\,k-1) = \hat{\mathbf{x}}(k-1\,|\,k-1) - \left(\sqrt{(L+\lambda)\mathbf{P}_{\hat{\mathbf{x}}}(k-1\,|\,k-1)}\right)_i \end{cases}, \qquad i = 1, \cdots, L$$

Time update:

$$\chi_i(k\,|\,k-1) = \mathbf{F}[\chi_i(k-1\,|\,k-1), \hat{\mathbf{w}}(k-1\,|\,k-1)], \qquad i = 0, \cdots, 2L$$

Predicted state-vector:

$$\hat{\mathbf{x}}(k\,|\,k-1) = \sum_{i=0}^{2L} W_i^{(m)} \chi_i(k\,|\,k-1)$$

Predicted state-error covariance matrix:

$$\mathbf{P}_{\tilde{\mathbf{x}}}(k\,|\,k-1) = \sum_{i=0}^{2L} W_i^{(c)} [\chi_i(k\,|\,k-1) - \hat{\mathbf{x}}(k\,|\,k-1)][\chi_i(k\,|\,k-1) - \hat{\mathbf{x}}(k\,|\,k-1)]^T + \mathbf{B}\sigma_v^2\mathbf{B}^T$$

Kalman gain:

$\mathbf{K}(k)$: obtained by solving equation (6.12)

Filtered estimate of the state vector:

$$\hat{\mathbf{x}}(k\,|\,k) = \hat{\mathbf{x}}(k\,|\,k-1) + \mathbf{K}(k)(\mathbf{y}(k) - \hat{\mathbf{x}}(k\,|\,k-1))$$

Filtered state-error covariance matrix:

$$\mathbf{P}_{\tilde{\mathbf{x}}}(k\,|\,k) = [\mathbf{I} - \mathbf{K}(k)] \times \mathbf{P}_{\tilde{\mathbf{x}}}(k\,|\,k-1)[\mathbf{I} - \mathbf{K}(k)]^T + \mathbf{K}(k)\mathbf{R}\mathbf{K}^T(k)$$

The UKF algorithm for equation (6.7) is shown as in Table 6.3:

Table 6.3 The UKF equations for equation (6.7)

Initialize with:

$$\hat{\mathbf{w}}(0 \mid 0) = E[\mathbf{w}(0)]$$

$$\mathbf{P}_{\hat{\mathbf{w}}}(0 \mid 0) = E\{[\mathbf{w}(0) - \hat{\mathbf{w}}(0 \mid 0)][\mathbf{w}(0) - \hat{\mathbf{w}}(0 \mid 0)]^T\}$$

Predicted weight vector:

$$\hat{\mathbf{w}}(k \mid k-1) = \hat{\mathbf{w}}(k-1 \mid k-1)$$

Predicted weight error covariance matrix:

$$\mathbf{P}_{\hat{\mathbf{w}}}(k \mid k-1) = \mathbf{P}_{\hat{\mathbf{w}}}(k-1 \mid k-1)$$

Calculate the sigma matrix:

$$\begin{cases} \varpi_0(k-1 \mid k-1) = \hat{\mathbf{w}}(k-1 \mid k-1) \\ \varpi_i(k-1 \mid k-1) = \hat{\mathbf{w}}(k-1 \mid k-1) + \left(\sqrt{(L+\lambda)\mathbf{P}_{\hat{\mathbf{w}}}(k-1 \mid k-1)}\right)_i \\ \varpi_{i+L}(k-1 \mid k-1) = \hat{\mathbf{w}}(k-1 \mid k-1) - \left(\sqrt{(L+\lambda)\mathbf{P}_{\hat{\mathbf{w}}}(k-1 \mid k-1)}\right)_i \end{cases}, \quad i = 1, \cdots, L$$

Time update:

$$\varpi_i(k \mid k-1) = \varpi_i(k-1 \mid k-1), \quad i = 0, \cdots, 2L$$

$$\mathrm{Y}_i(k \mid k-1) = \mathbf{F}[\hat{\mathbf{x}}(k-1 \mid k-1), \varpi_i(k-1 \mid k-1)], \quad i = 0, \cdots, 2L$$

$$\hat{\mathbf{y}}(k \mid k-1) = \sum_{i=0}^{2L} W_i^{(m)} \mathrm{Y}_i(k \mid k-1) + \bar{\mathbf{n}}, \quad i = 0, \cdots, 2L$$

Measurement update equations:

$$\mathbf{P}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}(k) = \sum_{i=0}^{2L} W_i^{(m)} [\mathrm{Y}_i(k \mid k-1) - \hat{\mathbf{y}}(k \mid k-1)][\mathrm{Y}_i(k \mid k-1) - \hat{\mathbf{y}}(k \mid k-1)]^T$$

$$\mathbf{P}_{\hat{\mathbf{w}}\hat{\mathbf{y}}}(k) = \sum_{i=0}^{2L} W_i^{(c)} [\varpi_i(k \mid k-1) - \hat{\mathbf{w}}(k \mid k-1)][\varpi_i(k \mid k-1) - \hat{\mathbf{w}}(k \mid k-1)]^T$$

$$\mathbf{\Theta}(k) = \mathbf{P}_{\hat{\mathbf{w}}\hat{\mathbf{y}}}(k)\mathbf{P}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}(k)$$

Filtered weight vector:

$$\hat{\mathbf{w}}(k \mid k) = \hat{\mathbf{w}}(k \mid k-1) + \mathbf{\Theta}(k)[\mathbf{y}(k) - \hat{\mathbf{y}}(k \mid k-1)]$$

Filtered weight error covariance matrix:

$$\mathbf{P}_{\hat{\mathbf{w}}}(k \mid k) = \mathbf{P}_{\hat{\mathbf{w}}}(k \mid k-1) - \mathbf{\Theta}(k)\mathbf{P}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}(k)\mathbf{\Theta}^T(k)$$

Note that the weight transition is linear in the weight filter, so the nonlinearity is restricted to the measurement equation.

The dual UKF estimation scheme is shown as in Fig. 6.1. Two UKFs are run simultaneously for signal and weight estimation. At every time-step, the current estimate of the weights is used in the signal-filter, and the current estimate of the signal-state is used in the weight-filter [Wan00]. For finite data sets, the algorithm is run iteratively for the data until the weights converge.

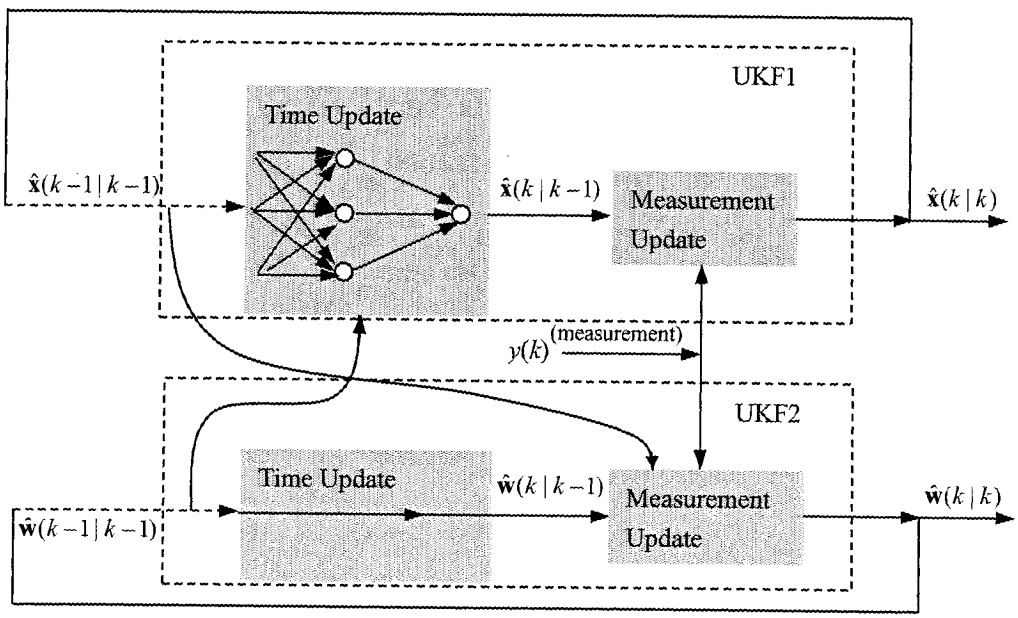


Fig. 6.1 The Perceptually constrained dual UKF for Speech Enhancement. UKF1 represents the perceptually constrained UKF filter for the states (enhanced speech) and UKF2 represents the UKF filter for the weights of the speech model.

6.4. Dual UKF with Perceptual Post-Filter Algorithm

The constrained dual UFK algorithm of Section 6.3 (PCDUKF) provided a theoretical foundation for combining the masking properties of human auditory systems with a dual UKF algorithm for speech enhancement. The simulations will also show the excellent performance of this approach.

However, the requirement of solving a nonlinear optimization problem at each step makes the computation load of the method quite large, especially if the algorithm is to be implemented for real time applications. A simpler heuristic implementation of combining the dual UKF and masking properties is considered in this section. A standard UKF is used and a post-filter based on frequency and time domain masking properties is applied at the output of the dual UKF. The method is named as DUKFPPF (Dual UKF with Perceptual Post-Filter). The post-filter adjusts the spectrum of the dual UKF filtered speech signal by making the power of the clean speech state estimate error (computed in one of the UKF filters) lower than the masking threshold. Fig. 6.2 shows the block diagram of the resulting system. While the dual UKF filtering is performed on a sample-by- sample basis, the perceptual post-filtering using masking effects is performed on a frame-by-frame basis, unlike the perceptual processing of the truly constrained Kalman filtering method in Section 6.3 (PCDUKF). This means that the masking thresholds are now computed only once per frame. The last element of $\hat{\mathbf{x}}(n|n)$ is called $\hat{s}(n)$, and it is considered as the output of the dual UKF filter.

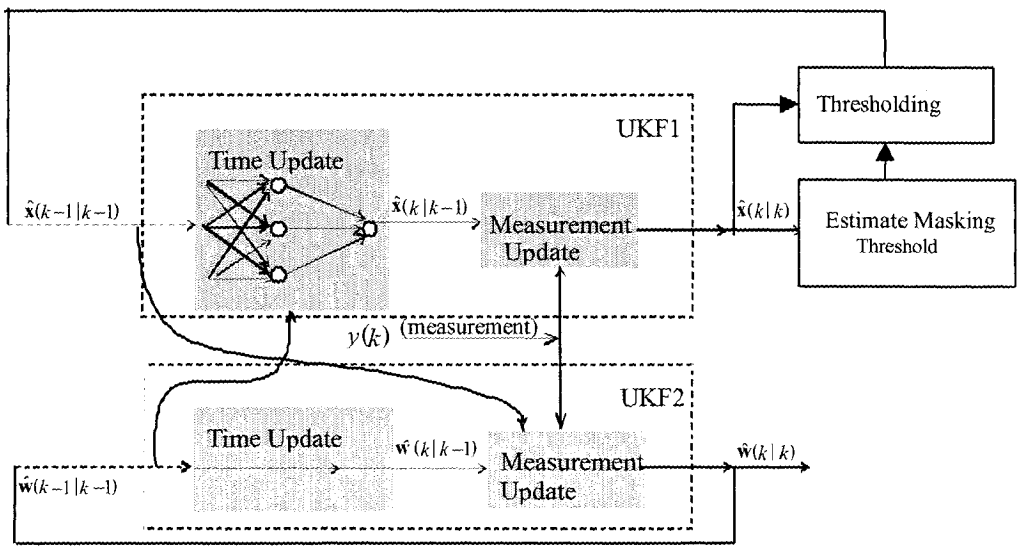


Fig. 6.2 The Dual UKF with a Perceptual Post-Filter for Speech Enhancement

6.4.1 Post-Filter Based on Masking Properties of Human Auditory Systems

The post-filter performs a thresholding on its input signal $\hat{s}(n)$, based on a computed total masking level $M_t(i)$ as described in Chapter 3. To perform the thresholding, the following procedure is used:

(1) Map the total masking level $M_t(i)$ in each critical band to the frequency domain (256 FFT bins, 129 freq.) to obtain the levels $T(\omega_j)$ ($j = 0, 1, \cdots, 128$).

(2) From the last 40 computed filtered state error covariance matrices, $\mathbf{P}(n-i \mid n-i)$ ($i = 0, \cdots, 39$), estimate the power spectrum density (PSD) of the filtered state error. The estimation method was described by Step 2 of algorithm KFSFMC which was described in Section 4.3.2.

(3) Estimate the signal to noise ratio of the current frame. Let $P_{\hat{s}}$ equal the power of the current frame of the Kalman filtered signal $\hat{\mathbf{s}}(n)$,

$$P_{\hat{s}} = \sum_{i=0}^{128} \left\{ \frac{\left|\hat{S}(\omega_i)\right|^2}{L} \right\} \tag{6.15}$$

Let $\hat{P}_s$ be the estimate power of the clean speech signal of the current frame. Then $\hat{P}_s$ is defined as

$$\hat{P}_s = P_{\hat{s}} - \sum_{i=0}^{128} P_e(\omega_i) \tag{6.16}$$

The estimated signal to noise ratio of the current frame is defined as

$$\overline{SNR} = \hat{P}_s / \sum_{i=0}^{128} P_e(\omega_i) \tag{6.17}$$

(4) Perform the thresholding on the speech spectrum:

$$\left|\tilde{\hat{S}}(\omega_i)\right| = \begin{cases} \left|\hat{S}(\omega_i)\right| \times (\alpha^{P_e(\omega_i)/T(\omega_i)} + \overline{SNR}^2) & \text{if } P_e(\omega_i) < T(\omega_i) \\ \left|\hat{S}(\omega_i)\right| \times (\alpha \times (1 + \alpha^{P_e(\omega_i)/T(\omega_i)}) + \overline{SNR}^2) & \text{otherwise} \end{cases} \tag{6.18}$$

where $i = 0,1,\cdots 128$, and α is a tonality coefficient. The range of α is from 0 to 1, where 0 corresponds to a completely noise-like signal and 1 corresponds to a completely tonal signal.

(5) Do an IFFT using $\left|\tilde{\tilde{S}}(\omega_i)\right|$ and the phase of $\hat{S}(\omega_i)$, keeping the last 80 values of the size-256 IFFT outputs to obtain the improved frame of speech signal $\tilde{\mathbf{s}}(n) = [\tilde{s}(n-79), \tilde{s}(n-78), \cdots, \tilde{s}(n)]$..

When the post-filter is designed, a tradeoff is made between signal distortion and residual noise. The scheme is the same as that in Chapter 4.

6.5 Simulation Results and Conclusions

Six different speech sentences were used in the simulations. In this chapter, the noise signals used were white car noise, babble noise, street noise, and a generated colored noise as shown in equation (3.99). The order of the nonlinear speech model is 10. Tables 6.4 and 6.5 show the average PESQ scores obtained by different methods for white noise and colored noise, respectively. In Appendix D, the detailed PESQ scores obtained by different methods for white noise and colored noise are shown by Table D.1 and Table D.2, respectively. Fig. D.1 shows the average PESQ scores obtained for white noise signals by different methods and different speakers (females and males). Fig. D.2 shows the average PESQ scores obtained for different colored noise signals by different methods and different speakers (females and males). Table D.3 shows the average PESQ scores obtained by the perceptually constrained dual UKF, and the dual UKF with a perceptual post-filter.

Table 6.4 Average PESQ scores obtained by different methods with white noise signal

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	1.17	0.66	1.27	0.70	1.66	1.58
0	1.40	1.02	1.54	1.07	1.96	1.87
5	1.64	1.44	1.81	1.47	2.24	2.30
10	1.94	1.76	2.16	1.80	2.52	2.60
15	2.27	1.97	2.49	2.00	2.79	2.86

Table 6.5 Average PESQ scores obtained by different methods with colored noise signals

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.25	0.79	1.36	0.83	0.83	1.71	1.61
0	1.56	1.15	1.68	1.20	1.20	1.97	1.88
5	1.84	1.53	1.98	1.57	1.57	2.25	2.31
10	2.15	1.81	2.27	1.85	1.85	2.49	2.56
15	2.47	2.02	2.57	2.06	2.06	2.75	2.82

The results of Tables D.1 and D.2 show that for high input SNR cases (SNR>5dB), the PESQ scores of the perceptually constrained dual UKF method introduced in this chapter (PCDUKF) have the best performance, compared with the methods of the previous chapters. However for low SNR environments, the method produces worse PESQ scores than the method from Chapter 3 (CKFSFM). The underlying reason is that we can only estimate the weight vector **w** from the noisy signal. When the input SNR is high, the estimation is more accurate. However for low SNR, the estimation of **w** may deviate far from the true value. Thus the accuracy of the nonlinear speech model is sensitive to the input SNR.

Table D.3 shows that with the dual UKF perceptual post-filter algorithm, the PESQ scores are about 0.05 less than that produced by the perceptually constrained dual UKF algorithm. From the point of the view of the computational cost, the dual UKF perceptual post-filter algorithm has a lower computational load compared to the perceptually constrained dual UKF algorithm. The post-filter method in Section 4.3.2 (KFSFMC) has a lower computational cost than the dual UKF post-filter algorithm. Thus, for the case of low input SNR (<5dB), KFSFMC may be the best algorithm for real time applications, while for high input SNR (>=5dB) cases, the best algorithm for real time application would seem to be the dual UKF with perceptual post-filter.

In the methods proposed in this chapter, there was typically an initial delay of 30ms required to find stable weights at the initialization of the methods. To conclude, the perceptually constrained dual-UKF method proposed in this chapter is an interesting option for offline speech enhancement (except for low SNR cases, where CKFSFM is more suitable), while for on-line real-time speech enhancement the dual UKF with post-filter algorithm (for higher input SNR) and KFSFMC (for lower input SNR) are more suitable.

Chapter 7 Subjective Test Results

A first informal subjective test was performed with eighteen students involved in the test. Five of them were female students, while the rest were male students. The age of the students was between 20 and 35. A female speech signal was used for the tests. Five different noise types including babble, car, street, white and synthetically generated colored noise were used in the first test. Two input SNRs were chosen for the tests: 0dB and 10dB. To simplify the testing, only the clean speech signal, the noisy speech signal and the signal enhanced by the method proposed in Section 4.3.2 (KFSFMC) were used in this first test. Table 7.1 shows the informal mean opinion scores (MOS) obtained by averaging the results given by the students.

Another informal subjective test was performed with eight people involved in the test. Two of them were females, while the rest were males. The age of the people was between 26 and 45. A female speech signal was used for the tests. Five different noise types including babble, car, street, white and synthetically generated colored noise were used in the tests. Two input SNRs were chosen for the tests: 0 dB and 10 dB. To simplify the testing, only the clean speech signal, the noisy speech signal, the signal enhanced by the method proposed in [Vir99], by the KFSFMC, and by the DUKFPPF were used in this second test. Table 7.2 shows the informal mean opinion scores (MOS) obtained by averaging the results given by the people.

From Table 7.1 and Table 7.2, it can be observed that the informal MOS scores match very well the trend of the PESQ scores, and that the gap between the enhanced speech and the noisy speech scores is also similar in both scales. This indicates that the use of the objective PESQ measure to evaluate the quality of noisy and enhanced speech files in this thesis is a valid indication of the

true speech quality, and that the new methods introduced in this thesis do seem to provide a meaningful performance gain compared to existing methods.

Table 7.1 Comparison between informal MOS and PESQ scores: first subjective test

	Informal MOS scores		PESQ scores	
Noise type and input SNR	Noisy signal	Enhanced Signal	Noisy signal	Enhanced signal
babble, 0dB	1.51	1.95	1.22	1.77
babble, 10dB	2.43	2.74	1.88	2.4
car, 0dB	1.59	1.90	1.42	1.88
car, 10dB	2.22	2.67	2.11	2.42
street, 0dB	1.73	2.15	1.63	1.98
street, 10dB	2.58	2.81	2.30	2.49
white, 0dB	1.35	1.84	1.24	1.95
white, 10dB	2.53	3.04	1.94	2.51
colored, 0dB	1.45	2.03	1.51	2.00
colored, 10dB	2.62	2.89	2.16	2.54

Table 7.2 Comparison between informal MOS and PESQ scores: second subjective test

	Informal MOS scores				PESQ scores			
Noise type and input SNR	Noisy signal	KFSFMC	DUKFPPF	Vir99	Noisy signal	KFSFMC	DUKFPPF	Vir99
babble, 0 dB	1.48	1.90	1.82	1.66	1.22	1.77	1.60	1.39
babble, 10 dB	2.45	2.72	2.82	2.58	1.88	2.4	2.39	2.05
car, 0 dB	1.57	1.91	1.79	1.71	1.42	1.88	1.81	1.65
car, 10 dB	2.20	2.70	2.85	2.36	2.11	2.42	2.58	2.25
street, 0 dB	1.73	2.18	2.05	1.92	1.63	1.98	1.91	1.83
street, 10 dB	2.50	2.80	2.88	2.65	2.30	2.49	2.69	2.39
white, 0 dB	1.36	1.79	1.70	1.55	1.24	1.95	1.75	1.48
white, 10 dB	2.51	3.01	3.10	2.64	1.94	2.51	2.57	2.17
colored, 0 dB	1.42	2.05	1.97	1.70	1.51	2.00	1.88	1.76
colored, 10 dB	2.60	2.89	2.96	2.77	2.16	2.54	2.65	2.37

Chapter 8 Conclusions and Future Work

8.1 Conclusions

This thesis focused on Kalman filter based speech enhancement approaches combined with masking properties of human auditory systems. It should also be noted that no voice activity detection is required in the proposed methods. Chapter 3 derived a modified Kalman filter constrained from the masking properties. The derivation gave a theoretical explanation for the feasibility of combining masking properties with the Kalman filter. A true perceptually constrained Kalman filtering algorithm for both white and colored noise was proposed in this chapter. However the algorithm requires a large amount of computations not only because the matrix inverse is calculated at each time step, but also a nonlinear optimization problem has to be solved in each step. Then in Chapter 4, to reduce the computaional cost, some heuristic methods were proposed. At first, the frequency domain masking level was used as a hard threshold to reshape the Kalman filtered signal. This method was used to enhance signal degraded with white noise and has proven to be a feasible solution. Then a post-filter on frame-by-frame basis was designed from the masking level where both time-domain and frequency domain masking properties are taken into account. The goal of the masking is to make the estimate state error energy smaller than the threshold. A soft threshold was proposed in the post-filter. The thresholding was designed to make a tradeoff between the amount of signal distortion and residual noise. Since the computation of the masking threshold is in the frequency domain and the Kalman filter is in time domain, the computation cost because of the switching between the two domains was then proposed to be reduced by the method described in

Chapter 5, with the wavelet domain Kalman filter combined with a perceptual post-filter, where the masking threshold is calculated in the wavelet packet domain. The multi-resolution property of human auditory systems can be mapped very well to the wavelet packet transform domain. By constructing the state-space model in the wavelet domain, the output of the wavelet Kalman filtering algorithm is directly used by the post-filter without any transform, reducing the complexity compared to a pure time domain Kalman filter approach. In Chapter 6, a nonlinear speech model was studied. Since the speech production cannot be accounted for by a linear model, a neural network is used to map such a nonlinear model in this chapter. Thus the state-space model for speech enhancement is a nonlinear one. To solve the nonlinear problem, a dual UKF is applied. By combining the masking properties, both a perceptually constrained dual UKF (which is time-consuming) and a cost-efficient dual UKF with a perceptual post-filter are proposed.

Overall, the simulation results have shown that the methods combining the Kalman filter and masking properties can produce promising results from the point of view of PESQ scores. Some informal subjective tests and results were described in Chapter 7, further confirming the potential of the algorithms introduced in this thesis.

8.2 Future work

The algorithms proposed in the thesis can be easily extended to wideband speech signal processing, which is becoming increasingly used. Also, this thesis only addressed the problem of additive noise, with types of noise having relatively slow varying statistics. For the case of highly non-stationary noise such as click noise or impulse noise, then other speech enhancement algorithms may be more suitable. Moreover, the algorithms to enhance speech degraded by reverberation, echo or clipping were not considered in this thesis. A system designed to enhance

speech under all these conditions would probably need to classify the noise type and use different algorithms based on the classification decision.

Appendix

A. Simulation Results of Chapter 3

Table A.1 PESQ scores obtained by different methods with white noise signal

Female1.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM
-5	1.03	0.86	1.09	0.89	1.61
0	1.22	1.30	1.44	1.36	1.96
5	1.53	1.73	1.86	1.75	2.29
10	1.92	2.07	2.22	2.12	2.54
15	2.29	2.32	2.56	2.34	2.87

Female2.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM
-5	1.16	0.79	1.28	0.82	1.63
0	1.38	1.16	1.62	1.22	1.98
5	1.69	1.57	1.98	1.60	2.24
10	2.04	1.85	2.31	1.87	2.54
15	2.38	2.02	2.64	2.05	2.82

Female3.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM
-5	0.98	0.67	1.12	0.71	1.63
0	1.24	1.08	1.48	1.11	1.96
5	1.58	1.53	1.83	1.57	2.28
10	1.94	1.84	2.17	1.89	2.62
15	2.27	2.04	2.52	2.11	2.80

Male1.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM
-5	1.16	0.47	1.19	0.48	1.62
0	1.30	0.80	1.37	0.85	1.89
5	1.51	1.26	1.61	1.29	2.18
10	1.81	1.59	2.03	1.64	2.48
15	2.15	1.84	2.34	1.88	2.75

Male2.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM
-5	1.24	0.74	1.47	0.77	1.89
0	1.66	1.02	1.68	1.08	2.10
5	1.82	1.39	1.83	1.43	2.24
10	1.99	1.69	2.12	1.70	2.46
15	2.24	1.89	2.41	1.91	2.73

Male3.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM
-5	1.44	0.48	1.48	0.52	1.62
0	1.57	0.74	1.64	0.81	1.89
5	1.72	1.18	1.77	1.20	2.23
10	1.96	1.51	2.10	1.54	2.49
15	2.27	1.71	2.46	1.71	2.75

Table A.2 PESQ scores obtained by different methods with some colored noise signals

Female1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	0.89	0.76	1.09	0.78	0.76	1.51
0	1.32	1.17	1.39	1.23	1.21	1.84
5	1.61	1.65	1.81	1.70	1.69	2.14
10	1.97	2.02	2.17	2.08	2.05	2.45
15	2.34	2.32	2.52	2.35	2.34	2.73

Female1.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.06	0.93	1.12	0.93	0.95	1.45
0	1.33	1.39	1.53	1.46	1.40	1.81
5	1.66	1.81	1.96	1.85	1.85	2.16
10	2.04	2.10	2.35	2.13	2.10	2.39
15	2.40	2.33	2.63	2.33	2.38	2.66

Female1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.19	1.16	1.30	1.22	1.22	1.58
0	1.49	1.60	1.72	1.60	1.65	1.90
5	1.85	1.96	2.11	2.00	2.02	2.18
10	2.22	2.23	2.35	2.24	2.30	2.48
15	2.56	2.41	2.62	2.42	2.45	2.70

Female1.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.29	1.11	1.37	1.12	1.17	1.64
0	1.54	1.57	1.71	1.61	1.65	1.99
5	1.85	1.95	2.05	2.00	1.99	2.24
10	2.20	2.23	2.49	2.28	2.26	2.51
15	2.54	2.45	2.71	2.52	2.50	2.85

Female2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.07	0.70	1.13	0.74	0.74	1.57
0	1.44	1.10	1.50	1.10	1.17	1.89
5	1.75	1.51	1.92	1.52	1.54	2.26
10	2.10	1.82	2.27	1.85	1.84	2.56
15	2.45	2.01	2.60	2.02	2.04	2.77

Female2.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.08	0.91	1.21	0.91	0.98	1.45
0	1.40	1.21	1.63	1.23	1.27	1.74
5	1.75	1.63	1.96	1.66	1.64	2.20
10	2.12	1.88	2.20	1.94	1.95	2.45
15	2.48	2.05	2.49	2.12	2.08	2.77

Female2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.25	1.03	1.34	1.05	1.04	1.57
0	1.56	1.38	1.68	1.42	1.40	1.82
5	1.92	1.74	2.98	1.80	1.74	2.15
10	2.28	1.97	2.30	2.00	1.97	2.43
15	2.61	2.11	2.57	2.12	2.16	2.76

Female2.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.32	0.91	1.44	0.94	0.94	1.68
0	1.59	1.34	1.73	1.38	1.40	1.98
5	1.93	1.70	2.10	1.73	1.74	2.29
10	2.29	1.94	2.43	2.00	1.97	2.51
15	2.63	2.09	2.71	2.10	2.13	2.78

Female3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	0.95	0.62	0.86	0.68	0.66	1.45
0	1.22	0.96	1.25	1.01	0.99	1.78
5	1.54	1.32	1.65	1.35	1.37	2.14
10	1.88	1.65	2.05	1.68	1.69	2.46
15	2.23	1.95	2.42	2.00	1.96	2.73

Female3.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.13	0.71	1.24	0.72	0.75	1.62
0	1.42	1.20	1.65	1.24	1.27	1.88
5	1.77	1.65	2.02	1.68	1.67	2.22
10	2.11	1.91	2.25	1.97	1.95	2.48
15	2.44	2.12	2.55	2.15	2.16	2.72

Female3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.34	1.06	1.43	1.10	1.11	1.63
0	1.63	1.47	1.83	1.51	1.49	2.00
5	1.97	1.79	2.18	1.79	1.81	2.31
10	2.30	2.04	2.39	2.04	2.06	2.55
15	2.62	2.21	2.69	2.26	2.22	2.77

Female3.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.26	0.98	1.42	1.03	1.05	1.79
0	1.51	1.43	1.76	1.48	1.43	2.01
5	1.84	1.75	2.07	1.80	1.80	2.34
10	2.16	1.94	2.37	2.01	2.00	2.61
15	2.49	2.12	2.67	2.17	2.12	2.83

Male1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	0.82	0.26	1.11	0.29	0.32	1.63
0	1.18	0.54	1.34	0.59	0.61	1.63
5	1.56	0.94	1.65	0.94	0.96	1.97
10	1.82	1.41	1.83	1.44	1.43	2.34
15	2.17	1.77	2.22	1.77	1.81	2.66

Male1.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.36	0.69	1.42	0.71	0.76	1.89
0	1.55	0.98	1.63	1.05	1.04	2.12
5	1.81	1.37	1.93	1.43	1.44	2.37
10	2.10	1.69	2.19	1.70	1.76	2.51
15	2.41	1.95	2.41	2.02	2.02	2.74

Male1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.50	0.88	1.50	0.89	0.92	1.97
0	1.74	1.19	1.79	1.25	1.20	2.25
5	1.99	1.55	2.11	1.62	1.57	2.46
10	2.29	1.80	2.43	1.86	1.80	2.62
15	2.60	2.00	2.71	2.05	2.05	2.85

Male1.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.37	0.60	1.40	0.67	0.62	1.86
0	1.55	0.98	1.65	1.01	1.06	2.05
5	1.79	1.39	1.95	1.44	1.42	2.27
10	2.08	1.68	2.22	1.75	1.71	2.51
15	2.40	1.93	2.51	1.96	1.97	2.74

Male2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	0.72	0.55	1.01	0.60	0.62	1.29
0	1.57	0.82	1.57	0.88	0.86	1.83
5	1.78	1.12	1.79	1.17	1.13	2.06
10	1.99	1.45	2.05	1.46	1.48	2.33
15	2.25	1.73	2.37	1.73	1.73	2.60

Male2.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.79	0.82	1.66	0.89	0.87	1.87
0	1.82	1.18	1.89	1.23	1.19	2.15
5	2.00	1.53	2.00	1.56	1.58	2.31
10	2.21	1.78	2.28	1.85	1.82	2.49
15	2.48	1.94	2.57	2.01	1.94	2.77

Male2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.68	0.96	1.73	0.96	1.01	2.00
0	1.92	1.31	1.92	1.36	1.32	2.19
5	2.12	1.65	2.16	1.65	1.66	2.42
10	2.38	1.87	2.42	1.88	1.87	2.59
15	2.66	2.08	2.72	2.11	2.13	2.83

Male2.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.71	0.89	1.79	0.95	0.96	2.11
0	1.87	1.22	1.96	1.28	1.28	2.22
5	2.04	1.56	2.05	1.62	1.59	2.37
10	2.25	1.82	2.33	1.86	1.84	2.57
15	2.50	1.98	2.64	2.00	2.02	2.80

Male3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	0.70	0.43	1.27	0.45	0.44	1.64
0	1.59	0.67	1.66	0.67	0.67	1.81
5	1.78	1.06	1.81	1.08	1.10	2.12
10	2.02	1.32	2.07	1.37	1.38	2.38
15	2.33	1.59	2.46	1.64	1.63	2.70

Male3.wav+car.wav

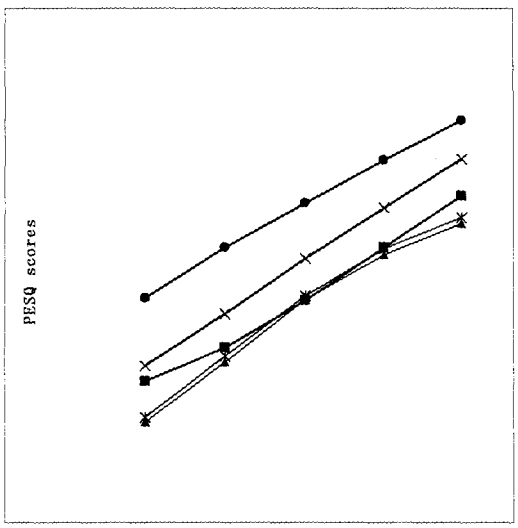
Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.43	0.63	1.45	0.65	0.64	1.99
0	1.64	0.90	1.71	0.93	0.90	2.10
5	1.92	1.33	1.95	1.35	1.35	2.30
10	2.15	1.61	2.26	1.65	1.64	2.44
15	2.44	1.79	2.55	1.79	1.82	2.65

Male3.wav+street.wav

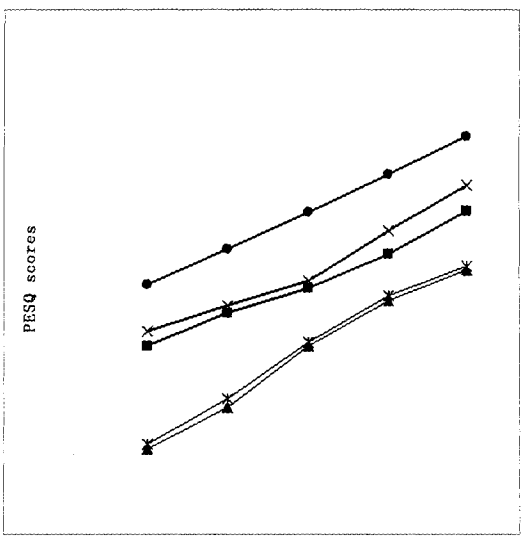
Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.53	0.80	1.59	0.84	0.80	1.85
0	1.81	1.09	1.87	1.10	1.12	2.15
5	2.03	1.45	2.09	1.49	1.49	2.30
10	2.31	1.69	2.41	1.74	1.69	2.53
15	2.62	1.85	2.71	1.91	1.86	2.80

Male3.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM
-5	1.64	0.69	1.70	0.77	0.70	1.93
0	1.81	1.04	1.86	1.10	1.11	2.18
5	2.02	1.39	2.11	1.42	1.44	2.33
10	2.27	1.64	2.35	1.67	1.71	2.53
15	2.55	1.79	2.62	1.85	1.80	2.78



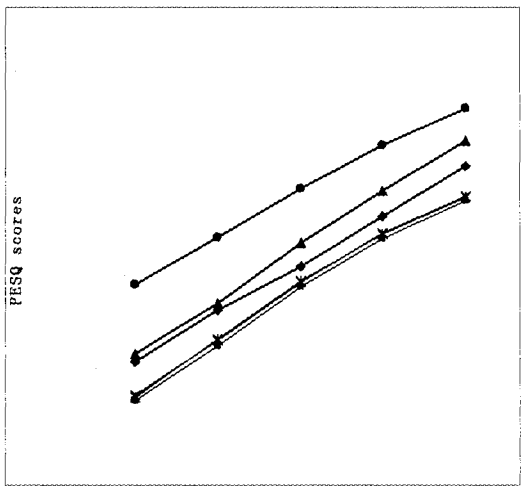
(a) female speech with white noise



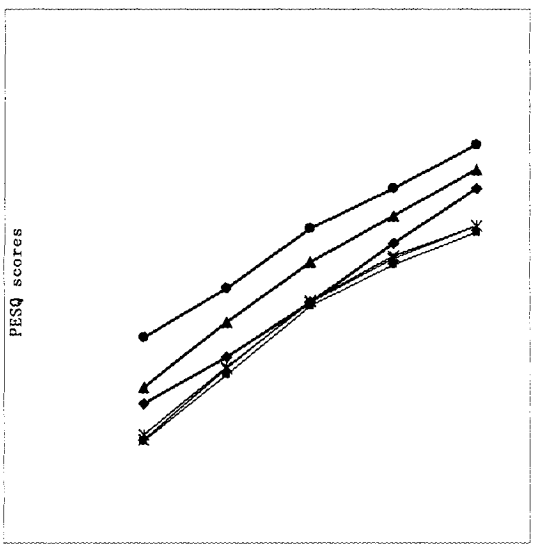
(b) male speech with white noise

¦ : noisy speech signal, ? : traditional spectral subtraction method, ×: perceptually weighted spectral subtraction method [Vir99], ✳: Kalman filtering method [Gab01], •: truly constrained Kalman filter with masking properties(CKFSFM).

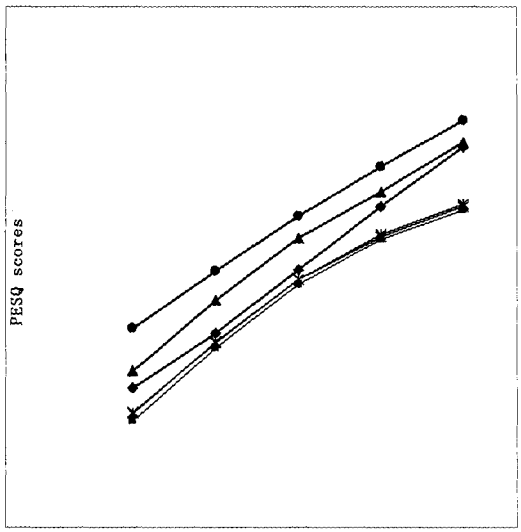
Fig. A.1 Average PESQ scores of different methods for white noise case



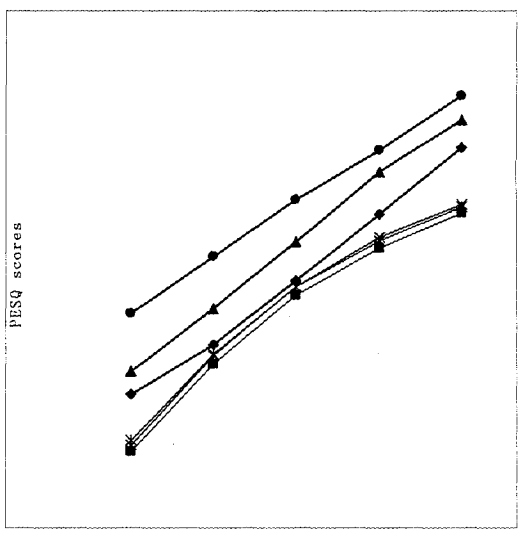
(a) female speech with babble noise



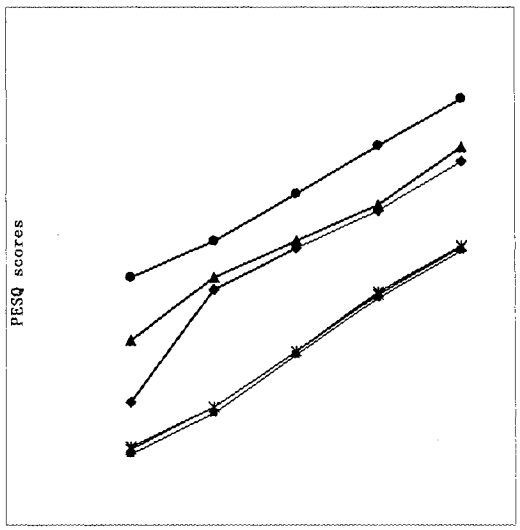
(b) female speech with car noise



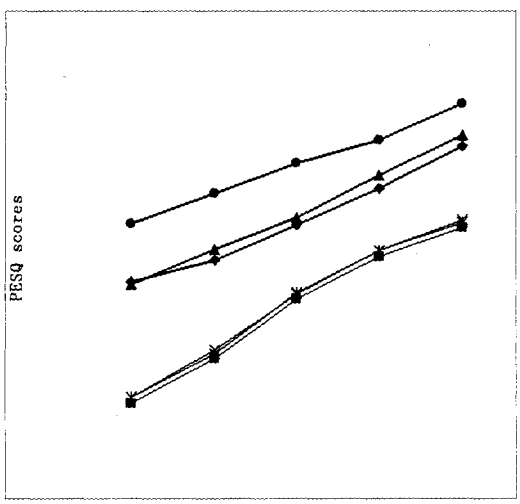
(c) female speech with street noise



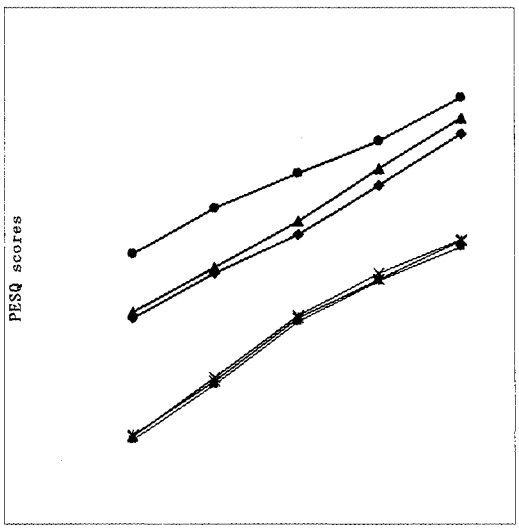
(d) female speech with generated colored noise



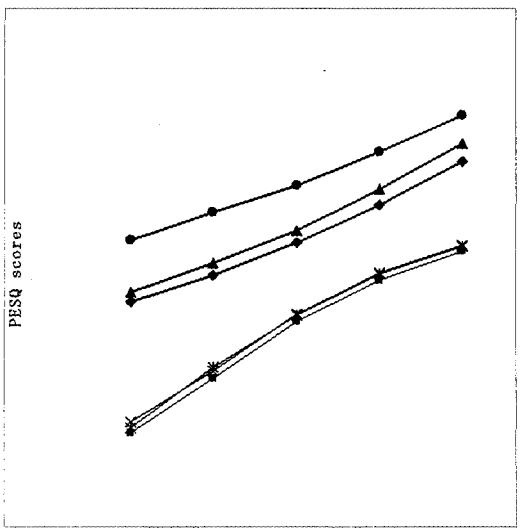
(e) male speech with babble noise



(f) male speech with car noise



(g) male speech with street noise



(h) male speech with generated colored noise

◆: noisy speech signal, ¦ : traditional spectral subtraction method, ? : perceptually weighted spectral subtraction method [Vir99], ×: KLT method [Pop98], ✷: [Mit00], •: truly constrained Kalman filter with masking properties(CKFSFM).

Fig. A.2 Average PESQ scores of different methods for colored noise case

B. Simulation Results of Chapter 4

Table B.1 PESQ scores obtained by different methods with white noise signal

Female1.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	1.03	0.86	1.09	0.89	1.61	1.41
0	1.22	1.30	1.44	1.36	1.96	1.69
5	1.53	1.73	1.86	1.75	2.29	2.13
10	1.92	2.07	2.22	2.12	2.54	2.40
15	2.29	2.32	2.56	2.34	2.87	2.64

Female2.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	1.16	0.79	1.28	0.82	1.63	1.40
0	1.38	1.16	1.62	1.22	1.98	1.76
5	1.69	1.57	1.98	1.60	2.24	2.15
10	2.04	1.85	2.31	1.87	2.54	2.44
15	2.38	2.02	2.64	2.05	2.82	2.69

Female3.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	0.98	0.67	1.12	0.71	1.63	1.42
0	1.24	1.08	1.48	1.11	1.96	1.87
5	1.58	1.53	1.83	1.57	2.28	2.13
10	1.94	1.84	2.17	1.89	2.62	2.38
15	2.27	2.04	2.52	2.11	2.80	2.63

Male1.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	1.16	0.47	1.19	0.48	1.62	1.42
0	1.30	0.80	1.37	0.85	1.89	1.70
5	1.51	1.26	1.61	1.29	2.18	2.00
10	1.81	1.59	2.03	1.64	2.48	2.23
15	2.15	1.84	2.34	1.88	2.75	2.54

Male2.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	1.24	0.74	1.47	0.77	1.89	1.68
0	1.66	1.02	1.68	1.08	2.10	1.85
5	1.82	1.39	1.83	1.43	2.24	2.08
10	1.99	1.69	2.12	1.70	2.46	2.28
15	2.24	1.89	2.41	1.91	2.73	2.52

Male3.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM
-5	1.44	0.48	1.48	0.52	1.62	1.56
0	1.57	0.74	1.64	0.81	1.89	1.75
5	1.72	1.18	1.77	1.20	2.23	1.90
10	1.96	1.51	2.10	1.54	2.49	2.33
15	2.27	1.71	2.46	1.71	2.75	2.61

Table B.2 PESQ scores obtained by different methods with white noise signal

Female1.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	1.03	0.86	1.09	0.89	1.61	1.41	1.53
0	1.22	1.30	1.44	1.36	1.96	1.69	1.94
5	1.53	1.73	1.86	1.75	2.29	2.13	2.25
10	1.92	2.07	2.22	2.12	2.54	2.40	2.52
15	2.29	2.32	2.56	2.34	2.87	2.64	2.78

Female2.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	1.16	0.79	1.28	0.82	1.63	1.40	1.55
0	1.38	1.16	1.62	1.22	1.98	1.76	1.92
5	1.69	1.57	1.98	1.60	2.24	2.15	2.23
10	2.04	1.85	2.31	1.87	2.54	2.44	2.53
15	2.38	2.02	2.64	2.05	2.82	2.69	2.78

Female3.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	0.98	0.67	1.12	0.71	1.63	1.42	1.59
0	1.24	1.08	1.48	1.11	1.96	1.87	1.95
5	1.58	1.53	1.83	1.57	2.28	2.13	2.25
10	1.94	1.84	2.17	1.89	2.62	2.38	2.51
15	2.27	2.04	2.52	2.11	2.80	2.63	2.79

Male1.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	1.16	0.47	1.19	0.48	1.62	1.42	1.57
0	1.30	0.80	1.37	0.85	1.89	1.70	1.87
5	1.51	1.26	1.61	1.29	2.18	2.00	2.15
10	1.81	1.59	2.03	1.64	2.48	2.23	2.44
15	2.15	1.84	2.34	1.88	2.75	2.54	2.71

Male2.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	1.24	0.74	1.47	0.77	1.89	1.68	1.83
0	1.66	1.02	1.68	1.08	2.10	1.85	2.03
5	1.82	1.39	1.83	1.43	2.24	2.08	2.22
10	1.99	1.69	2.12	1.70	2.46	2.28	2.43
15	2.24	1.89	2.41	1.91	2.73	2.52	2.68

Male3.wav+white.wav

Input SNR (dB)	Noisy Speech	Spectral Sub.	Vir99	Gab01	CKFSFM	KFSM	KFSFMW
-5	1.44	0.48	1.48	0.52	1.62	1.56	1.60
0	1.57	0.74	1.64	0.81	1.89	1.75	1.87
5	1.72	1.18	1.77	1.20	2.23	1.90	2.16
10	1.96	1.51	2.10	1.54	2.49	2.33	2.45
15	2.27	1.71	2.46	1.71	2.75	2.61	2.73

Table B.3 PESQ scores obtained by different methods with some colored noise signals

Female1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	0.89	0.76	1.09	0.78	0.76	1.51	1.50
0	1.32	1.17	1.39	1.23	1.21	1.84	1.83
5	1.61	1.65	1.81	1.70	1.69	2.14	2.13
10	1.97	2.02	2.17	2.08	2.05	2.45	2.44
15	2.34	2.32	2.52	2.35	2.34	2.73	2.72

Female1.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.06	0.93	1.12	0.93	0.95	1.45	1.39
0	1.33	1.39	1.53	1.46	1.40	1.81	1.75
5	1.66	1.81	1.96	1.85	1.85	2.16	2.08
10	2.04	2.10	2.35	2.13	2.10	2.39	2.38
15	2.40	2.33	2.63	2.33	2.38	2.66	2.61

Female1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.19	1.16	1.30	1.22	1.22	1.58	1.51
0	1.49	1.60	1.72	1.60	1.65	1.90	1.86
5	1.85	1.96	2.11	2.00	2.02	2.18	2.17
10	2.22	2.23	2.35	2.24	2.30	2.48	2.43
15	2.56	2.41	2.62	2.42	2.45	2.70	2.69

Female1.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.29	1.11	1.37	1.12	1.17	1.64	1.63
0	1.54	1.57	1.71	1.61	1.65	1.99	1.92
5	1.85	1.95	2.05	2.00	1.99	2.24	2.20
10	2.20	2.23	2.49	2.28	2.26	2.51	2.50
15	2.54	2.45	2.71	2.52	2.50	2.85	2.78

Female2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.07	0.70	1.13	0.74	0.74	1.57	1.49
0	1.44	1.10	1.50	1.10	1.17	1.89	1.86
5	1.75	1.51	1.92	1.52	1.54	2.26	2.19
10	2.10	1.82	2.27	1.85	1.84	2.56	2.48
15	2.45	2.01	2.60	2.02	2.04	2.77	2.75

Female2.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.08	0.91	1.21	0.91	0.98	1.45	1.38
0	1.40	1.21	1.63	1.23	1.27	1.74	1.70
5	1.75	1.63	1.96	1.66	1.64	2.20	2.22
10	2.12	1.88	2.20	1.94	1.95	2.45	2.32
15	2.48	2.05	2.49	2.12	2.08	2.77	2.60

Female2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.25	1.03	1.34	1.05	1.04	1.57	1.49
0	1.56	1.38	1.68	1.42	1.40	1.82	1.82
5	1.92	1.74	2.98	1.80	1.74	2.15	2.13
10	2.28	1.97	2.30	2.00	1.97	2.43	2.41
15	2.61	2.11	2.57	2.12	2.16	2.76	2.68

Female2.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.32	0.91	1.44	0.94	0.94	1.68	1.61
0	1.59	1.34	1.73	1.38	1.40	1.98	1.92
5	1.93	1.70	2.10	1.73	1.74	2.29	2.23
10	2.29	1.94	2.43	2.00	1.97	2.51	2.51
15	2.63	2.09	2.71	2.10	2.13	2.78	2.76

Female3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	0.95	0.62	0.86	0.68	0.66	1.45	1.40
0	1.22	0.96	1.25	1.01	0.99	1.78	1.77
5	1.54	1.32	1.65	1.35	1.37	2.14	2.08
10	1.88	1.65	2.05	1.68	1.69	2.46	2.40
15	2.23	1.95	2.42	2.00	1.96	2.73	2.72

Female3.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.13	0.71	1.24	0.72	0.75	1.62	1.56
0	1.42	1.20	1.65	1.24	1.27	1.88	1.88
5	1.77	1.65	2.02	1.68	1.67	2.22	2.16
10	2.11	1.91	2.25	1.97	1.95	2.48	2.42
15	2.44	2.12	2.55	2.15	2.16	2.72	2.68

Female3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.34	1.06	1.43	1.10	1.11	1.63	1.61
0	1.63	1.47	1.83	1.51	1.49	2.00	1.98
5	1.97	1.79	2.18	1.79	1.81	2.31	2.24
10	2.30	2.04	2.39	2.04	2.06	2.55	2.49
15	2.62	2.21	2.69	2.26	2.22	2.77	2.76

Female3.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.26	0.98	1.42	1.03	1.05	1.79	1.74
0	1.51	1.43	1.76	1.48	1.43	2.01	2.00
5	1.84	1.75	2.07	1.80	1.80	2.34	2.27
10	2.16	1.94	2.37	2.01	2.00	2.61	2.54
15	2.49	2.12	2.67	2.17	2.12	2.83	2.80

Male1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	0.82	0.26	1.11	0.29	0.32	1.63	1.58
0	1.18	0.54	1.34	0.59	0.61	1.63	1.61
5	1.56	0.94	1.65	0.94	0.96	1.97	1.91
10	1.82	1.41	1.83	1.44	1.43	2.34	2.28
15	2.17	1.77	2.22	1.77	1.81	2.66	2.61

Male1.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.36	0.69	1.42	0.71	0.76	1.89	1.81
0	1.55	0.98	1.63	1.05	1.04	2.12	2.08
5	1.81	1.37	1.93	1.43	1.44	2.37	2.30
10	2.10	1.69	2.19	1.70	1.76	2.51	2.49
15	2.41	1.95	2.41	2.02	2.02	2.74	2.70

Male1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.50	0.88	1.50	0.89	0.92	1.97	1.94
0	1.74	1.19	1.79	1.25	1.20	2.25	2.18
5	1.99	1.55	2.11	1.62	1.57	2.46	2.40
10	2.29	1.80	2.43	1.86	1.80	2.62	2.60
15	2.60	2.00	2.71	2.05	2.05	2.85	2.82

Male1.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.37	0.60	1.40	0.67	0.62	1.86	1.82
0	1.55	0.98	1.65	1.01	1.06	2.05	2.03
5	1.79	1.39	1.95	1.44	1.42	2.27	2.24
10	2.08	1.68	2.22	1.75	1.71	2.51	2.45
15	2.40	1.93	2.51	1.96	1.97	2.74	2.68

Male2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	0.72	0.55	1.01	0.60	0.62	1.29	1.20
0	1.57	0.82	1.57	0.88	0.86	1.83	1.79
5	1.78	1.12	1.79	1.17	1.13	2.06	2.06
10	1.99	1.45	2.05	1.46	1.48	2.33	2.33
15	2.25	1.73	2.37	1.73	1.73	2.60	2.59

Male2.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.79	0.82	1.66	0.89	0.87	1.87	1.96
0	1.82	1.18	1.89	1.23	1.19	2.15	2.14
5	2.00	1.53	2.00	1.56	1.58	2.31	2.31
10	2.21	1.78	2.28	1.85	1.82	2.49	2.46
15	2.48	1.94	2.57	2.01	1.94	2.77	2.65

Male2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.68	0.96	1.73	0.96	1.01	2.00	1.95
0	1.92	1.31	1.92	1.36	1.32	2.19	2.18
5	2.12	1.65	2.16	1.65	1.66	2.42	2.35
10	2.38	1.87	2.42	1.88	1.87	2.59	2.57
15	2.66	2.08	2.72	2.11	2.13	2.83	2.79

Male2.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.71	0.89	1.79	0.95	0.96	2.11	2.02
0	1.87	1.22	1.96	1.28	1.28	2.22	2.20
5	2.04	1.56	2.05	1.62	1.59	2.37	2.36
10	2.25	1.82	2.33	1.86	1.84	2.57	2.52
15	2.50	1.98	2.64	2.00	2.02	2.80	2.73

Male3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	0.70	0.43	1.27	0.45	0.44	1.64	1.58
0	1.59	0.67	1.66	0.67	0.67	1.81	1.77
5	1.78	1.06	1.81	1.08	1.10	2.12	2.11
10	2.02	1.32	2.07	1.37	1.38	2.38	2.38
15	2.33	1.59	2.46	1.64	1.63	2.70	2.66

Male3.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.43	0.63	1.45	0.65	0.64	1.99	1.94
0	1.64	0.90	1.71	0.93	0.90	2.10	2.03
5	1.92	1.33	1.95	1.35	1.35	2.30	2.22
10	2.15	1.61	2.26	1.65	1.64	2.44	2.42
15	2.44	1.79	2.55	1.79	1.82	2.65	2.65

Male3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.53	0.80	1.59	0.84	0.80	1.85	1.82
0	1.81	1.09	1.87	1.10	1.12	2.15	2.09
5	2.03	1.45	2.09	1.49	1.49	2.30	2.30
10	2.31	1.69	2.41	1.74	1.69	2.53	2.51
15	2.62	1.85	2.71	1.91	1.86	2.80	2.77

Male3.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	KFSFMC
-5	1.64	0.69	1.70	0.77	0.70	1.93	1.92
0	1.81	1.04	1.86	1.10	1.11	2.18	2.11
5	2.02	1.39	2.11	1.42	1.44	2.33	2.28
10	2.27	1.64	2.35	1.67	1.71	2.53	2.48
15	2.55	1.79	2.62	1.85	1.80	2.78	2.72

Table B.4 Average PESQ score gains obtained by applying the Kalman filter with a perceptual post-filter (KFSFMC) as a pre-processing step or a post-processing step to common speech codecs

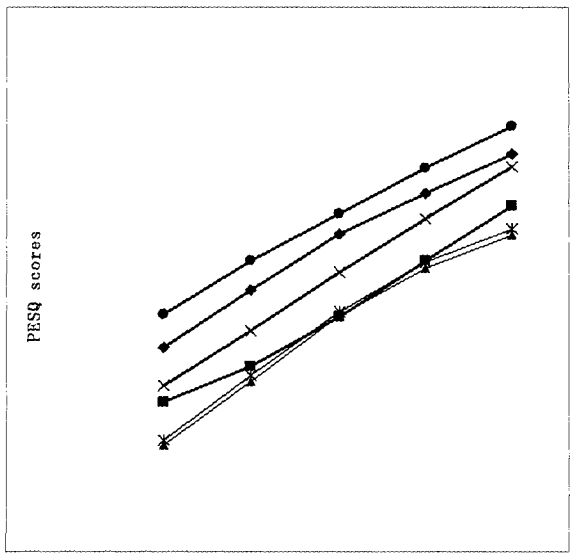
input speech SNR → / algorithm ↓	-5 dB	0 dB	5 dB	10 dB	15 dB
original noisy speech PESQ	1.26	1.54	1.80	2.11	2.43
PESQ gain with G729 codec (without enh.)	-0.11	-0.01	0.09	0.10	0.10
PESQ gain with G729 and post-processing	0.27	0.33	0.39	0.33	0.24
PESQ gain with G729 and pre-processing	0.41	0.45	0.45	0.38	0.26
PESQ gain with G723.1 codec (without enh.)	0.06	0.07	0.10	0.10	0.07
PESQ gain with G723.1 and post-processing	0.41	0.40	0.41	0.32	0.20
PESQ gain with G723.1 and pre-processing	0.44	0.46	0.44	0.36	0.20
PESQ gain with G711codec (without enh.)	0.01	0.00	0.00	0.00	0.00
PESQ gain with G711 and post-processing	0.36	0.40	0.40	0.35	0.29
PESQ gain with G711 and pre-processing	0.41	0.39	0.39	0.34	0.27

Table B.5 Details of average PESQ score gains obtained by the proposed Kalman filter with perceptual post-filter method (KFSFMC) and the perceptual spectral subtraction method [Vir99] for 5 noise types

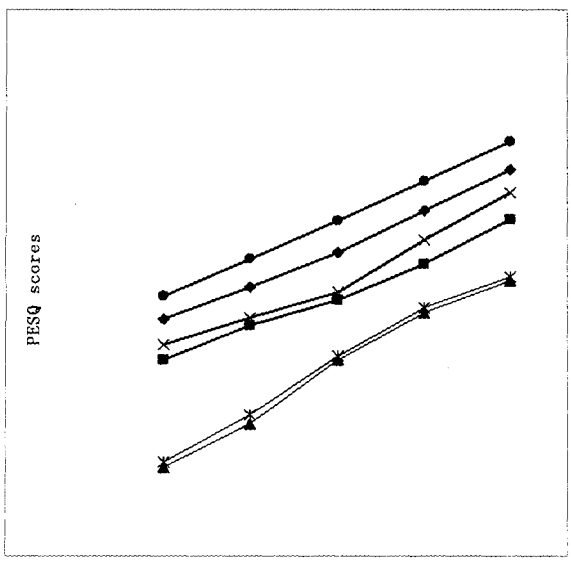
input speech SNR → / algorithm ↓	-5 dB	0 dB	5 dB	10 dB	15 dB
original noisy speech PESQ	1.26	1.54	1.80	2.11	2.43
Babble noise					
PESQ gain with KFSFMC	0.33	0.32	0.41	0.42	0.38
PESQ gain with method in [Vir99]	0.16	-0.08	0.06	0.14	0.17
Car noise					
PESQ gain with KFSFMC	0.38	0.40	0.37	0.29	0.21
PESQ gain with method in [Vir99]	0.05	0.15	0.20	0.22	0.19
Street noise					
PESQ gain with KFSFMC	0.30	0.32	0.28	0.21	0.14
PESQ gain with method in [Vir99]	0.09	0.13	0.19	0.18	0.14
White noise					
PESQ gain with KFSFMC	0.44	0.54	0.57	0.54	0.48
PESQ gain with method in [Vir99]	0.12	0.15	0.20	0.24	0.25
Generated colored noise					
PESQ gain with KFSFMC	0.36	0.39	0.35	0.29	0.23
PESQ gain with method in [Vir99]	0.09	0.16	0.19	0.20	0.18

Table B.6 PESQ score gains obtained by the proposed Kalman filter with perceptual post-filter method (KFSFMC) and the perceptual spectral subtraction method [Vir99], under a reverberation environment

speech signals ? / algorithm ?	Female1	Female2	Female3	Male1	Male2	Male3	Ave.
original reverberant speech PESQ	2.12	2.25	2.39	2.39	2.46	2.26	2.31
PESQ gain with proposed method	0.11	0.02	0.11	0.14	0.11	0.08	0.09
PESQ gain with method in [Vir99]	0.01	0.08	0.04	-0.03	-0.02	0.00	0.01



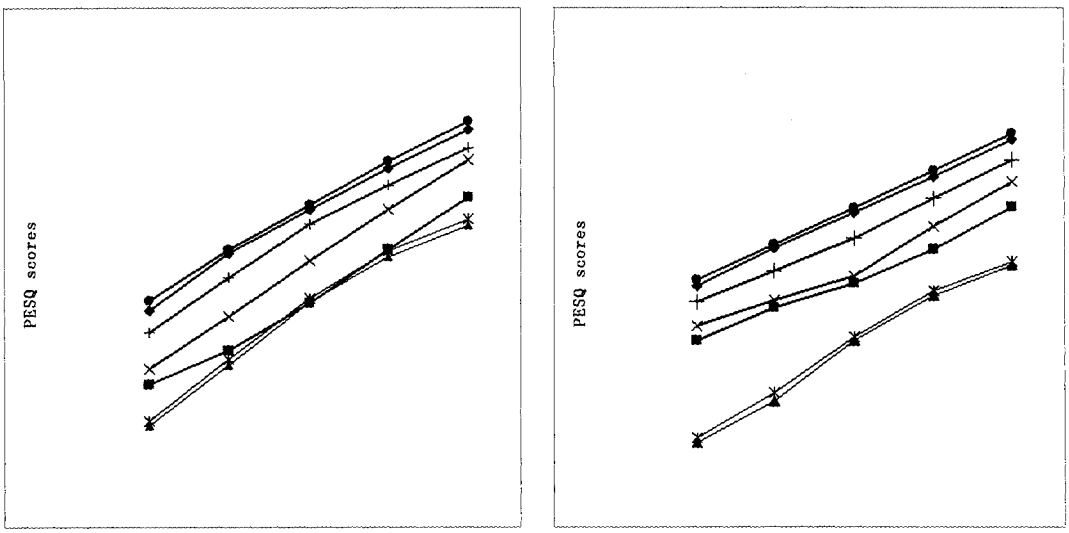
(a) Female Speech Signals



(b) Male Speech Signals

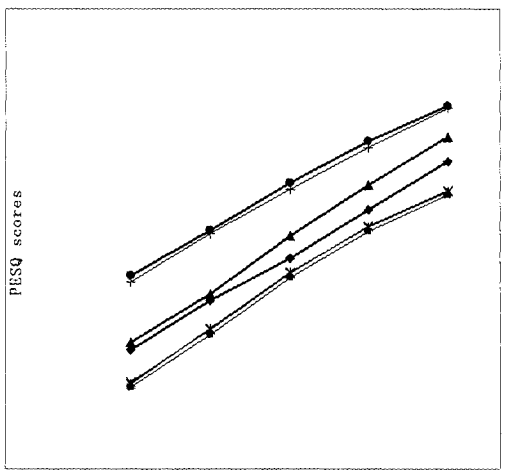
┆ : noisy speech signal, ? : traditional spectral subtraction method, ×: perceptually weighted spectral subtraction method [Vir99],∗: Kalman filtering method [Gab01], •: truly constrained Kalman filter with masking properties(CKFSFM), ◆: Kalman filter with simultaneous masking (KFSM).

Fig. B.1 Average PESQ scores of different methods

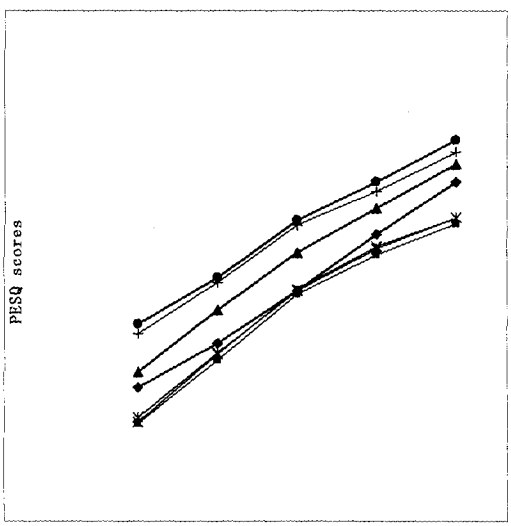


¦ : noisy speech signal, ? : traditional spectral subtraction method, ×: perceptually spectral subtraction method [Vir99],∗: Kalman filtering method [Gab01], •: truly constrained Kalman filter with masking properties(CKFSFM), ⋆: Kalman filter with perceptual post-filter for white noise (KFSFMW), +: Kalman filter with simultaneous masking (KFSM).

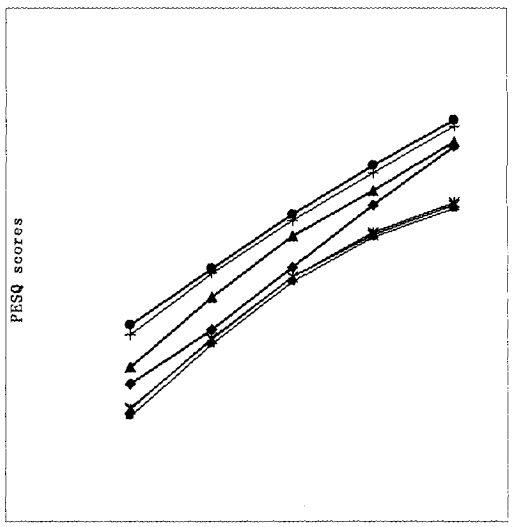
Fig. B.2 Average PESQ scores of different methods



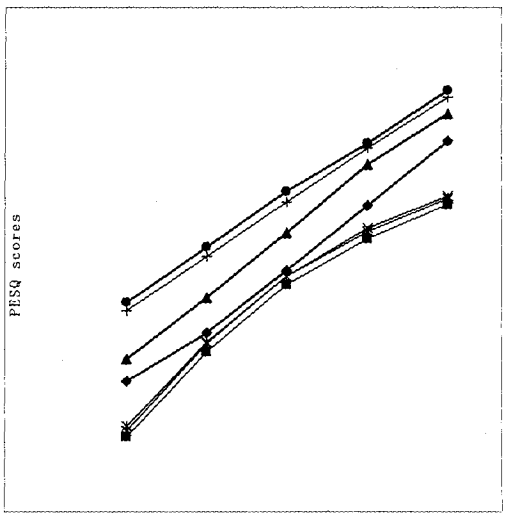
(a) female speech with babble noise



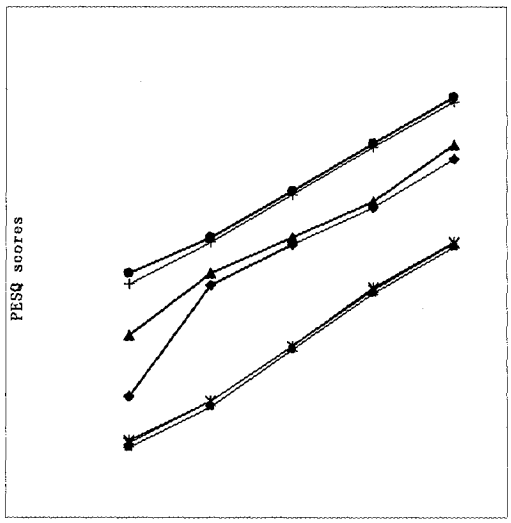
(b) female speech with car noise



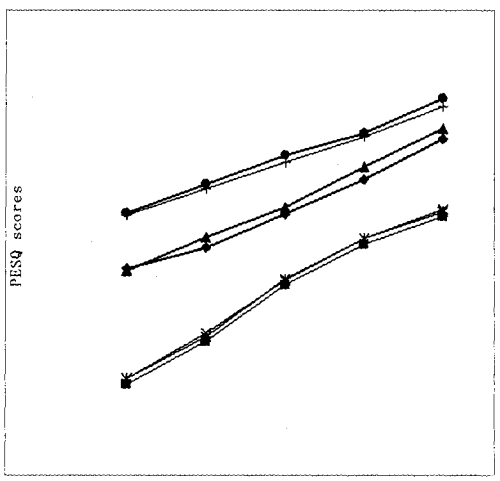
(c) female speech with street noise



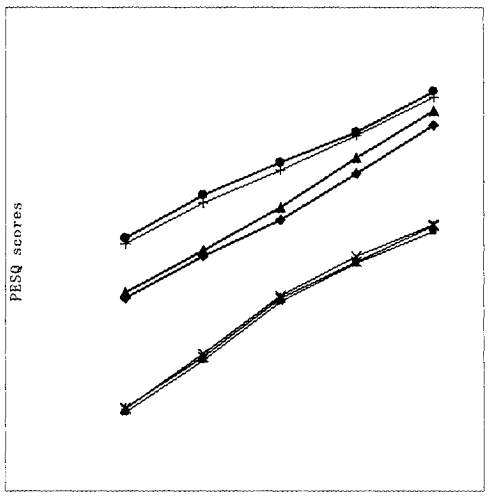
(d) female speech with generated colored noise



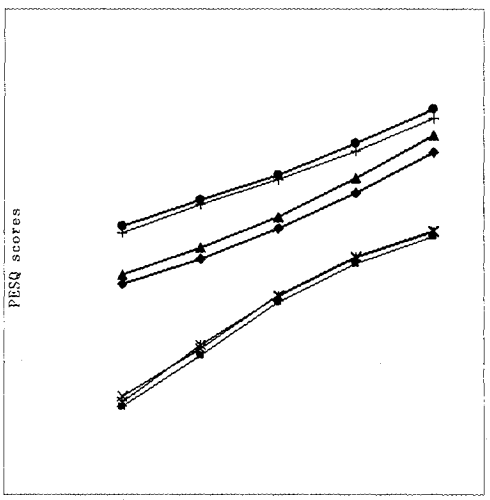
(e) male speech with babble noise



(f) male speech with car noise



(g) male speech with street noise



(h) male speech with generated colored noise

◆: noisy speech signal, ┆ : traditional spectral subtraction method, ? : perceptually weighted spectral subtraction method [Vir99], ×: KLT method [Pop98], ✳: [Mit00], •: truly constrained Kalman filter with masking properties(CKFSFM), +: Kalman filter with perceptual post-filter for colored noise (KFSFMC).

Fig. B.3 Average PESQ scores of different methods for colored noise case

C. Simulation Results of Chapter 5

Table C.1 Detailed PESQ scores obtained by different methods

Female1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	0.89	0.76	1.09	0.78	0.76	1.50	1.50
0	1.32	1.17	1.39	1.23	1.21	1.83	1.82
5	1.61	1.65	1.81	1.70	1.69	2.13	2.14
10	1.97	2.02	2.17	2.08	2.05	2.44	2.45
15	2.34	2.32	2.52	2.35	2.34	2.72	2.71

Female1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.19	1.16	1.30	1.22	1.22	1.51	1.51
0	1.49	1.60	1.72	1.60	1.65	1.86	1.85
5	1.85	1.96	2.11	2.00	2.02	2.17	2.17
10	2.22	2.23	2.35	2.24	2.30	2.43	2.44
15	2.56	2.41	2.62	2.42	2.45	2.69	2.69

Female2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.07	0.70	1.13	0.74	0.74	1.49	1.50
0	1.44	1.10	1.50	1.10	1.17	1.86	1.85
5	1.75	1.51	1.92	1.52	1.54	2.19	2.19
10	2.10	1.82	2.27	1.85	1.84	2.48	2.49
15	2.45	2.01	2.60	2.02	2.04	2.75	2.75

Female2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.25	1.03	1.34	1.05	1.04	1.49	1.49
0	1.56	1.38	1.68	1.42	1.40	1.82	1.81
5	1.92	1.74	2.98	1.80	1.74	2.13	2.13
10	2.28	1.97	2.30	2.00	1.97	2.41	2.42
15	2.61	2.11	2.57	2.12	2.16	2.68	2.68

Female3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	0.95	0.62	0.86	0.68	0.66	1.40	1.40
0	1.22	0.96	1.25	1.01	0.99	1.77	1.76
5	1.54	1.32	1.65	1.35	1.37	2.08	2.08
10	1.88	1.65	2.05	1.68	1.69	2.40	2.40
15	2.23	1.95	2.42	2.00	1.96	2.72	2.71

Female3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.34	1.06	1.43	1.10	1.11	1.61	1.61
0	1.63	1.47	1.83	1.51	1.49	1.98	1.98
5	1.97	1.79	2.18	1.79	1.81	2.24	2.25
10	2.30	2.04	2.39	2.04	2.06	2.49	2.49
15	2.62	2.21	2.69	2.26	2.22	2.76	2.76

Male1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	0.82	0.26	1.11	0.29	0.32	1.58	1.59
0	1.18	0.54	1.34	0.59	0.61	1.61	1.60
5	1.56	0.94	1.65	0.94	0.96	1.91	1.92
10	1.82	1.41	1.83	1.44	1.43	2.28	2.29
15	2.17	1.77	2.22	1.77	1.81	2.61	2.61

Male1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.50	0.88	1.50	0.89	0.92	1.94	1.94
0	1.74	1.19	1.79	1.25	1.20	2.18	2.17
5	1.99	1.55	2.11	1.62	1.57	2.40	2.40
10	2.29	1.80	2.43	1.86	1.80	2.60	2.61
15	2.60	2.00	2.71	2.05	2.05	2.82	2.81

Male2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	0.72	0.55	1.01	0.60	0.62	1.20	1.20
0	1.57	0.82	1.57	0.88	0.86	1.79	1.78
5	1.78	1.12	1.79	1.17	1.13	2.06	2.06
10	1.99	1.45	2.05	1.46	1.48	2.33	2.33
15	2.25	1.73	2.37	1.73	1.73	2.59	2.59

Male2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.68	0.96	1.73	0.96	1.01	1.95	1.96
0	1.92	1.31	1.92	1.36	1.32	2.18	2.17
5	2.12	1.65	2.16	1.65	1.66	2.35	2.36
10	2.38	1.87	2.42	1.88	1.87	2.57	2.57
15	2.66	2.08	2.72	2.11	2.13	2.79	2.79

Male3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	0.70	0.43	1.27	0.45	0.44	1.58	1.59
0	1.59	0.67	1.66	0.67	0.67	1.77	1.76
5	1.78	1.06	1.81	1.08	1.10	2.11	2.11
10	2.02	1.32	2.07	1.37	1.38	2.38	2.38
15	2.33	1.59	2.46	1.64	1.63	2.66	2.65

Male3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	KFSFMC	WKFSFM
-5	1.53	0.80	1.59	0.84	0.80	1.82	1.82
0	1.81	1.09	1.87	1.10	1.12	2.09	2.09
5	2.03	1.45	2.09	1.49	1.49	2.30	2.31
10	2.31	1.69	2.41	1.74	1.69	2.51	2.51
15	2.62	1.85	2.71	1.91	1.86	2.77	2.76

D. Simulation Results of Chapter 6

Table D.1 PESQ scores obtained by different methods with white noise signal

Female1.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	1.03	0.86	1.09	0.89	1.61	1.59
0	1.22	1.30	1.44	1.36	1.96	1.88
5	1.53	1.73	1.86	1.75	2.29	2.35
10	1.92	2.07	2.22	2.12	2.54	2.60
15	2.29	2.32	2.56	2.34	2.87	2.94

Female2.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	1.16	0.79	1.28	0.82	1.63	1.54
0	1.38	1.16	1.62	1.22	1.98	1.88
5	1.69	1.57	1.98	1.60	2.24	2.29
10	2.04	1.85	2.31	1.87	2.54	2.66
15	2.38	2.02	2.64	2.05	2.82	2.93

Female3.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	0.98	0.67	1.12	0.71	1.63	1.53
0	1.24	1.08	1.48	1.11	1.96	1.87
5	1.58	1.53	1.83	1.57	2.28	2.33
10	1.94	1.84	2.17	1.89	2.62	2.69
15	2.27	2.04	2.52	2.11	2.80	2.89

Male1.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	1.16	0.47	1.19	0.48	1.62	1.53
0	1.30	0.80	1.37	0.85	1.89	1.80
5	1.51	1.26	1.61	1.29	2.18	2.24
10	1.81	1.59	2.03	1.64	2.48	2.54
15	2.15	1.84	2.34	1.88	2.75	2.82

Male2.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	1.24	0.74	1.47	0.77	1.89	1.79
0	1.66	1.02	1.68	1.08	2.10	2.01
5	1.82	1.39	1.83	1.43	2.24	2.30
10	1.99	1.69	2.12	1.70	2.46	2.52
15	2.24	1.89	2.41	1.91	2.73	2.78

Male3.wav+white.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Gab01	CKFSFM	PCDUKF
-5	1.44	0.48	1.48	0.52	1.62	1.53
0	1.57	0.74	1.64	0.81	1.89	1.80
5	1.72	1.18	1.77	1.20	2.23	2.29
10	1.96	1.51	2.10	1.54	2.49	2.56
15	2.27	1.71	2.46	1.71	2.75	2.83

Table D.2 PESQ scores obtained by different methods with some colored noise signals

Female1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	0.89	0.76	1.09	0.78	0.76	1.51	1.50
0	1.32	1.17	1.39	1.23	1.21	1.84	1.80
5	1.61	1.65	1.81	1.70	1.69	2.14	2.20
10	1.97	2.02	2.17	2.08	2.05	2.45	2.52
15	2.34	2.32	2.52	2.35	2.34	2.73	2.79

Female1.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.06	0.93	1.12	0.93	0.95	1.45	1.34
0	1.33	1.39	1.53	1.46	1.40	1.81	1.69
5	1.66	1.81	1.96	1.85	1.85	2.16	2.22
10	2.04	2.10	2.35	2.13	2.10	2.39	2.45
15	2.40	2.33	2.63	2.33	2.38	2.66	2.73

Female1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.19	1.16	1.30	1.22	1.22	1.58	1.46
0	1.49	1.60	1.72	1.60	1.65	1.90	1.79
5	1.85	1.96	2.11	2.00	2.02	2.18	2.24
10	2.22	2.23	2.35	2.24	2.30	2.48	2.54
15	2.56	2.41	2.62	2.42	2.45	2.70	2.75

Female1.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.29	1.11	1.37	1.12	1.17	1.64	1.53
0	1.54	1.57	1.71	1.61	1.65	1.99	1.90
5	1.85	1.95	2.05	2.00	1.99	2.24	2.34
10	2.20	2.23	2.49	2.28	2.26	2.51	2.60
15	2.54	2.45	2.71	2.52	2.50	2.85	2.97

Female2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.07	0.70	1.13	0.74	0.74	1.57	1.47
0	1.44	1.10	1.50	1.10	1.17	1.89	1.80
5	1.75	1.51	1.92	1.52	1.54	2.26	2.32
10	2.10	1.82	2.27	1.85	1.84	2.56	2.63
15	2.45	2.01	2.60	2.02	2.04	2.77	2.83

Female2.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.08	0.91	1.21	0.91	0.98	1.45	1.36
0	1.40	1.21	1.63	1.23	1.27	1.74	1.69
5	1.75	1.63	1.96	1.66	1.64	2.20	2.26
10	2.12	1.88	2.20	1.94	1.95	2.45	2.52
15	2.48	2.05	2.49	2.12	2.08	2.77	2.81

Female2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.25	1.03	1.34	1.05	1.04	1.57	1.47
0	1.56	1.38	1.68	1.42	1.40	1.82	1.79
5	1.92	1.74	1.98	1.80	1.74	2.15	2.21
10	2.28	1.97	2.30	2.00	1.97	2.43	2.51
15	2.61	2.11	2.57	2.12	2.16	2.76	2.82

Female2.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.32	0.91	1.44	0.94	0.94	1.68	1.59
0	1.59	1.34	1.73	1.38	1.40	1.98	1.88
5	1.93	1.70	2.10	1.73	1.74	2.29	2.35
10	2.29	1.94	2.43	2.00	1.97	2.51	2.59
15	2.63	2.09	2.71	2.10	2.13	2.78	2.80

Female3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	0.95	0.62	0.86	0.68	0.66	1.45	1.35
0	1.22	0.96	1.25	1.01	0.99	1.78	1.68
5	1.54	1.32	1.65	1.35	1.37	2.14	2.19
10	1.88	1.65	2.05	1.68	1.69	2.46	2.57
15	2.23	1.95	2.42	2.00	1.96	2.73	2.81

Female3.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.13	0.71	1.24	0.72	0.75	1.62	1.52
0	1.42	1.20	1.65	1.24	1.27	1.88	1.78
5	1.77	1.65	2.02	1.68	1.67	2.22	2.28
10	2.11	1.91	2.25	1.97	1.95	2.48	2.54
15	2.44	2.12	2.55	2.15	2.16	2.72	2.81

Female3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.34	1.06	1.43	1.10	1.11	1.63	1.53
0	1.63	1.47	1.83	1.51	1.49	2.00	1.91
5	1.97	1.79	2.18	1.79	1.81	2.31	2.38
10	2.30	2.04	2.39	2.04	2.06	2.55	2.63
15	2.62	2.21	2.69	2.26	2.22	2.77	2.83

Female3.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.26	0.98	1.42	1.03	1.05	1.79	1.69
0	1.51	1.43	1.76	1.48	1.43	2.01	1.91
5	1.84	1.75	2.07	1.80	1.80	2.34	2.38
10	2.16	1.94	2.37	2.01	2.00	2.61	2.67
15	2.49	2.12	2.67	2.17	2.12	2.83	2.91

Male1.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	0.82	0.26	1.11	0.29	0.32	1.63	1.54
0	1.18	0.54	1.34	0.59	0.61	1.63	1.53
5	1.56	0.94	1.65	0.94	0.96	1.97	2.02
10	1.82	1.41	1.83	1.44	1.43	2.34	2.42
15	2.17	1.77	2.22	1.77	1.81	2.66	2.74

Male1.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.36	0.69	1.42	0.71	0.76	1.89	1.79
0	1.55	0.98	1.63	1.05	1.04	2.12	2.02
5	1.81	1.37	1.93	1.43	1.44	2.37	2.42
10	2.10	1.69	2.19	1.70	1.76	2.51	2.58
15	2.41	1.95	2.41	2.02	2.02	2.74	2.82

Male1.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.50	0.88	1.50	0.89	0.92	1.97	1.87
0	1.74	1.19	1.79	1.25	1.20	2.25	2.15
5	1.99	1.55	2.11	1.62	1.57	2.46	2.51
10	2.29	1.80	2.43	1.86	1.80	2.62	2.67
15	2.60	2.00	2.71	2.05	2.05	2.85	2.89

Male1.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.37	0.60	1.40	0.67	0.62	1.86	1.77
0	1.55	0.98	1.65	1.01	1.06	2.05	1.96
5	1.79	1.39	1.95	1.44	1.42	2.27	2.33
10	2.08	1.68	2.22	1.75	1.71	2.51	2.56
15	2.40	1.93	2.51	1.96	1.97	2.74	2.81

Male2.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	0.72	0.55	1.01	0.60	0.62	1.29	1.13
0	1.57	0.82	1.57	0.88	0.86	1.83	1.73
5	1.78	1.12	1.79	1.17	1.13	2.06	2.13
10	1.99	1.45	2.05	1.46	1.48	2.33	2.39
15	2.25	1.73	2.37	1.73	1.73	2.60	2.67

Male2.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.79	0.82	1.66	0.89	0.87	1.87	1.78
0	1.82	1.18	1.89	1.23	1.19	2.15	2.05
5	2.00	1.53	2.00	1.56	1.58	2.31	2.36
10	2.21	1.78	2.28	1.85	1.82	2.49	2.55
15	2.48	1.94	2.57	2.01	1.94	2.77	2.85

Male2.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.68	0.96	1.73	0.96	1.01	2.00	1.89
0	1.92	1.31	1.92	1.36	1.32	2.19	2.09
5	2.12	1.65	2.16	1.65	1.66	2.42	2.48
10	2.38	1.87	2.42	1.88	1.87	2.59	2.66
15	2.66	2.08	2.72	2.11	2.13	2.83	2.91

Male2.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.71	0.89	1.79	0.95	0.96	2.11	2.01
0	1.87	1.22	1.96	1.28	1.28	2.22	2.13
5	2.04	1.56	2.05	1.62	1.59	2.37	2.43
10	2.25	1.82	2.33	1.86	1.84	2.57	2.65
15	2.50	1.98	2.64	2.00	2.02	2.80	2.86

Male3.wav+babble.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	0.70	0.43	1.27	0.45	0.44	1.64	1.55
0	1.59	0.67	1.66	0.67	0.67	1.81	1.71
5	1.78	1.06	1.81	1.08	1.10	2.12	2.17
10	2.02	1.32	2.07	1.37	1.38	2.38	2.47
15	2.33	1.59	2.46	1.64	1.63	2.70	2.77

Male3.wav+car.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.43	0.63	1.45	0.65	0.64	1.99	1.90
0	1.64	0.90	1.71	0.93	0.90	2.10	2.01
5	1.92	1.33	1.95	1.35	1.35	2.30	2.36
10	2.15	1.61	2.26	1.65	1.64	2.44	2.52
15	2.44	1.79	2.55	1.79	1.82	2.65	2.73

Male3.wav+street.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.53	0.80	1.59	0.84	0.80	1.85	1.76
0	1.81	1.09	1.87	1.10	1.12	2.15	2.06
5	2.03	1.45	2.09	1.49	1.49	2.30	2.37
10	2.31	1.69	2.41	1.74	1.69	2.53	2.61
15	2.62	1.85	2.71	1.91	1.86	2.80	2.86

Male3.wav+colored1.wav

Input SNR (dB)	Noisy Speech	Conventional SS	Vir99	Pop98	Mit00	CKFSFM	PCDUKF
-5	1.64	0.69	1.70	0.77	0.70	1.93	1.84
0	1.81	1.04	1.86	1.10	1.11	2.18	2.08
5	2.02	1.39	2.11	1.42	1.44	2.33	2.39
10	2.27	1.64	2.35	1.67	1.71	2.53	2.60
15	2.55	1.79	2.62	1.85	1.80	2.78	2.85

Table D.3 Average PESQ scores obtained by the DUKFPPF and comparison with other algorithms

Female+white

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.05	1.56	1.49	1.55
0	1.28	1.94	1.86	1.88
5	1.60	2.24	2.30	2.33
10	1.97	2.52	2.58	2.65
15	2.32	2.78	2.85	2.92

Male+white

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.28	1.67	1.57	1.61
0	1.51	1.93	1.84	1.87
5	1.68	2.18	2.22	2.27
10	1.92	2.44	2.49	2.54
15	2.22	2.70	2.76	2.81

Female+babble

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	0.97	1.46	1.37	1.44
0	1.33	1.82	1.74	1.76
5	1.63	2.13	2.20	2.24
10	1.98	2.44	2.51	2.57
15	2.34	2.73	2.77	2.81

Female+car

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.09	1.44	1.35	1.41
0	1.38	1.78	1.67	1.72
5	1.72	2.15	2.20	2.25
10	2.09	2.37	2.45	2.50
15	2.44	2.63	2.70	2.78

Female+street

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.26	1.53	1.45	1.49
0	1.56	1.88	1.79	1.83
5	1.91	2.18	2.23	2.28
10	2.27	2.45	2.48	2.56
15	2.60	2.71	2.77	2.80

Female+colored1

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.29	1.66	1.56	1.60
0	1.55	1.95	1.85	1.90
5	1.87	2.23	2.31	2.36
10	2.21	2.52	2.58	2.62
15	2.55	2.78	2.84	2.90

Male+babble

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	0.74	1.45	1.37	1.40
0	1.45	1.72	1.63	1.66
5	1.71	2.03	2.06	2.11
10	1.94	2.33	2.37	2.43
15	2.25	2.62	2.67	2.72

Male+car

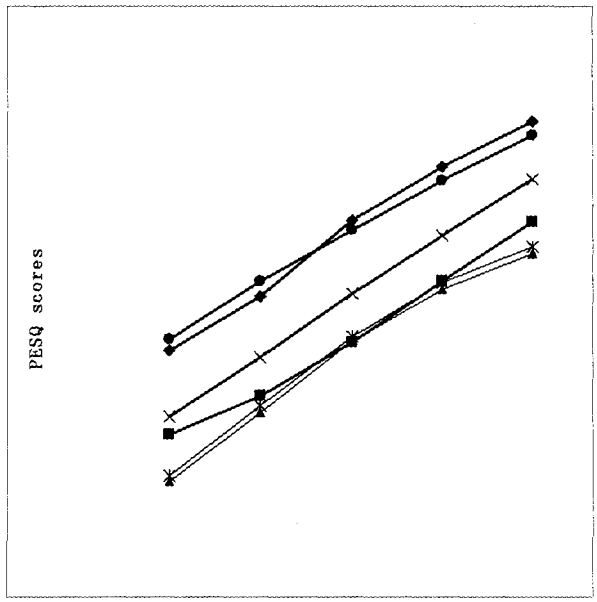
Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.53	1.90	1.75	1.82
0	1.67	2.09	1.96	2.03
5	1.91	2.28	2.30	2.38
10	2.15	2.46	2.49	2.55
15	2.44	2.67	2.71	2.80

Male+street

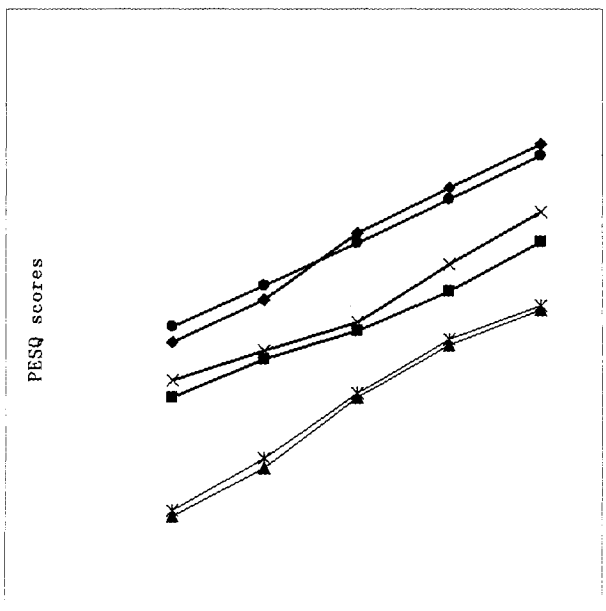
Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.57	1.90	1.81	1.84
0	1.82	2.15	2.07	2.10
5	2.05	2.35	2.38	2.45
10	2.33	2.56	2.60	2.65
15	2.63	2.79	2.82	2.89

Male+colored1

Input SNR (dB)	Noisy Speech	KFSFMC	DUKFPPF	PCDUKF
-5	1.57	1.92	1.83	1.88
0	1.74	2.11	2.02	2.06
5	1.95	2.29	2.32	2.38
10	2.20	2.48	2.55	2.60
15	2.48	2.71	2.78	2.84



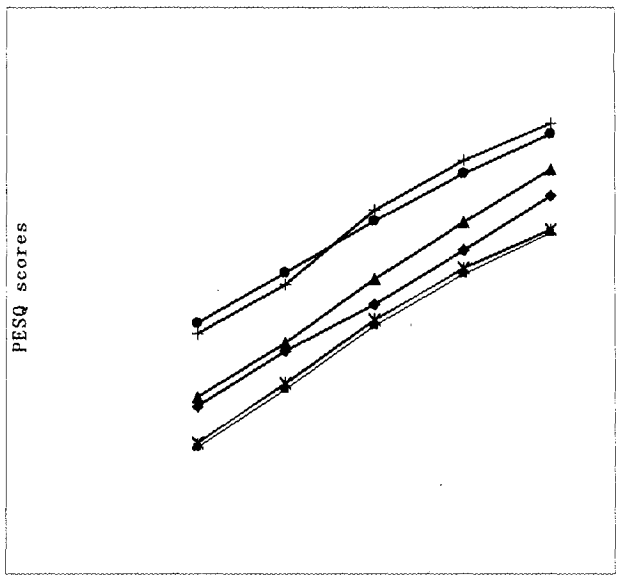
(a) female speech with white noise



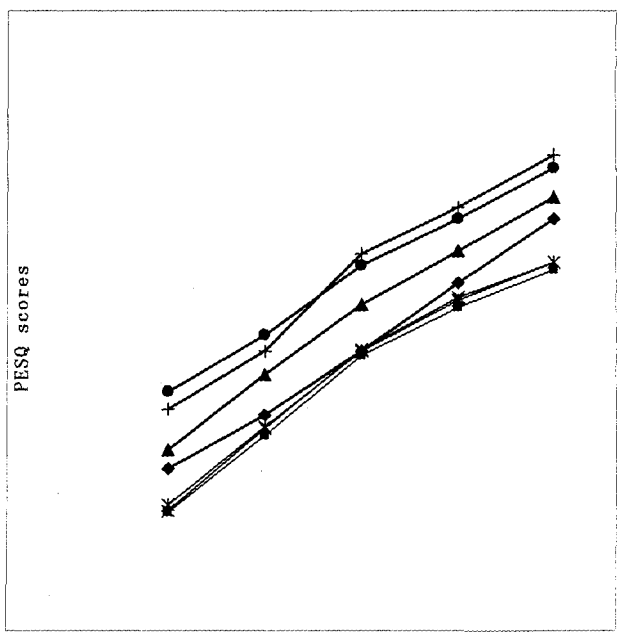
(b) male speech with white noise

¦ : noisy speech signal, ? : traditional spectral subtraction method, ×: perceptually weighted spectral subtraction method [Vir99], ∗: Kalman filtering method [Gab01], •: truly constrained Kalman filter with masking properties(CKFSFM), ◆: perceptually constrained dual UKF (PCDUKF)

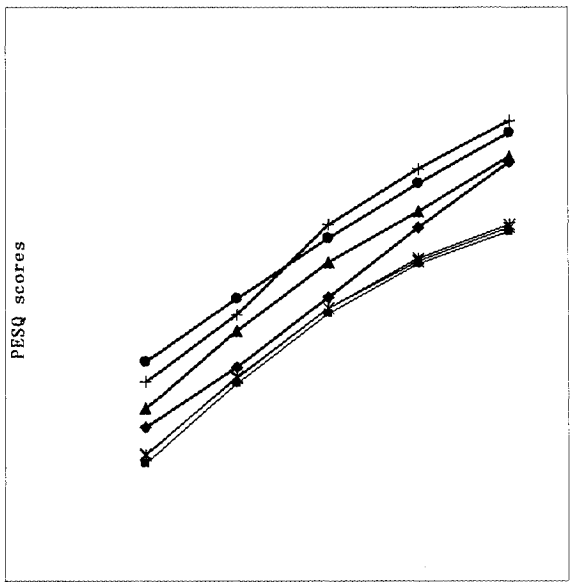
Fig. D.1 Average PESQ scores of different methods for white noise case



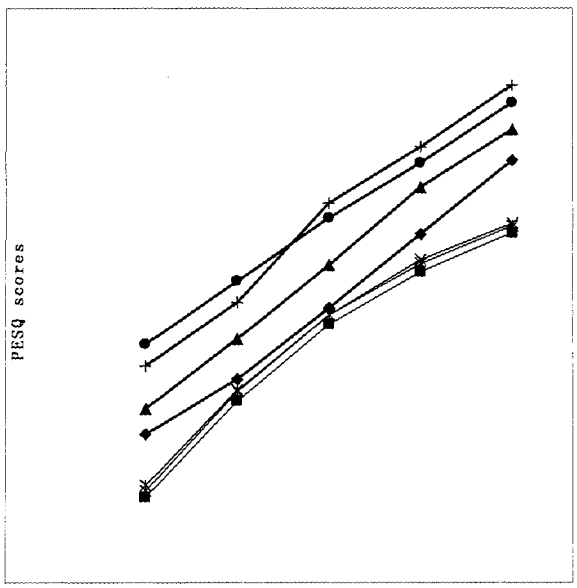
(a) female speech with babble noise



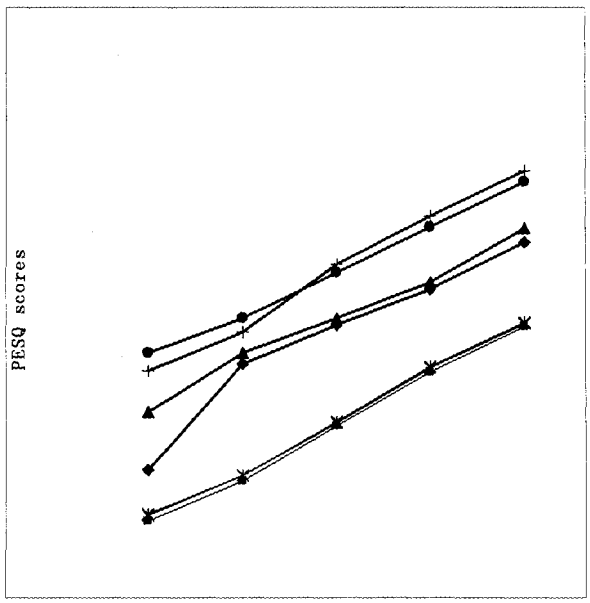
(b) female speech with car noise



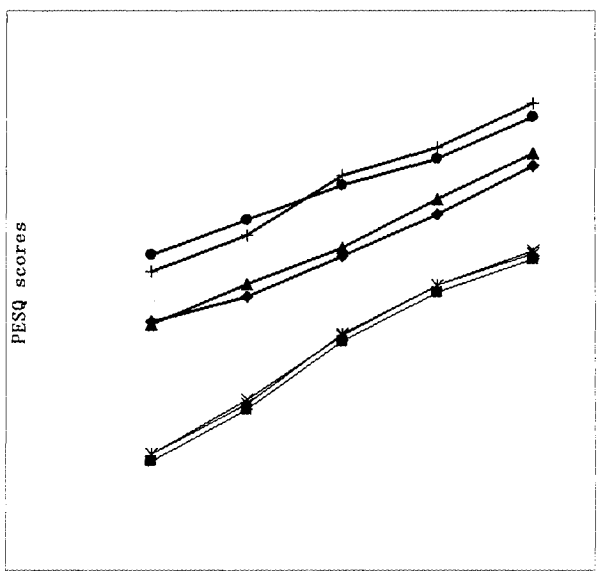
(c) female speech with street noise



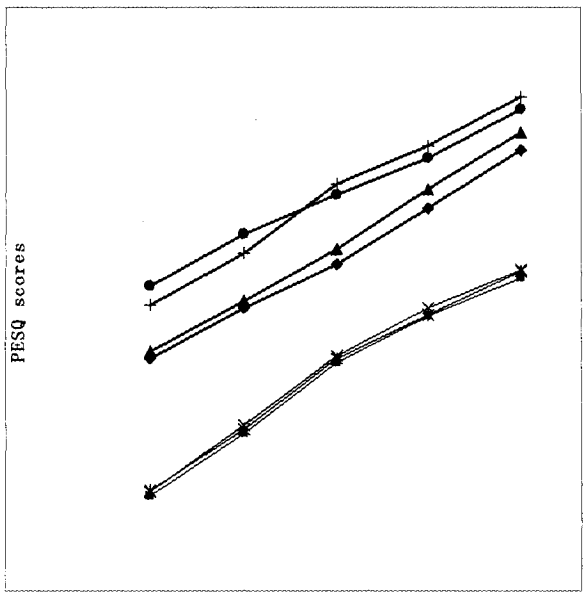
(d) female speech with generated colored1 noise



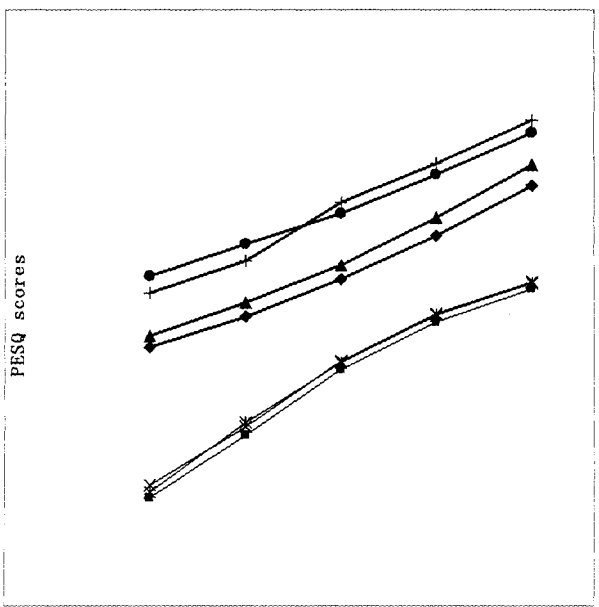
(e) male speech with babble noise



(f) male speech with car noise



(g) male speech with street noise



(h) male speech with generated colored1 noise

•: noisy speech signal, ¦ : traditional spectral subtraction method, ? : perceptually weighted spectral subtraction method [Vir99], ×: KLT method [Pop98], ✳: [Mit00], •: truly constrained Kalman filter with masking properties(CKFSFM), +: perceptually constrained dual UKF (PCDUKF).

Fig. D.2 Average PESQ scores of different methods for colored noise case

References

[Ami91] M. G. Amin, "A Frequency-Domain LMS Comb Filter," *IEEE Tran. Circuits Systems*, Vol. 38, No. 12, December, 1991, pp.1573-1576.

[Bah01] M. Bahoura and J. Rouat, "Wavelet Speech Enhancement Based on the Teager Energy Operator," *IEEE Signal Processing Letters*, Vol.8, No.1, January, 2001, pp.10-12.

[Ban99] M. Banbrook, S. McLaughlin, and I. Mann, "Speech Characterization and Synthesis by Nonlinear Methods," *IEEE Trans. Speech and audio Processing*, Vol. 7, No. 1, January, 1999, pp. 1-17.

[Bee94] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation," *Journal of the Audio Engineering Society*, Vol. 42, No. 3, 1994, pp. 115-123.

[Ber79] 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*, ICASSP'79, Vol. 4, April, 1979, pp. 208-211.

[Bol79] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No. 2, April, 1979, pp.113-120.

[Bro92] R. G. Brown and P. Y. C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*, 2nd Edition. John Wiley and Sons, Inc., New York. 1992.

[Che04] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies", *Proc. of IEEE ICASSP'2004*, Vol. 1, Montreal, Canada, May 2004, pp. I – 709-12.

[Coh04] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator", *IEEE Signal Processing Letters*, Vol. 11, Iss. 9 , Sept. 2004, pp725 - 728

[Chu99] C.K.Chui, G. Chen, *Kalman Filtering with Real-Time Applications*, 3rd Edition, Springer-Verlag, 1999

[Eph84] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 32, No. 6, December, 1984, pp. 1109-1121,.

[Eph89] Y. Ephraim, D. Malah and B. Juang, "On the Application of Hidden Markov Models

for enhancing Noisy Speech," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 37, No. 12, December, 1989, pp.1846-1856.

[Eph92] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proceedings of IEEE*, Vol. 80, No. 10, December, 1992, pp.1526-1555.

[Eph95] Y. Ephraim, H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 4, July, 1995, pp.251-266.

[Fun89] K. Funahashi,"On the Approximate Realization of Continuous Mappings by Neural Networks," *Neural Networks*, Vol. 2, 1989, pp. 183-192

[Gab01] Marcel Gabrea, "Adaptive Kalman Filtering-Based Speech enhancement Algorithm," *Canadian Conference on Electrical and Computer Engineering 2001*, Fredericton, New-Brunswick, Vol.1, May, 2001, pp.521-526.

[Gan98] Sharon Gannot, David Burshtein and Ehud Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 4, July, 1998, pp. 373-385.

[Gib91] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding", *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp.1732-1742, Aug. 1991

[Gor97] M. J. Goris, D. A. Gray and I. M. Y. Mareels, "Reducing the computational load of a Kalman filter", *Electronics Letters*, Volume 33, Issue 18, Aug. 1997, pp. 1539 – 1541.

[Gra93] J. T. Graf and N. Hubing, "Dynamic Time Warping Comb Filter for the Enhancement fo Speech Degraded by White Gaussian Noise," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'93, Vol. 2, 1993, pp.339-342.

[Har97] W. M. Hartmann, *Signals, Sound, and Sensation*. Springer Verlag, 1997.

[Hay96] S. Haykin, *Adaptive Filter Theory*, Third Edition, Prentice Hall, 1996

[Hel72] R. P. Hellman, "Asymmetry of masking between noise and tone," *Percept. and Psychophys.*, vol. 11, 1972, pp. 241-246.

[Hor91] K. Hornik, "Approximation Capabilities of Multilayer Feedforward Network", *Neural Networks*, Vol. 4, 1991, pp. 251-257

[Hua02] Y. Huang, T. Chiueh, "A New Audio Coding Scheme Using a Forward Masking

Model and Perceptually Weighted Vector Quantization", *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, July, 2002, pp.325-335.

[Huh98] H. T. Hu, "Comb Filtering of Noisy Speech Using Overlap-and-Add Approach," *Electronics Letters*, Vol.34, No.1, January, 1998, pp.16-18.

[ITUT01] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P.862, February 2001

[Jab02] F. Jabloun and B. Champagne, "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'2002, Vol. 1, 2002, pp.569-572

[Jes82] W. Jesteadt, S. P. Bacon, J. R. Lehman, "Forward Masking as a Function of Frequency, Masker Level, and Signal Delay", *J. Acoust. Soc. Amer.*, vol. 71, no. 4, April, 1982, pp. 950-962.

[Joh88] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Selected Areas in Communications*, Vol. 6, No. 2, February, 1988, pp.314-323.

[Joh03] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, and X. Yuan, "Performance of Nonlinear Speech enhancement Using Phase Space Reconstruction", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'03,Vol. 1, Apr. 2003, pp.920-923.

[Jul97] S. J. Julier and J. K. Uhlmann. "A New Extension of the Kalman Filter to Nonlinear Systems." *Proc. of AeroSense: The 11th Int. Symp. On Aerospace/Defence Sensing, Simulation and Controls*, Orlando FL, USA, Multi Sensor Fusion Tracking and Resource Management II, Vol. 3068, Apr. 1997, pp.182-193.

[Kid81] G. Kidd, Jr. and L. L. Feth, "Effects of masker duration in pure-tone forward masking", *J. Acoust. Soc. Amer.*, vol. 72, November, 1982, pp. 1384-1386.

[Lea81] J. F. Leathrum, "On Sequential Estimation of State Noise Variances," *IEEE Tran. Automatic Control*, Vol. AC-26, June, 1981, pp. 745-746.

[Lim79] J. S. Lim, A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proceedings of the IEEE*, Vol. 67, No. 12, December, 1979, pp. 1586-1604.

[Ma04a] N. Ma, M. Bouchard and R. A. Goubran, "Perceptual Kalman Filtering for Speech

Enhancement in Colored Noise", *Proc. of IEEE ICASSP'2004*, Vol. 1, Montreal, Canada, May 2004, pp. I - 717-20.

[Ma04b] N. Ma, M. Bouchard and R. Goubran, "Dual Perceptually Constrained Unscented Kalman Filter for Enhancing Speech Degraded by Colored Noise", *Proceedings of IEEE Seventh International Conference on Signal Processing (ICSP'04)*, vol. 3, pp. 2522-2525, Beijing, China, Aug. 31-Sept. 4 2004

[Ma04c] 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

[Ma04d] 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, n. 3, pp. 138-139, Ottawa, Ontario, Oct. 2004

[MaN03] N. Ma, M. Bouchard and R. Goubran,. "A perceptual Kalman filtering-based approach for speech enhancement," *Proceedings of Seventh International Symposium on Signal Processing and its Applications (ISSPA) 2003*, Vol. 1, July, 2003, pp. 373-376.

[MaN05] N. Ma, M. Bouchard and R. Goubran, "A Wavelet Kalman Filter with Perceptual Masking for Speech Enhancement in Colored Noise", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2005*, Philadelphia, USA, Mar. 2005, pp. I -149-152.

[MaP02] P. Maragos, A. G. Dimakis and I. Kokkinos, "Some Advances in Nonlinear Speech Modeling Using Modulations, Fractals, and Chaos," *Proc. 2002 14th International Conference on Digital Signal Processing,* DSP 2002, Vol. 1, 1-3, July, 2002, pp. 325-332.

[Mit00] U. Mittal and N. Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise," *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 2, March, 2000, pp. 159-167.

[Moo97] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 4th ed., 1997.

[Mye76] K. A. Myers and B. D. Tapley, "Adaptive sequential estimation with unknown noise

statistics," *IEEE Tran. Automatic Control*, Vol. AC-21, August, 1976, pp. 520-523.

[Nel97] A. T. Nelson, E. A. Wan, "Neural Speech Enhancement Using Dual Extended Kalman Filtering", *IEEE International Conference on Neural Netowrks*, vol. 4, 1997, pp. 2171-2175.

[Nem99] E. J. Nemer, *Speech Analysis and Quality Enhancement Using Higher Order Cumulants*, Ph. D. Thesis,Carleton University, Canada, November, 1999

[Nie92] R. J. Niederjohn and J. A. Heinen, "Understanding Speech Corrupted by Noise," *Proceedings of the IEEE International Conference on Industrial Technology*, 1996, pp. 1-5.

[O'Sh89] D. O'Shaughnessy, "Enhancing Speech Degraded by Additive Noise of Interfering Speakers," *IEEE Communications Magazine*, Vol. 27, No. 2, February, 1989, pp. 46-52.

[Opp89] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 1989.

[Pal87] 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*, ICASSP'87, Vol. 12, April, 1987, pp. 297-300.

[Pop98] D. C. Popescu, I. Zeljkovic, "Kalman Filtering of Colored Noise for Speech Enhancement", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'98, vol.2, Seattle, USA, May 1998, pp. 997-1000.

[Rix01] A. W. Rix, J. G. Beerends, M. P. Hollier, etc., "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2001, vol. 2, May, 2001, pp.749-752.

[Sch70] B. Scharf, ch. 5 in Foundations of Modern Auditory Theory. New York: Academic, 1970.

[Sch79] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, Dec. 1979, pp. 1647-1651.

[Sam98] H. Sameti, H. Sheikhzadeh, L. Deng et. al., "HMM-Based Stragegies for

Enhancement of Speech Signals Embedded in Nonstationary Noise," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 5, September, 1998, pp. 445-454.

[Tak81] F. Takens, "Dynamical systems and turbulence," *Lecture Notes in Mathematics*. Berlin, Germany: Springer-Verlag, 1981, vol. 898, pp. 366-381.

[Ter82] E. Terhardt, G. Stoll, and M. Seewann, Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, Vol. 71, March, 1982. pp.679-688.

[Thi00] T. Thiede, W. C. Treurniet, R. Bitto, et. al., "PEAQ-the ITU standard for objective measurement of perceived audio quality," J. Audio Eng. Soc., Vol. 48, No. 1/2, January, 2000, pp.3-29.

[Vir99] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 2, March, 1999, pp.126-137.

[Wan98] E. A. Wan, A. T. Nelson, "Removal of Noise from Speech Using the Dual EKF Algorithm", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* , ICASSP'98, vol. 1, 1998, pp. 381-384.

[Wan00] E. A. Wan and R. van der Merwe, "The Unscented Kalman Filter for Nonlinear Estimation", *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications and Control Symposium (AS-SPCC)*, 2000, pp. 153-158.

[Wit94] L. I. Witzke, R. J. Niederjohn, J. A. Heinen, and M. G. Sommerville, "Speech Synthesis Based on Feature Extraction to Enhance Noise-Corrupted Speech," *20th International Conference on Industrial Electronics, Control and Instrumentation*, Vol. 3, 1994, pp. 1946-1951.

[Zwi80] E. Zwicker, and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth asa function of frequency," *J. Acoust. Soc. Am.*, Vol. 68, November, 1980, pp. 1523-1525.

[Zwi84] E. Zwicker, "Dependence of post-masking on masker duration and its relation to temporal effects in loudness", *J. Acoust. Soc. Amer.*, vol. 75, January, 1984, pp. 219-223.

[Zwi99] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models*, Springer Verlag, 2nd ed., 1999.