Transform Domain Model-Based Wideband Speech Enhancement with Hearing Aid Applications

by

Brady N. M. Laska, B.Sc., M.A.Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy in Electrical Engineering

Ottawa-Carleton Institute for
Electrical and Computer Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada

Your file Votre référence
ISBN: 978-0-494-67867-1
Our file Notre référence
ISBN: 978-0-494-67867-1

Abstract

This thesis studies transform-domain model-based algorithms to improve the feedback and noise control performance of hearing aids processing wideband speech in non-stationary environments. Subband adaptive filter structures are investigated for continuous and non-continuous adaptation feedback compensation, and the particle filter framework is used to develop state-space model-based speech enhancement algorithms.

Subband feedback compensation is shown to offer several advantages. The flexibility offered by the frequency division allows subband systems to offer faster and more stable convergence for wideband signals, and better tracking of changing acoustic feedback paths. Robustness is also improved, as divergence caused by path changes or input signal correlation is confined to individual frequency bands.

A Rao-Blackwellized particle filter (RBPF) algorithm is proposed for enhancement of speech discrete cosine transform (DCT) coefficients, and is evaluated in comparison to the standard fullband RBPF. The DCT subband decomposition is shown to enable improved modeling of wideband speech signals, especially in spectral troughs, thereby decreasing intra-speech noise in both best-case and real-world conditions.

A novel particle filter algorithm is also proposed for short-time spectral amplitude (STSA) speech enhancement. A dynamic model of spectral amplitude evolution is used to allow for speech signal correlation in the frequency domain. Two variants

of the basic algorithm are presented: the first incorporates phase information to improve the spectral amplitude estimates; the second uses interacting multiple models to account for speech presence uncertainty. The monaural algorithm is extended to the binaural case with a filter based speech and noise parameter estimator. The estimator exploits knowledge of the diffuse noise field coherence to separate the clean speech and noise power spectra without external noise or voice-activity estimation.

In both feedback and noise control, the transform domain allows for processing strategies that take into account the diverse frequency-dependent characteristics of the wideband signals. Incorporating models of speech, noise and the acoustic environment allows the addition of time-domain constraints and *a-priori* knowledge to address signal non-stationarity. The developed algorithms are evaluated using real speech and noise signals recorded with a commercial hearing aid.

Acknowledgments

First and foremost I would like to thank my supervisors, Dr. Rafik Goubran and Dr. Miodrag Bolić; they left me free to explore when I wanted to and provided me with direction when I needed it, the whole time offering friendly guidance, advice, and support.

I would also like to acknowledge the financial support of the Ontario Graduate Scholarship (OGS) program, the Natural Sciences and Engineering Research Council (NSERC) and Siemens AG. The Siemens Audio team also provided me with valuable feedback and helpful suggestions.

The adminstrative and techical staff of the Department of Systems and Computer Engineering deserve thanks for ensuring that my applications were complete and accurate, and that our network and lab ran smoothly. I would also like to thank my fellow graduate students in the DSP lab for the advice and helpful distraction they provided.

Finally, I reserve special thanks to Susan for her support and patience, and to my son Erik for the joy and motivation he has brought me.

Table of Contents

List of Figures

List of Acronyms

SNR	Signal-to-noise ratio
PSD	Power spectral density
MSC	Magnitude-squared coherence
NLMS	Normalized least-mean square
IPNLMS	Improved proportionate NLMS
FFT	Fast Fourier transform
DFT	Discrete Fourier transform
DCT	Discrete cosine transform
STFT	Short-time Fourier transform
STSA	Short-time spectral amplitude
OLA	Overlap-add
VAD	Voice-activity detector
MSE	Mean-squared error
MMSE	Minimum mean-squared error
MAP	Maximum *a-posteriori*
ML	Maximum likelihood
PDF	Probability density function
AR	Auto-regressive
TVAR	Time-varying AR

ITD	Interaural time difference
ILD	Interaural level difference
SIR	Sampling importance resampling
RBPF	Rao-Blackwellized particle filter
CSII	Coherence speech intelligibility index
WPESQ	Wideband perceptual evaluation of speech quality
LLR	Log-likelihood ratio
LPC	Linear predictive coding
AWGN	Additive white Gaussian Noise

Chapter 1

Introduction

1.1 Problem Background

Individuals with hearing loss possess a decreased ability to detect or perceive sounds. The severity of hearing loss ranges considerably: individuals with mild to moderate hearing loss may be able to hear speech at close range in a quiet room but have difficulty following a group conversation, while an individual with more severe hearing loss may not be able to perceive any speech at conversational volume [1]. Hearing loss is classified as conductive or sensorineural. Conductive hearing loss is the result of an impairment in the outer or middle ear, and can frequently be resolved through surgical means or simple amplification. Sensorineural hearing loss is a broader category, encompassing hearing loss arising from deficiencies in the sensory or neural components of the auditory system. Sensorineural hearing loss is most commonly caused by damage to or abnormalities in the hair cells within the cochlea resulting from genetic conditions or damage incurred through exposure to intense sound, disease, biochemical agents, or aging [2]. Sensorineural hearing loss is frequency and amplitude dependent, so it affects not only the absolute level, but also the quality and character of sound. As a result some audible sounds may not be properly understood,

degrading speech intelligibility. As the baby-boom generation ages, the proportion of the population with sensorineural hearing impairments is growing. Between the years 2000 and 2004, the hearing impaired population in the United States grew by 9.9% to an estimated 31 million people, compared to an 6.8% population increase; by 2025 the hearing impaired population in the United States is expected to reach 40 million people [3].

Hearing aids work to improve life for people with hearing loss. In addition to amplifying sounds, enabling users to hear sounds that lie beneath their impaired threshold of hearing, they also reduce ambient noise levels, reducing the cognitive effort required to hear, improving listening comfort and the ability to hear in noisy situations. Hearing aid users report that the devices improve their quality of life by enhancing their ability to communicate effectively and to participate in groups, as well as increasing their sense of personal safety and independence [4]. Despite the benefits and the large potential user-base, the percentage of the hearing impaired population actually adopting hearing aids continues to lie between 20 and 25%. More distressing is the fact that over 15% of hearing aid owners do not use their devices and only 70% of active users are satisfied with their hearing aid devices [3]. These factors indicate that there are situations where the performance of current hearing aids is insufficient, and new processing strategies are needed to increase satisfaction amongst existing users.

The most common complaint among hearing aid users is the poor performance of the device in noisy situations, especially large groups and in school classrooms [4]. While a hearing aid can successfully compensate for hearing loss in quiet, in noisy situations the noise is amplified along with the desired signal, negating the benefit. Another common complaint relates to acoustic feedback in hearing aids [4]. Low levels of acoustic feedback manifest as reverberation which affects the spectral

energy distribution of the signal, distorting the desired speech. High levels of feedback manifest as annoying "whistling" or "howling" sounds that are very disturbing to the user.

The fact that large groups, school classrooms and reverberation due to acoustic feedback are the most problematic listening situations is not surprising. A competing sound with the same frequency content as a desired signal reduces the sensitivity of the ear to another sound through an effect known as masking. As a result, the signal energy level required to hear a sound in noise is greater than the level in quiet, and the intelligibility of the desired signal is largely governed by the signal-to-noise ratio (SNR) [5]. Most commonly encountered noises such as car road noise and noise from ventilation systems are narrowband so there exists large spectral regions where the speech and noise do not overlap and from which the speech information can be extracted. Also, the power spectrum of these signals is stationary, facilitating statistical recovery of the clean speech. In contrast multi-talker babble and reverberation are composed of speech, so they are broadband and occupy the same spectral regions as speech. The characteristics of the speech signals processed by hearing aids are also non-stationary and frequency dependent, degrading the performance of algorithms that assume uniformity in time or frequency domains.

1.1.1 Wideband speech enhancement

Most speech processing algorithms are designed for telephony applications where the bandwidth is limited to the 300 - 3300 Hz frequency range, while hearing aids operate at higher sampling frequencies and include wideband speech content. Fig. 1.1, presents a spectrogram of a segment of wideband speech along with the short-time power spectral densities (PSDs) from segments containing harmonic content and a speech onset; the dashed line at 4 kHz represents the high frequency cut-off for

traditional narrowband speech. Wideband speech has a greater spectral slope than narrowband, and while narrowband speech is dominated by harmonic formants, the high-frequency content in wideband speech is spectrally flatter and temporally more transient. These properties create problems for fullband processing algorithms. In order to capture the characteristics of wideband speech, time-domain speech enhancement algorithms require very high-order autoregressive models that can vary with time, but accurately fitting high-order models requires long signal segments, reducing the time-resolution. Similarly, feedback compensation is typically performed using an adaptive filter to model the feedback path [6], but it is well known that the uneven distribution of wideband speech energy leads to slow convergence and poor tracking performance of adaptive filters [7]. Acoustic feedback paths can change frequently in response to a telephone, hand, or other reflecting object being placed near the ear; therefore, feedback cancellation algorithms need to respond quickly and offer fast tracking of these changing paths. Enhancement algorithms must be designed to accommodate the frequency-dependent nature of the wideband speech signal.

Processing of wideband speech is also complicated by its non-stationary nature in the time domain. The speech signal exhibits fluctuations in energy levels and frequency content that depend on the speaker and what is being said. These fluctuations reduce the performance of models that rely on long-term statistics and signal stationarity. As a result traditional means of noise reduction have to choose between capturing the time-variations of the signal or obtaining a low-variance estimate. This presents a trade-off between higher residual noise and artifact levels, or smoothing of speech leading to temporal blurring, reduced intelligibility and removal of low-energy speech content [8], leaving the problems of perceptual noise level and intelligibility improvement at odds.

The degradations in algorithm performance caused by the time and frequency

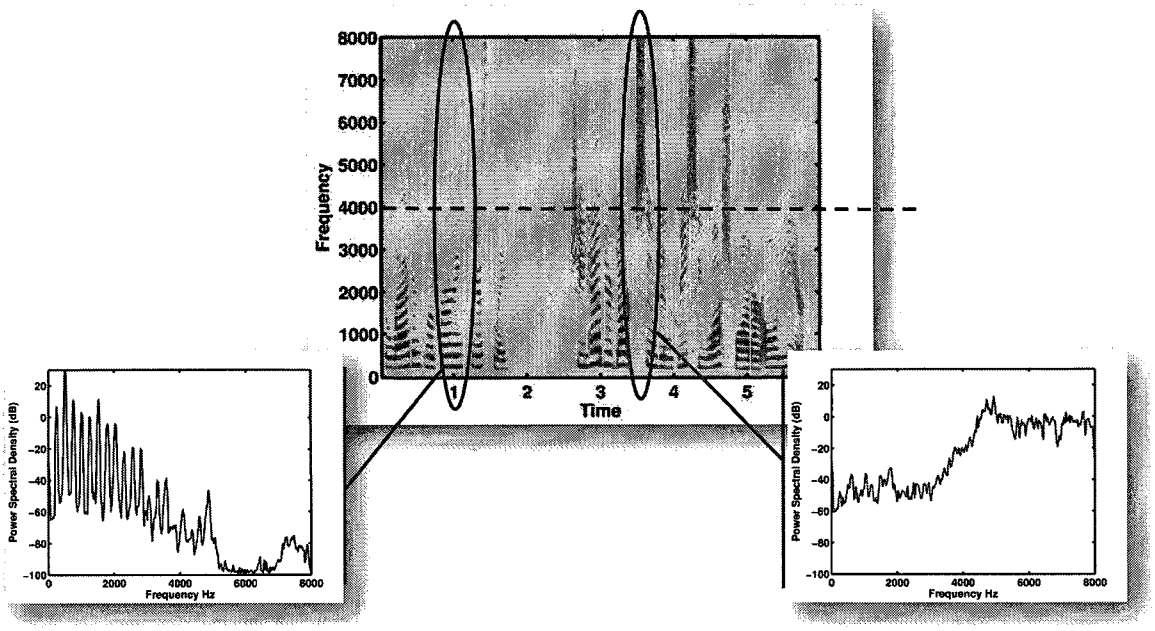


Figure 1.1: Wideband speech spectrogram with short-time PSDs of voiced (left) and unvoiced (right) segments. The dashed line is the narrowband high-frequency cutoff.

fluctuations of wideband speech as a desired signal are compounded when the interfering signal is also composed of wideband speech. The impact of noise masking on speech intelligibility is greater for individuals with hearing loss than for those with ordinary hearing. The effective level of hearing loss for speech in noise is only weakly related to hearing loss for speech in quiet [9]. With a single target in quiet, simple amplification can bring the target speech to a level where it can be heard. However in multi-talker situations a hearing aid amplifying both speech and noise will not improve the SNR and therefore intelligibility. Furthermore, while the auditory systems of ordinary-hearing listeners employ a number of strategies to focus on and understand a single speaker in a multi-talker environment, the performance of these strategies is compromised by sensorineural hearing loss.

The time-varying energy modulation of a competing talker allows the listener to

catch high SNR glimpses of the desired speech when the instantaneous energy of the competing talker is low. Even though average SNR may be low, the instantaneous SNR can be high during competing speech pauses. This results in a "release from masking" whereby the absolute signal energy level required for intelligibility is lower in the presence of a competing talker compared to steady-state noise with the same average energy. The release from masking degrades as the number of talkers increases because both the frequency and duration of the gaps used for glimpsing decreases. Reverberation, whether it is caused by reflections in the acoustic environment or acoustic feedback, also fills in some of the glimpsing gaps, reducing the release from masking provided by masker modulation and decreasing intelligibility. Hearing impaired listeners benefit less from masker modulations as the high SNR intervals used for glimpsing may occur when the target speech is below the impaired listeners' hearing threshold. The decreased time and frequency resolution characteristic of sensorineural hearing loss also has an impact, as temporal smearing eliminates the small gaps used for glimpsing, and decreased frequency resolution can lead to a decrease in head shadowing benefit and modulation release from masking [10]. The result is that hearing impaired listeners have less information to work with in noisy situations, adding to the absolute level effects of hearing loss. Consequently the increase in speech reception threshold for a hearing impaired listener compared to a normal hearing listener is greater in a fluctuating noise condition than in steady state noise or in quiet.

Since the effects of hearing loss are increased in the presence of fluctuating noise, the *potential* intelligibility benefit of a hearing instrument is higher in fluctuating noise than stationary. However, as the satisfaction rates indicate, the benefit realized by current devices is low. Several reasons for this performance gap can be identified. Traditional frequency domain speech enhancement algorithms work by subtracting an

estimate of the of the noise spectrum from the noisy speech [11], or by estimating the clean speech spectral amplitude using a statistical model of the speech and noise [12]. Time-domain algorithms use an autoregressive model of speech and assume white noise [13] or autoregressively modeled noise [14]. These methods require robust estimates of the noise statistics and model parameters, requiring long-term averaging which creates a lag in estimation for fluctuating noise environments. Common interference sources such as reverberation and babble noise are variable, with an overall background baseline energy level that is dependent on the number of speakers, interspersed with occasional higher energy bursts from nearby talkers. Fluctuations of the noise from its expected value can lead to residual noise when the noise is under-estimated and, more critically, speech distortion when the noise is over-estimated. Furthermore, successive over- and under-estimation of the noise spectrum can lead to annoying and distracting time-varying tonal artifacts known as musical noise [8]. Similarly, an interfering noise signal with a fluctuating power spectrum disturbs the adaptive filters used for feedback suppression, requiring robust algorithms to control their adaptation. Also, interfering speech correlated with the feedback signal can cause the filter to diverge completely, leading to cancellation of the desired speech signal or system instability. In order to realize the potential intelligibility benefits to hearing impaired listeners, new speech enhancement strategies are needed when both the desired and the interfering signals are composed of wideband speech. These strategies must take into consideration the non-uniform energy distribution in the time and frequency domains that is characteristic of the wideband speech signal.

1.1.2 Binaural processing

Spatial sound processing is a strategy used by normal-hearing and hearing impaired listeners that has the potential to be enhanced and exploited by systems operating on

non-stationary speech and noise signals [9]. Having two ears placed on opposite sides of the head provides the listener with two versions of the desired speech signal and any competing sounds, thereby providing spatial information that can be exploited to distinguish the signal from the noise. The delay between the signal received in each ear corresponds to the location of the source around the head, and frequency-dependent head shadowing provides additional source location information, enabling the listener to perform spatial filtering to focus on a source from a given target direction. With binaural hearing the release from masking by spatial separation of target and interferer can be up to 10 dB, and in multi-talker situations the spatial configuration of the sources has more of an impact on intelligibility than number of talkers [9]. While some of the frequency dependent spatial cues are lost to hearing impaired listeners, the time and level difference information is still present and is actively exploited to improve listening in noise.

In current two-ear hearing aid systems the hearing aids work independently as monaural devices in a bilateral configuration. Future hearing aids will include an interaural wireless link, enabling the shift from bilateral to true binaural hearing aid systems [15]. This shift will require the development of speech enhancement systems that take binaural factors into account. While some multi-channel speech enhancement algorithms exist, most are not designed specifically to take advantage of, or even preserve, binaural information. Applying existing monaural or multi-channel algorithms in a binaural situation has the potential to impair binaural hearing-in-noise strategies. Distortion of time and level difference cues by bilateral hearing aids has been shown to reduce the localization ability of hearing impaired users [16], potentially further reducing intelligibility. True binaural systems open up the possibility of greater spatial awareness in a noise reduction system. This could lead to higher noise reduction, better preservation of speech content and the ability to distinguish

between a target and a nearby talker. Binaural algorithms could also be developed to mimic or enhance the binaural processing performed by the human auditory system.

1.2 Thesis Statement and Objectives

While current hearing aids offer intelligibility improvements in some environments, in complex noisy situations, where the potential benefit to users is the greatest, the potential is largely unrealized. Feedback control and noise reduction performance have been identified as two performance bottlenecks. Hearing aids have recently made the shift from analog to digital devices, which has enabled the introduction of digital signal processing algorithms for feedback and noise control; however, the traditional speech processing algorithms employed are largely designed for telephony applications and the assumptions invoked in their development do not reflect the reality of current and future hearing aids. This thesis aims to improve the state of the art of speech enhancement for hearing aids by developing algorithms that take into account the challenges faced in real environments. Specifically, the algorithms will address the following issues identified in the previous section:

Wideband speech and noise: Multi-band compression in hearing aids has increased the frequency content of the signal that can be comfortably presented to hearing aid users. While the increased signal content has the potential to improve intelligibility, the varied spectral content poses problems for speech enhancement and feedback reduction algorithms designed for telephone bandwidth.

Non-stationarity: As with signal frequency variability, time variability also degrades the performance of existing speech enhancement algorithms which are

commonly designed for stationary noise and long term speech statistics. Fluctuating noises and non-stationary acoustic enclosures also degrade the convergence speed and overall performance of feedback compensation algorithms. Poor performance with non-stationary signals is especially problematic in hearing aid applications, as the effects of hearing impairment on intelligibility are increased when the noise source is fluctuating.

Time/frequency resolution: Model fitting in speech enhancement algorithms results in a trade-off between time and frequency resolution. The large models required for good frequency performance require large amounts of data to accurately estimate, degrading performance when the signal statistics or acoustic environment are non-stationary.

Residual noise character: Among single channel noise reduction algorithms transform-domain statistical estimation algorithms offer good noise reduction but un-natural sounding noise; while model-based approaches offer natural sounding noise, but noise reduction can be limited. To achieve intelligibility improvements, hearing aid noise reduction algorithms must produce a speech output with low-level but natural sounding residual noise.

To address these issues and improve hearing aid performance in the complex environments experienced in practice, this thesis will combine transform domain and model-based speech enhancement techniques to exploit the advantages of both types of processing. This approach is depicted in Fig. 1.2. Applying a frequency transform will allow the developed algorithms to account for the frequency dependent speech and noise characteristics by tailoring the processing strategy based on the characteristics of the signal content in each transform bin. Incorporating model-based approaches can account for the fluctuations and changing statistics of wideband speech and noise

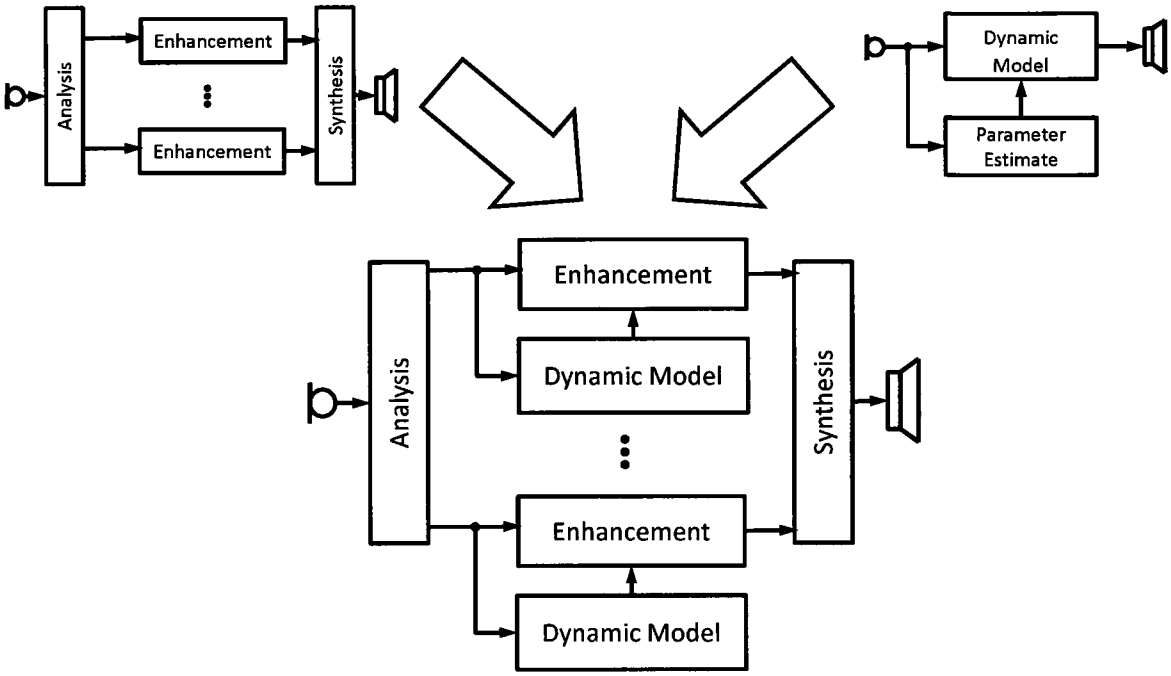


Figure 1.2: Application of model-based speech enhancement algorithms in transform domains.

signals, enabling the developed algorithms to separate the desired signal from the noise, even when the signals overlap in frequency. Applying models in transform domains rather than the time domain may also improve model parameter estimation, as the transform domain models can be smaller, requiring fewer parameters to be estimated. The parameters of the model in each transform bin can be adjusted independently, offering additional degrees of freedom to handle the time/frequency resolution trade-off. In some cases the transform domain and dynamic model combinations have not been explored because they lead to statistical estimators that are intractable in closed-form; in this thesis new algorithm paradigms, such as the particle filter framework, will be explored to enable the use of new information and more sophisticated models of speech and noise to improve performance.

1.3 Thesis Contributions

The primary contributions of this thesis are outlined as follows:

Subband acoustic feedback compensation: Acoustic feedback compensation in hearing aids is typically performed on the fullband signal, before the subband decomposition used for noise and gain compensation. The use of a subband structure is investigated for modeling wideband acoustic feedback paths in continuous and non-continuous adaptation feedback compensation systems. It is demonstrated that, compared to traditional fullband structures, the flexibility offered by the frequency division allows subband systems to offer faster and more stable convergence and better feedback path tracking when presented with wideband speech signals and changing acoustic environments. Robustness is also improved, as divergence caused by feedback path changes or input signal correlation is confined to individual frequency bands. This work is presented in Chapter 4 and is published in part in:

[17] B. Laska, R. Goubran, and M. Bolić, "Improved proportionate subband NLMS for acoustic echo cancellation in changing environments," *IEEE Sig. Process. Letters*, vol. 15, pp. 337 – 340, 2008.

[18] B. Laska, M. Bolić, and R. Goubran, "Subband adaptive filtering for acoustic feedback compensation in hearing aids," *Canadian Acoustics*, vol. 37, pp. 134 – 135, Sept. 2009.

DCT RBPF speech enhancement: A subband version of the Rao-Blackwellized particle filter speech enhancement algorithm is proposed and evaluated. The use of a model-based statistically motivated algorithm that takes the noise distribution into account addresses the issue of changing speech and noise statistics;

while the subband decomposition enables the algorithm to better enhance wideband speech. The limitations of fullband and subband RBPF algorithms are investigated and it is shown that the subband decomposition enables improved modeling of a wideband speech signal, and that the best-case and real-world performance of the proposed subband algorithm are significantly better than the equivalent fullband system. A hybrid particle/Wiener filter is also proposed that takes advantage of the fact that wideband speech characteristics are non-uniform across frequency. They hybrid system offers significantly reduced complexity with a modest reduction in performance. This work is presented in Chapter 5 and has been published in part in:

[19] B. Laska, R. Goubran, and M. Bolić, "Subband autoregressive modelling for speech enhancement," *Canadian Acoustics*, vol. 37, pp. 62 – 63, Sept. 2009. (CAA 2009 student presentation award winner).

[20] B. Laska, M. Bolić, and R. A. Goubran, "Discrete cosine transform particle filter speech enhancement." Manuscript accepted for publication in *Elsevier Speech Communication*, April 2010.

Spectral amplitude particle filter speech enhancement: A novel particle filter algorithm is proposed for short-time spectral amplitude (STSA) speech enhancement. In order to facilitate closed-form solutions, traditional STSA enhancement algorithms assume independence of successive speech frames and rely on static statistical modeling of the spectral amplitudes. The particle filter framework allows the use of dynamic models of spectral amplitude evolution which account for speech signal correlation, reducing speech distortion in fluctuating noise environments. A basic algorithm framework is developed, and two algorithm variants are presented: the first incorporates phase information

to improve the spectral amplitude estimates; the second uses an interacting multiple model approach to account for speech presence uncertainty. This work is presented in Chapter 6 and has been published in:

[21] B. Laska, M. Bolić, and R. A. Goubran, "Particle filter enhancement of speech spectral amplitudes." To appear in IEEE. Trans. Audio Speech Lang. Process., vol. 18, Aug. 2010.

Binaural speech enhancement: The proposed STSA particle filter algorithm is extended to the binaural case using a binaural Wiener filter based signal and noise parameter estimator. The estimator assumes a frontal or near-frontal target and diffuse noise, it exploits knowledge of the diffuse noise field interaural coherence to estimate the left and right channel clean speech and noise power spectra. The left and right channel spectral gains from the particle filters are combined to produce a system with a binaural output that preserves interaural time and level differences. The system does not require external noise estimation or voice activity detection and is able to effectively attenuate even fluctuating noise sources with a diffuse character. This work is presented in Chapter 7 and has been published in part in:

[22] B. Laska, M. Bolić, and R. A. Goubran, "Coherence-assisted Wiener filter binaural speech enhancement." To appear in Proc. IEEE I^2MTC., (Austin, TX, USA), May 2010.

1.4 Organization

This thesis is organized into eight Chapters. Chapter 2 presents a review of relevant speech processing background. The hearing aid noise reduction and feedback cancellation problems are outlined and solution approaches from literature are discussed. Chapter 3 describes the experimental setup used to conduct the simulations, and the performance measures that are used to evaluate the presented algorithms. The thesis contributions are described in Chapters 4 – 7. Chapter 4 proposes a subband adaptive filtering structure to address the issues of convergence speed and robustness of feedback cancellation systems. Chapter 5 presents a particle filter algorithm for enhancement of speech DCT coefficients that reduces complexity and improves performance over the fullband algorithm from which it is derived. Chapter 6 presents a dynamic model-based approach to spectral amplitude speech enhancement that uses the particle filter framework to account for the non-Gaussian statistics and time correlation observed in speech spectral amplitudes. Chapter 7 extends the spectral amplitude particle filter to the binaural case, exploiting the spatial coherence of sound sources to obtain instantaneous estimates of the required speech and noise parameters. Chapter 8 summarizes the conclusions of the previous Chapters and offers suggestions for future extensions to the work presented in this thesis.

Chapter 2

Background

2.1 Speech Enhancement for Hearing Aids

Fig. 2.1 presents a simplified block diagram of a modern digital hearing aid. The movement from analog to digital signal processing, combined with the increasing speed a decreasing power consumption of processors has enabled rapid improvements in hearing aid technology [23]. Rather than providing simple amplification, the device performs multi-band amplification including dynamic range compression to compensate for the "recruitment phenomenon" associated with sensorineaural hearing loss [15]. The recruitment phenomenon results in the perceived loudness curve of a hearing impaired listener not being a linear shift of a normal listener. While soft sounds may sound attenuated or even fall below the impaired listener's perception threshold, loud sounds are perceived at the same volume as a normal hearing listener. Simple amplification without considering the recruitment phenomenon can lead to loud sounds being pushed above the uncomfortable level. The use of digital hearing aids also enables sophisticated noise processing algorithms and more robust identification and compensation of feedback. Furthemore, signal and environment classification and control logic can be to determine appropriate settings and modes of

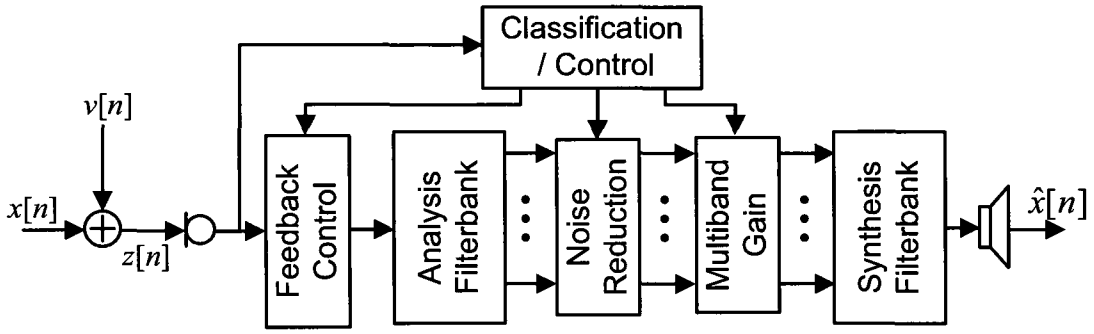


Figure 2.1: Simplified block diagram of hearing aid, after [23].

operation; for example a user listening to music would wish to have noise reduction disabled to prevent any possible distortion.

2.2 Feedback Control

Acoustic feedback occurs in hearing aids when the amplified sound signal played out to the user is coupled back to the input microphones. Feedback could be prevented by completely blocking the ear canal, but this would lead to the occlusion effect, which is when a users own voice sounds hollow and echoed due to amplification of low frequency speech components within the blocked ear canal. A vent is included in the ear-mold to reduce the occlusion effect and also to prevent moisture buildup. Larger vent sizes are desirable as they provide a more natural listening environment for the user, however they also increase the level of feedback.

The feedback problem is illustrated in Fig. 2.2, where $G(z)$ represents the signal processing performed by the hearing aid, $H(z)$ is the feedback path $s[n]$ is the input signal which is mixed with the feedback signal $d[n]$ and the measurement noise $v[n]$ to give the measurement signal $y[n]$ and $x[n]$ is the amplified signal sent to the user.

In the absence of measurement noise, feedback creates a closed loop system with the transfer function:

$$X(z) = Y(z)G(z) + X(z)H(z)G(z) \tag{2.1}$$

$$\frac{X(z)}{Y(z)} = \frac{G(z)}{1 - H(z)G(z)} \tag{2.2}$$

If $|H(z)G(z)| > 1$, the system has a positive feedback loop and can become unstable. In practice the gain limiting of the hearing aid prevents complete instability but leads to oscillations at the unstable frequencies. These oscillations manifest as annoying "whistling" or "howling" sounds and are very disturbing to the user. Low levels of feedback appear as reverberation and can still cause audible distortion and reduction of speech intelligibility, as the reverberation masks the incoming speech. Next to performance in noisy situations, feedback performance is the most common complaint among hearing aid users, with almost 1/3 of users reporting dissatisfaction [4].

The simplest method to prevent feedback is to limit the gain of the hearing aid so that the closed-loop gain is less than unity and positive feedback is prevented. This method is obviously undesirable as limiting the gain limits the utility of the device and changing feedback paths can render a previously stable gain unstable. One approach that is used in practice is to use adaptive notch filters to attenuate the oscillating frequencies when feedback occurs [15]. While effective this approach has several drawbacks: first while the notches can be narrow, the filtering invariably introduces some signal distortion; second the approach is reactive, the notch filters are only applied when oscillation is detected, so there is a still a period of feedback while the oscillatory frequencies are determined.

The most desirable method for feedback control is direct compensation of the feedback signal. If the feedback path is known, a replica of the feedback signal can

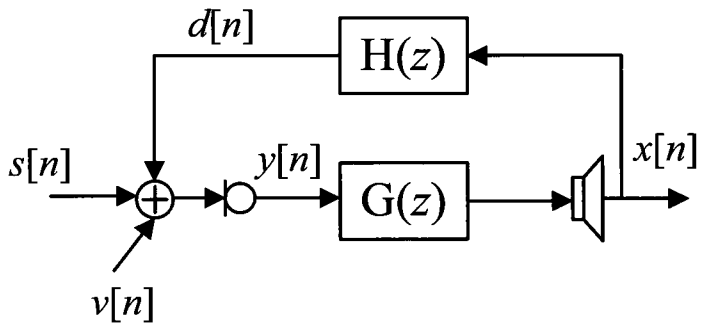


Figure 2.2: Acoustic feedback in a hearing aid system, after [15].

be created and subtracted from the microphone signal, enabling arbitrary gain in the hearing aid without signal distortion. The difficulty with this approach is obtaining an accurate model of the feedback path. Since the feedback path changes with movement of the hearing aid within the ear, and with the presence of reflecting surfaces around the ear, such as hands, telephones and hats; static modeling is insufficient so dynamic modeling of the feedback path with an adaptive filter is required.

2.2.1 Adaptive System Identification

A length-N adaptive filter $\mathbf{w}[n]$ attempts to model a linear, length L, finite impulse response plant $\mathbf{h}$ by filtering the input vector $\mathbf{x}[n]$ to create an estimate of the plant output, and adjusting the tap weights of $\mathbf{w}[n]$ to minimize the error $e[n]$ between the that estimate and the measured plant output. In the absence of measurement noise:

$$e[n] = d[n] - \hat{d}[n] \tag{2.3}$$

$$= (\mathbf{h} - \mathbf{w}[n])^T \mathbf{x}[n] \tag{2.4}$$

If $L = N$ it is possible for $\mathbf{w}[n] = \mathbf{h}$, resulting in perfect modeling of the plant. In

practice L is typically larger than N and measurement noise always corrupts the error signal. It can be shown [7] that the linear filter which minimizes the mean-square error (MSE), between the output and the desired signal, $\mathbb{E}\{|d[n]-\hat{d}[n]|^2\} = \mathbb{E}\{e^2[n]\}$, is given by:

$$\mathbf{w}_{opt} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xd} \tag{2.5}$$

where $\mathbf{R}_{xx} = \mathbb{E}\{\mathbf{x}[n]\mathbf{x}^H[n]\}$ is the correlation matrix of the input signal, and $\mathbf{r}_{xd} = \mathbb{E}\{\mathbf{x}[n]d^*[n]\}$ is the cross-correlation between the input and the desired signal. The solution in (2.5) is known as the Wiener solution. A number of algorithms have been derived to adapt the modeling filter coefficients online so that the tap weight vector converges to the Wiener solution, and there is generally a tradeoff between algorithm complexity and speed of convergence. In practice the plant $\mathbf{h}[n]$ is time varying; acoustic paths can change rapidly in response to movements of the reflecting surfaces surrounding the speaker and microphone, so the adaptive algorithm must also be able to provide fast tracking of plant changes.

The normalized least-mean square (NLMS) algorithm, belongs to the family of stochastic gradient algorithms that minimize the expected value of the instantaneous squared error. The NLMS tap weight update equation is given by

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu\frac{1}{\mathbf{x}^H[n]\mathbf{x}[n]+\delta}\mathbf{x}[n]e^*[n] \tag{2.6}$$

where $\mu \in [0,2]$ is a parameter which controls the size of the "step" taken by the weight vector at each iteration. The convergence time, stability, and steady state error are dependent on the step-size parameter. Larger values of μ lead to faster convergence but, due to the noisy nature of the gradient estimate, larger steady-state error. Also, since the tap weight vector affects the output error, which in turn

controls the adaptation, there is feedback in the adaptation process, so the algorithm can become unstable if μ is too large. A small μ effectively lowpass filters the fluctuations in the gradient estimate, keeping the algorithm stable and yielding less steady-state error, but slower convergence. In the absence of near-end disturbances, the NLMS-adapted coefficient vector $\mathbf{w}[n]$ converges to the Wiener solution $\mathbf{h}_{opt}$ in the mean-square sense. The NLMS algorithm is one of the most frequently used adaptive filtering algorithm for acoustic echo cancellation due to its robustness, ease of implementation, and low complexity.

The NLMS algorithm does not make any assumptions about the nature of the system being modeled, making it a suitable, but not necessarily the most efficient, algorithm for all types of system identification problems. By using individual step-sizes that are chosen adaptively, the shape of the echo path can be exploited without a-priori knowledge of the characteristics of the acoustic environment. An example of an individual step-size algorithm that adaptively modifies the step-sizes is the Improved Proportionate NLMS (IPNLMS) algorithm [24]. In IPNLMS the available adaptation gain is distributed in proportion to the tap energy. This is achieved by using a time-varying $N \times N$ step-size matrix with diagonal elements that are proportional to the absolute value of the corresponding adaptive filter tap weight. A fixed element is added to the step-size matrix to smooth the energy distribution and thereby improve the initial convergence when the coefficient estimate is not accurate. The tap weight update for IPNLMS is given by:

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu \frac{\mathbf{A}[n]\mathbf{x}[n]}{\mathbf{x}^H[n]\mathbf{A}[n]\mathbf{x}[n] + \delta} e^*[n] \tag{2.7}$$

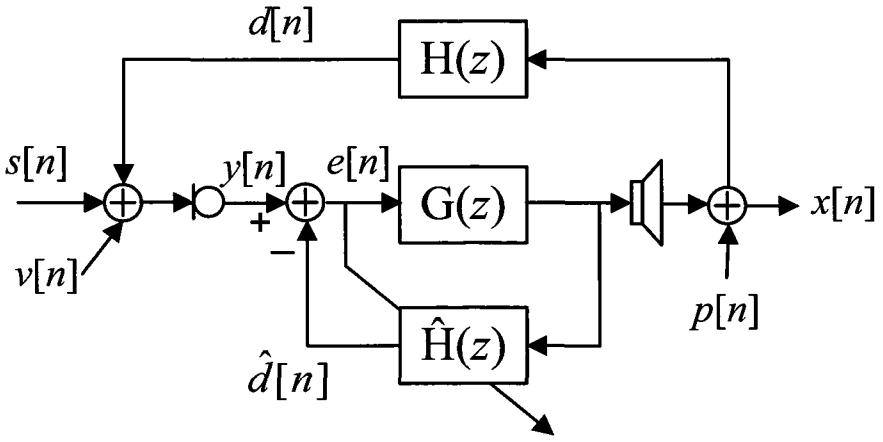


Figure 2.3: Feedback compensation using an adaptive filter.

where $\mathbf{A}[n]$ is a diagonal step-size matrix with elements:

$$a_{ii}[n] = \frac{1-\alpha}{2N} + (1+\alpha)\frac{|w_i[n]|}{2\|\mathbf{w}[n]\|_1 + \epsilon}, i \in \{0, 1, \ldots, N-1\} \tag{2.8}$$

and α is a parameter that controls the ratio of fixed to proportionate step-size. For $\alpha = -1$ the algorithm reduces to standard NLMS, and for $\alpha = 1$ the tap weight allocation is completely proportionate. According to [24], good choices of α are 0 or -0.5, and a value of $\alpha = -0.5$ is used for the simulations in this work.

2.2.2 Adaptive Feedback Compensation

A block diagram of an adaptive feedback cancellation system is presented in Fig. 2.3. Feedback modeling schemes can apply either continuous or non-continuous adaptation. In non-continuous adaptation schemes the feedback signal is monitored to detect oscillations or a changes in the feedback path. Periodically, or when a change is detected the hearing aid processing is halted, a white noise sequence $p[n]$ is injected into

$x[n]$ and the feedback path is identified using an adaptive filter [25]. The adaptive filter contains a linear estimate of the feedback path and can therefore not compensate for any non-linear feedback component introduced due to loudspeaker or microphone distortion. Despite accounting only for the linear contribution, this system offers good performance in quiet, but the system identification ability degrades in the presence of a competing signal, such as incoming speech.

Continuous adaptation schemes do not halt the signal processing to perform the adaptation, instead the system adapts using the speech signal presented to the user as a probe signal. Unfortunately, standard adaptive filtering algorithms perform poorly in continuous adaptation hearing aid feedback situations. Least mean square (LMS) filter adaptation is driven by the error signal, therefore frequency regions where the error signal is strong will receive the most adaptation energy reducing the power of the output signal in those regions. When speech is used as the input to an LMS-based hearing aid adaptive feedback canceler the microphone and feedback signals are correlated and the error signal contains the desired input signal. When the filter adapts, using the feedback signal, to reduce the error power, it partially cancels and distorts the desired input signal. Rather than modeling the feedback path, the adaptive filter acts as a prediction error filter and removes the correlated part of the input signal. As shown in eg. [26], under LMS adaptation the filter aims to minimize the error signal:

$$e[n] = y[n] - \mathbf{w}[n]^T\mathbf{x}[n] \tag{2.9}$$

$$= \underbrace{d[n] - \mathbf{w}[n]^T\mathbf{x}[n]} + s[n] \tag{2.10}$$

the first term corresponds to the feedback signal and is the part of the error signal we wish to drive to zero, however it is corrupted by the incoming desired signal $s[n]$,

so the filter will converge towards the biased Wiener solution:

$$\mathbf{w} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xy} \tag{2.11}$$

$$= \underbrace{\mathbf{R}_{xx}^{-1}\mathbf{r}_{xd}}_{\mathbf{w}_{opt}} + \mathbf{R}_{xx}^{-1}\mathbf{r}_{xs} \tag{2.12}$$

The bias term $\mathbf{r}_{xs} = \mathbb{E}\{x[n]s[n]\}$ is due to the correlation between the feedback signal and the incoming desired speech, which acts as a correlated disturbance signal. If the input signal is white, its auto-correlation is approximately zero for nonzero lags. Removing the auto-correlation of the input signal also removes the cross-correlation between the input and feedback signals, since the feedback signal is a delayed and scaled version of the input. In this case $\mathbf{r}_{xs} = 0$ and the adaptive filter can safely converge to an unbiased estimate of the feedback path, without distorting the input signal.

Several methods have been proposed to reduce the (linear) correlation between the desired input and the feedback signals, including performing non-linear processing in the signal path and adding de-correlating noise. The problem with both of these approaches is that the non-linear distortion or added noise levels required to de-correlate the speech and feedback signals results in severe degradation of the desired signal [6]. Since speech signal auto-correlation reduces with increasing lag, a simple but effective method is to add a delay to the signal processing path. This approach has been shown to improve convergence and reduce desired signal cancellation [26]. This approach does not achieve complete correlation as the delay must be limited to prevent comb filtering effects that occur when the delayed signal is added to the direct-path signal bypassing the hearing aid [6].

2.3 Noise Reduction

In order to remove noise from a speech signal, the speech and noise must be distinguishable in some way. Various combinations of time, frequency and statistical properties of speech and noise signals can be exploited to differentiate the signals. In the time domain speech is spontaneous and discontinuous with many pauses of varying length while many noise sources have a more continuous power envelope. Many noises also have different frequency content than speech or the time-frequency variations of the noise signal are different from speech. Speech and noise signals may also exhibit different time-correlation properties, frequently noise is considered random and non-predictable.

The most commonly exploited information is the time-frequency content of the signals, motivated by the frequency-domain transformation performed by the cochlea in the inner-ear. In speech enhancement applications the frequency division is generally achieved using fast Fourier transform (FFT) techniques to implement the discrete Fourier transform (DFT) . To account for time variations of the signal and to allow for real-time processing, the DFT is computed in overlapping windowed frames, where the frame length is sufficiently short that the speech can be considered statistically stationary for the frame duration. This short-time Fourier transform (STFT) approach gives a time-frequency division of the input signal. The N-point STFT of a signal x at time n windowed with a function w is computed as:

$$X_n(k) = \sum_{t=n-N}^{n} w[t]x[n-t]e^{-\frac{j2t\pi}{N}k}. \tag{2.13}$$

Simple speech enhancement approaches work in the STFT domain by attenuating frequency bins with low SNRs, leaving bins with high SNRs unmodified. Fig. 2.4 presents a block diagram of a STFT speech enhancement system. The clean speech

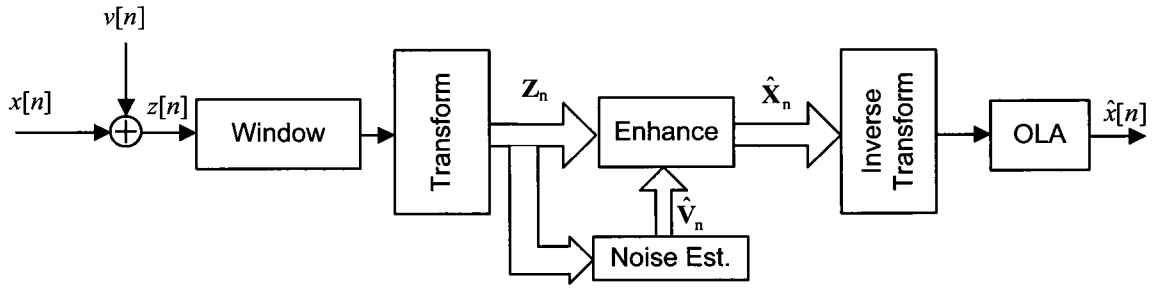


Figure 2.4: Block diagram of STFT speech enhancement.

signal $x[n]$ is mixed with additive measurement noise $v[n]$ to give the measurement signal $z[n]$. The noisy signal is enhanced in the transform domain using an estimate of the noise spectrum and the clean signal estimate is reconstructed from the inverse transformed frames using overlap-add (OLA) synthesis. STFT-based enhancement approaches exploit the fact that additive statistically independent noise in the time-domain remains additive statistically independent noise in the frequency domain; and while speech and noise overlap in time, they may not overlap in all time-frequency bins. STFT speech enhancement algorithms are commonly used because they offer consistently high noise attenuation and the availability of the FFT makes them computationally cheap to implement.

2.3.1 Spectral Subtraction

Spectral subtraction [11] is one of the earliest speech enhancement algorithms; it is also one of the simplest, both conceptually and computationally. Spectral subtraction attempts to recover the clean signal by subtracting an estimate of the noise signal in

the short-time Fourier transform domain:

$$z[n] = x[n] + v[n] \tag{2.14}$$

$$Z(k) = X(k) + V(k) \tag{2.15}$$

$$|Z(k)|^2 = |X(k) + V(k)|^2 \tag{2.16}$$

$$= |X(k)|^2 + |V(k)|^2 + 2|X(k)||V(k)| \cos(\angle X(k) - \angle V(k)). \tag{2.17}$$

If the signal and noise are in phase, then $\cos(\angle X(k) - \angle V(k)) = 1$ and

$$|Z(k)|^2 = (|X(k)| + |V(k)|)^2 \tag{2.18}$$

$$|Z(k)| = |X(k)| + |V(k)|. \tag{2.19}$$

If instead the speech and noise signal are out of phase, then $\cos(\angle X(k) - \angle V(k)) = 0$ and

$$|Z(k)|^2 = |X(k)|^2 + |V(k)|^2. \tag{2.20}$$

In the first case, the clean signal magnitude can be recovered by subtracting an estimate of the noisy signal magnitude $|\hat{V}(k)|$. In the second case the clean signal power spectrum can be recovered by subtracting an estimate of the noise power $|\hat{V}(k)|^2$. These two approaches are known as magnitude and power spectral subtraction respectively. Since the subtraction can produce negative spectral components, half-wave rectification is used to ensure a valid spectrum.

Spectral subtraction is appealing because it is both conceptually and computationally simple, as it requires only a subtraction, and because it can offer high levels of noise reduction. However, the performance of spectral subtraction relies heavily

upon the accuracy of the noise estimate, $|\hat{V}(k)|$. The original implementation in [11] used a voice activity detector (VAD) to estimate the noise spectrum during speech pauses, invoking the assumption of quasi-stationary, slowly varying noise. Since the noise estimate is a smoothed average, there will generally be a mis-match between it and the true spectrum. In addition, ignoring the cross terms in (2.17) will introduce additional error even if the noise magnitude estimate is exact. This leads to under or over-subtraction of the noise resulting in speech distortion or residual noise.

The main drawback of the spectral subtraction algorithm is the nature of the residual noise. In time-frequency regions where the noisy signal spectral amplitude is close to the estimated noise amplitude, successive over and under-estimation of the true noise spectrum leads to fluctuating, narrowband residual noise components in the enhanced speech known as musical noise or musical tones. Musical noise is problematic because it occurs in time-frequency regions where speech energy is low, so it is not masked, and its un-natural quality makes it disturbing to listeners, so that the original noisy speech is often preferable to enhanced speech with musical noise [27].

In [28] a number of modifications are proposed to reduce the level of musical noise introduced by the basic spectral subtraction algorithm. An over-estimate of the noise is used to attenuate the spurious spectral peaks that lead to musical noise, and a spectral floor was introduced to mask any residual peaks. A generalized exponent is also considered to complement the magnitude and power spectral subtraction algorithms. These modifications combined give the generalized spectral subtraction estimate of the clean speech FFT coefficient as:

$$\hat{X}(k) = \max[(|Z(k)|^{\gamma} - \alpha|V(k)|^{\gamma})^{1/\gamma} e^{j\angle Z(k)}, \beta|Z(k)|] \tag{2.21}$$

where α is the over-subtraction factor, which is generally SNR dependent, and β is the spectral floor, which is commonly frequency dependent. Due to the variance of the noise magnitude about its mean value, α has to be large to fully prevent the musical noise phenomenon; values of 3 to 6 are used in [28], corresponding to over-subtraction of up to 8 dB for power subtraction. However, this same noise spectral variance means that when the instantaneous noise spectrum is below its expected value, the desired speech will be severely distorted and low energy speech can be removed completely. The over-subtraction factor α can therefore be used to trade-off between speech distortion, musical noise artifacts and residual noise level; and different algorithms have been proposed to control the over-subtraction and spectral floor parameters to control this trade-off. Despite these attempts, while spectral subtraction can reduce the perceptual impact of noise, the modifications required to prevent musical noise result in a processed speech signal which does not improve and can actually reduce intelligibility compared to unprocessed speech [27].

Spectral subtraction algorithms only enhance the speech magnitude or power spectrum, the noisy signal phase spectrum is used to reconstruct the clean signal phase. This is often justified by noting that improving the phase estimate has a relatively small impact on overall SNR speech quality compared to improving the magnitude spectrum estimate [29], and that the perceptual impact of phase distortion is lower than magnitude distortion [30]. However, the presence of cross-terms in (2.17) means that a perfect estimate of the clean signal spectrum cannot be obtained by subtracting the noise spectrum from the noisy signal spectrum, even if a perfect estimate of the noise spectrum is available.

2.3.2 Statistical Model Speech Enhancement

Spectral subtraction can be viewed as a deterministic speech enhancement approach. It does not consider statistical fluctuations in the noise magnitude, but rather it assumes that the noise estimate is exact and can be used to directly compensate for a measurement bias in the noisy speech. A different approach is taken by statistical model speech enhancement algorithms, which treat the speech signal and noise as random processes and formulate estimators for the clean signal by minimizing a statistical distortion measure.

While spectral subtraction works with the single frame (periodogram) estimate of the speech spectrum, statistical model approaches use smoothed estimates of the speech and noise spectra, $\lambda_x(k) = \mathbb{E}\{|X(k)|^2\}$, and $\lambda_v(k) = \mathbb{E}\{|V(k)|^2\}$. Using expectations rather than direct measurements of the quantities of interest results in lower variance estimates of the enhanced signal, helping to reduce musical noise. While the signals are not stationary, they are assumed to be short-time stationary, and estimation of the required expectations is a trade-off between tracking the time-varying characteristics and obtaining low-variance estimates.

Using expectations rather than individual measurements means that rather than assuming *geometric* orthogonality of the clean speech and noise signals in a given frame, statistical approaches assume *statistical* orthogonality – independence – of the speech and noise signals:

$$\mathbb{E}\{Z(k)^2\} = \mathbb{E}\{X(k) + V(k)^2\} \tag{2.22}$$

$$= \mathbb{E}\{X(k)^2\} + \mathbb{E}\{V(k)\}^2 + \cancelto{0}{2\mathbb{E}\{X(k)V^*(k)\}} \tag{2.23}$$

$$= \mathbb{E}\{X(k)^2\} + \mathbb{E}\{V(k)\}^2 \tag{2.24}$$

STFT enhancement algorithms are most commonly expressed in terms of the *a-priori* and *a-posteriori* SNRs defined as:

$$\xi(k) = \frac{\lambda_x(k)}{\lambda_v(k)} \tag{2.25}$$

$$\gamma(k) = \frac{|Y(k)|^2}{\lambda_v(k)} \tag{2.26}$$

While $\gamma(k)$ can be computed directly, $\xi(k)$ requires knowledge of the clean signal.

Wiener Filter

A common choice for a statistical distortion function is the mean-squared error (MSE) between the true clean signal (in the time or frequency domain) and its estimated value:

$$E(k) = \mathbb{E}\{(X(k) - \hat{X}(k))^2\} \tag{2.27}$$

Wiener filters provide the a linear estimator of the clean signal complex DFT coefficient:

$$\hat{X}(k) = H(k)Z(k) \tag{2.28}$$

that minimizes the MSE cost function:

$$J(H) = \mathbb{E}\{(X(k) - H(k)Z(k))^2\} \tag{2.29}$$

$$J(H) = \mathbb{E}\{X^2(k)\} - |H(k)|^2\mathbb{E}\{Z^2(k)\} - H^*(k)\mathbb{E}\{X(k)Z^*(k)\} - H(k)\mathbb{E}\{X^*(k)Z(k)\} \tag{2.30}$$

$$= \lambda_x(k) - |H(k)|^2(\lambda_x(k) + \lambda_v(k)) - (H^*(k) + H(k))\lambda_x(k) \tag{2.31}$$

taking the derivative with respect to $H(k)$ and setting it to zero gives the Wiener

estimator:

$$H(k) = \frac{\lambda_x(k)}{\lambda_x(k) + \lambda_v(k)} \tag{2.32}$$

Alternatively, in terms of the *a-priori* SNR:

$$H(k) = \frac{\xi(k)}{1 + \xi(k)} \tag{2.33}$$

If the real and imaginary parts of the speech and noise DFT coefficients are assumed to be independent and jointly Gaussian, the MSE optimal linear estimate provided by the Wiener filter is the minimum MSE (MMSE) estimator among all linear or non-linear estimators. If the statistical constraints are not met, the Wiener filter still provides the MSE optimal linear estimate, but there may be better performing non-linear estimators.

In [31] it is shown that if the maximum likelihood estimate of the *a-priori* SNR, given as:

$$\xi_{ML}(k) = \frac{|Z(k)|^2 - \lambda_v(k)}{\lambda_v(k)} \tag{2.34}$$

is used in the Wiener filter expression (2.33), then gain function used to estimate the clean signal:

$$H_n(k) = \frac{|Z_n(k)|^2 - \lambda_v(k)}{|Z_n(k)|^2} \tag{2.35}$$

is the square of the power spectral subtraction gain function. The Wiener gain function therefore offers more attenuation for a given SNR than power spectral subtraction and also risks introducing more speech distortion if the SNR estimate is inaccurate.

In [32] a speech enhancement algorithm using Wiener filtering in the discrete cosine transform (DCT) domain was proposed and evaluated. The Wiener filter expressions in the DCT domain are the same as those in the DFT domain. In [32] it is argued that the DCT is well suited to speech processing as it provides better de-correlation and

energy compaction of speech than the DFT, validating the assumption that individual bands are independent and can be enhanced separately, while DCT transformed noise is approximately white in each band, simplifying its removal. Furthermore, as the DCT is a real transform the estimation error incurred by using the phase from the noisy signal coefficients will be lower for DCT-domain estimation than DFT-domain. The results presented in [32] show the DCT Wiener filter providing comparable SNR performance to the non-linear DFT spectral amplitude estimator described in section 2.3.2.

MMSE-STSA

Under the Gaussian assumptions, the Wiener filter is the optimal estimator of the complex DFT coefficients. However, since spectral amplitude plays a dominant role in determining the perceived quality of the enhanced speech, in [12] an MMSE short-time spectral amplitude (MMSE-STSA) estimator is derived. A shortened version of the derivation will be presented here, as the steps in the derivation and the role and effects of the assumed statistical model will be relevant for later discussions.

To facilitate the derivation, the following assumptions are made in [12]:

- The speech DFT coefficients are statistically independent. This permits each spectral amplitude to be estimated individually.
- The real and imaginary components of the speech and noise DFT coefficients are independent zero-mean Gaussian distributed random variables with time-varying variances.

Let $A_n(k) \triangleq |X_n(k)|$ be a shorthand notation treating the spectral amplitude as a random variable, and let a_k be a realization of $A_n(k)$. The MMSE-STSA treats STFT enhancement as a Bayesian estimation problem, so the estimate is the mean

(expectation) of the posterior probability density function (PDF) $p(A_n(k)|Z_n(k))$, which can be expressed using Bayes rule as:

$$p(A_n(k)|Z_n(k)) = \frac{p(Z_n(k))|A_n(k))p(A_n(k))}{\int_0^\infty p(Z_n(k)|a_k)p(a_k)da_k} \tag{2.36}$$

expanding the expectation in terms of the amplitude $A_n(k)$ and the phase θ_x:

$$|\hat{X}_n(k)| = \mathbb{E}\{A_n(k)|Z_n(k)\} \tag{2.37}$$

$$= \frac{\int_0^\infty a_k p(Z_n(k)|a_k)p(a_k)da_k}{\int_0^\infty p(Z_n(k)|a_k)p(a_k)da_k}\} \tag{2.38}$$

$$= \frac{\int_0^\infty \int_0^{2\pi} a_k p(Z_n(k)|a_k,\theta_x)p(a_k,\theta_x)d\theta_x da_k}{\int_0^\infty \int_0^{2\pi} p(Z_n(k)|a_k,\theta_x)p(a_k,\theta_x)d\theta_x da_k}. \tag{2.39}$$

The terms in the expectation can be determined using the assumed Gaussian statistical model, and the integrals can be solved in closed form using the Gamma function $\Gamma(\cdot)$ and the confluent hypergeometric function $\Phi(a,b;c)$ giving the non-linear MMSE-STSA estimate:

$$|\hat{X}_n(k)| = \sqrt{\lambda_k}\Gamma(1.5)\Phi(-0.5,1;-v_n(k)) \tag{2.40}$$

where

$$\lambda_k = \frac{\lambda_x(k)}{1+\xi(k)} \tag{2.41}$$

$$v_k = \frac{\xi(k)}{1+\xi(k)}\gamma(k) \tag{2.42}$$

The expression in (2.40) is commonly re-written as a gain function to facilitate comparisons with the Wiener, spectral subtraction and other gain functions. In this form:

$$\hat{X}_n(k) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma(k)}e^{-v_k/2}\left[(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})\right]Y_n(k) \tag{2.43}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the zeroth and first order modified Bessel functions of the first kind.

Log-MMSE STSA

In [33] another STSA estimator is derived that minimizes the MSE between the log-spectra of the clean and enhanced signals, on the grounds that differences in the log spectrum are more perceptually relevant for human listeners. Using the same Gaussian and independence assumptions as the standard MMSE-STSA, the log-MMSE STSA estimate can be shown to be:

$$\hat{X}_n(k) = \exp(\mathbb{E}\{\log A_n(k)\big|Y_n(k)\}) \tag{2.44}$$

$$= \frac{\xi(k)}{1+\xi(k)}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}Y_n(k) \tag{2.45}$$

The log-MMSE STSA gain function provides more attenuation than the MMSE-STSA for the same values of ξ and γ, providing more noise attenuation if the SNR estimates are accurate.

SNR Estimation

The work in [12] also introduced a recursive estimator for *a-priori* SNR, known as the "decision-directed" approach. The maximum likelihood (ML) SNR estimate (2.34) is based on the periodogram estimate of $X_n(k)$, and is therefore a high variance estimate,

making the enhanced signal susceptible to musical noise. To compensate for this the decision-directed approach creates a smoothed SNR estimate by convexly combining the ML estimate with an estimate obtained from the enhanced signal in the previous frame:

$$\xi_n(k) = \alpha \frac{|\hat{X}_{n-1}(k)|^2}{\lambda_v(k)} + (1-\alpha)\max\left[\gamma_n(k) - 1, 0\right] \tag{2.46}$$

The max function in the second term is a half-wave rectification to ensure $\xi_n(k)$ remains positive, as with spectral subtraction a spectral floor could be used instead. The weighting factor α determines the relative weight of the two terms, and acts as a recursive smoothing parameter. In [12] a value of $\alpha = 0.98$ is suggested, offering a high degree of smoothing. It is important to note that although the MMSE-STSA and log-MMSE-STSA estimators were derived assuming independence of successive speech frames; the decision-directed estimator uses the enhanced signal from the previous frame to estimate the SNR in the current frame, thereby assuming a dependence between the two.

The MMSE and log-MMSE STSA estimators both offer good noise reduction performance without suffering from the musical noise phenomenon present in spectral subtraction. In [8] it was shown that the suppression of musical noise is the result of several factors, with the dominant factor being the smoothed SNR estimate provided by the decision-directed estimator. At low SNRs, the decision-directed *a-priori* SNR estimate is a highly smoothed version of the ML estimate. In contrast, at high SNRs there is little attenuation, so $\hat{X}_n(k) \approx Z_n(k)$ and $\xi_{n-1}(k) \approx \gamma_n(k)$, so the decision-directed estimate tracks the ML estimate. The smoothing parameter α in (2.46) controls the smoothing at low SNRs. A value close to 1 is desired to provide a smooth SNR estimate, however this can also lead to over-attenuation of the desired signal,

especially during abrupt changes such as speech onsets. Another factor contributing to musical noise reduction is SNR flooring. The original MMSE-STSA restricted the range of the *a-priori* and *a-posteriori* SNRs to -15 –15 dB; this is equivalent to spectral flooring in spectral subtraction, where the floor is a function of the noisy signal magnitude. The musical noise suppression capability of the decision-directed estimator has been exploited by other algorithms, such as the DCT Wiener filter in [32].

By recursively smoothing the SNR estimate the decision-directed estimator implicitly assumes an inter-frame correlation, however the smoothing is done heuristically and the choice of α is made through experimentation to optimize trade-off between musical noise and smoothing of speech onsets and low energy segments. In [34] an attempt is made to explicitly recognize inter-frame correlation and replace the heuristic choice of α with a statistically motivated one.

In [34] a Gaussian model is used to derive an expression for the conditional variance of the spectral power $\lambda_x(k)$, given the set of measurements. The conditional variance is assumed to be time-varying and is estimated using propagate and update steps, analogous to Kalman filtering. The *a-priori* SNR estimate is obtained by dividing the conditional variance estimate by the noise power estimate. The SNR estimate is given by:

$$\xi_{n|n-1}(k) = \max\left[\beta\frac{|\hat{X}_{n-1}(k)|^2}{\lambda_v(k)} + (1-\beta)\xi_{n-1|n-1}(k), \xi_{min}\right] \tag{2.47}$$

$$\xi_{n|n}(k) = \frac{\xi_{n|n-1}(k)}{1+\xi_{n|n-1}(k)}\left(1 + \frac{\xi_{n|n-1}(k)}{1+\xi_{n|n-1}(k)}\gamma_n(k)\right) \tag{2.48}$$

where β is a parameter related to the stationarity of the spectral power process: a large value of β indicates a non-stationary process, and a low inter-frame correlation, while

a small value indicates a stationary process. Results in [34] show that the predict-update approach can track changes in the SNR faster than the decision-directed estimator, but this did not lead to any improvement in SNR or log spectral distortion scores.

Estimators with Non-Gaussian Speech Priors

The choice of a Gaussian distribution for speech and noise DFT coefficients is generally made to ease the derivation of a closed form solution, and is supported by the central limit theorem, since the DFT coefficient is a weighted sum of random variables, ie. the time-domain samples [12]. Although the time-domain samples are not independent, it is assumed that the span of their correlation is insignificant compared to the frame length, so they are only weakly dependent and the central limit theorem still holds. The validity of the Gaussian model is challenged in [35] and [36] on the basis that the required assumptions regarding speech sample correlation are not met in practice. Both works carry out analyses that shows speech DFT coefficients are better modeled by heavy-tailed *supergaussian* distributions such as the gamma and Laplace distributions.

In [35], MMSE-optimal complex DFT estimators are derived assuming the real and imaginary parts of the speech and noise coefficients are independent and Laplace or gamma distributed. While the Wiener filter provides the MSE-optimal linear estimate, the derived estimators are non-linear. The results presented in [35] show that the new estimators offer only modest segmental SNR improvements compared to the Wiener filter, and the improvements are not consistent across SNR and noise conditions.

Owing to the difficulties in obtaining an MMSE-estimate of the super-Gaussian amplitudes, in [36] a maximum *a-posteriori* (MAP) spectral amplitude estimator is

derived that models the speech spectral amplitude with a parametric probability density function that can approximate the density arising from Laplace or gamma distributed real and imaginary parts of the speech DFT coefficients; the noise is assumed to follow the standard Gaussian model. The derived estimators are found to offer comparable sound quality and a modest SNR improvement compared to the MMSE-STSA.

Despite the fact that the Laplace and gamma distributions are shown to fit the speech data much better than the Gaussian distribution, the improvements offered by the non-Gaussian estimators is generally modest and inconsistent. One possible reason for this is that order to make the derivation tractable, the real and imaginary components of the DFT coefficients are assumed to be statistically independent. Under the Gaussian assumption this also means that the amplitude and phase are independent, however for non-Gaussian distributions the amplitude and phase, and real and imaginary parts cannot be simultaneously independent [35]. The use of non-Gaussian distributions therefore removes the assumption of amplitude/phase independence. In [35] it is acknowledged that Kullback-Liebler discrimination information reveals that the amplitude and phase are more independent than the real and imaginary parts.

Another possible reason for the relatively small benefit is that the non-Gaussian models may not be a better fit to the actual speech DFT coefficient distribution. The support for the non-Gaussian distributions comes from fitting histograms to large data sets, however this approach assumes speech signals to be ergodic, which is clearly not the case [12]. In fact the observed histograms may not conflict with the Gaussian model assumed in [12]. As shown in [37] in practice the speech signal variance is assumed known (via the *a-priori* SNR), so the Gaussian model is actually

a conditionally Gaussian model, conditioned on the signal power. Assuming an exponential distribution for the DFT coefficient variance results in a marginal Laplace distribution for the conditionally Gaussian speech DFT coefficients.

2.3.3 Signal Subspace

The signal subspace family of algorithms relies on linear algebra techniques, rather than statistical estimation, to separate the speech and noise signals. The noisy speech vectors define a Euclidian vector space which is decomposed into signal plus noise and noise-only subspaces, either through singular value decomposition (SVD) of the noisy signal data matrix [38] or eigen-decomposition (ED) of the covariance matrix [39]. By nulling the noise-only subspace and constructing a reduced-rank approximation of the clean signal from the remaining signal plus noise subspace, some noise attenuation can be achieved with no speech distortion. Additional noise attenuation can be achieved by compensating the singular or eigen-values corresponding to the signal plus noise subspace, allowing this approach to effectively trade-off signal distortion and residual noise levels. However, the compensation combined with frame-to-frame noise power fluctuations can result in musical noise. While the original algorithms assumed white noise, they have since been extended to colored noises. For example [40] proposes a generalized version of the algorithm in [39] that projects the signal vector using a non-unitary transform that simultaneously diagonalizes the noisy signal and noise covariance matrices, pre-whitening the noise signal in the projection process.

2.3.4 Speech Production Model-based

STFT-based methods are the most common speech enhancement approach because of computational simplicity;the assumption of inter-frame independence allows closed

form solutions, contributing to the simplicity. However, a lack of inter-frame constraints leads to musical noise, ad-hoc smoothing methods smooth speech onsets and can eliminate low energy speech regions. STFT methods are essential non-parametric spectrum estimation methods, the harmonic nature of speech means that parametric (model-based) methods can be used to obtain lower variance estimates. Furthermore, speech production-based parametric modelling is used in high-quality low-rate speech coding.

Kalman Filtering

In [13] an auto-regressive (AR) model for speech is assumed. The speech signal at time n, x_n is assumed to be corrupted by additive measurement noise v_n giving the measurement signal z_n:

$$z_n = x_n + v_n \tag{2.49}$$

The clean speech signal is assumed to follow a p^{th} order AR model:

$$x_n = \sum_{i=1}^{p} a(i)x_{n-i} + d_n \tag{2.50}$$

$$= \mathbf{a}^T\mathbf{x}_n + d_n \tag{2.51}$$

where:

$$\mathbf{x}_n = [x_{n-1} \dots x_{n-p}]^T \tag{2.52}$$

is the signal vector,

$$\mathbf{a} = [a(1) \dots a(p)]^T \tag{2.53}$$

is the vector of auto-regression coefficients, and d_n is the excitation sequence.

Defining the state-transition and measurement matrices $\mathbf{F}_n$ and $\mathbf{G}$, (2.50) can be

described by the following linear state-space model:

$$\mathbf{F} = \begin{bmatrix} & \mathbf{a}^T & \\ \mathbf{I}_{p-1} & & \mathbf{0}_{p-1\times 1} \end{bmatrix} \tag{2.54}$$

$$\mathbf{G} = \begin{bmatrix} 1 \\ \\ \mathbf{0}_{p-1\times 1} \end{bmatrix} \tag{2.55}$$

$$\mathbf{x}_n = \mathbf{F}\mathbf{x}_{n-1} + \begin{bmatrix} d_n \\ \\ \mathbf{0}_{p-1\times 1} \end{bmatrix} \tag{2.56}$$

$$z_n = \mathbf{G}\mathbf{x}_n + v_n \tag{2.57}$$

where (2.56) and (2.57) are the state-space transition and measurement equations respectively.

If the process and measurement disturbances d_n and v_n are Gaussian and independent of the state, and if the matrices $\mathbf{F}, \mathbf{G}$ as well as the disturbance variances σ_d^2 and σ_v^2 are known, the Kalman filter can be used to obtain an MMSE-optimal estimate of the clean signal state $\mathbf{x}_n$. The Kalman estimate is unbiased and is a minimum variance estimate, and the Kalman filter is compact in that it only calculates and propagates the mean and covariance matrix of the state vector. Under the Gaussian assumption, these are sufficient statistics, ie. they completely describe the density, so any statistical question regarding the state can be answered using them. The Kalman filter uses a predictor-corrector algorithm to recursively estimate a state vector from a series of noisy measurements. An estimate (prediction) of the state at time n is made using the measurements up to time $n-1$; once the measurement at time n is

available, the estimate is corrected to reflect the new information.

The Kalman filter equations are:

$$\hat{\mathbf{x}}_{n|n-1} = \mathbf{F}\hat{\mathbf{x}}_{n-1} \qquad \text{Propagate the state mean}$$

$$\mathbf{P}_{n|n-1} = \mathbf{F}\mathbf{P}_{n-1}\mathbf{F}^T + [\sigma_d^2\mathbf{I} \qquad \text{Propagate the covariance matrix}$$

$$\mathbf{K}_n = \mathbf{P}_{n|n-1}\mathbf{G}^T(\mathbf{G}\mathbf{P}_{n|n-1}\mathbf{G}^T + \sigma_v^2\mathbf{I})^{-1} \qquad \text{Compute the Kalman gain}$$

$$\mathbf{P}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{G})\mathbf{P}_{n|n-1} \qquad \text{Update the covariance matrix}$$

$$\hat{\mathbf{x}}_n = \hat{\mathbf{x}}_{n|n-1} + \mathbf{K}_n(\mathbf{y}_n - \mathbf{G}\hat{\mathbf{x}}_{n|n-1}) \qquad \text{Update the state estimate}$$

The Kalman filter approach to speech enhancement has several advantages. Operating in the time domain and imposing an autoregressive speech production model on the speech signal results in natural sounding enhanced speech and residual noise, without any musical noise or robotic sounding artifacts. The use of the Kalman filter enables the modeling of non-stationary characteristics, unlike the Wiener filter which assumes signal stationarity. The main obstacle to using Kalman filtering for speech enhancement is obtaining an estimate of the clean speech AR coefficients $\mathbf{a}$ and the clean speech and measurement noise excitations σ_d^2 and σ_v^2 from the noisy speech signal. In the original Kalman filter speech enhancement work [13] the AR parameters and the speech and noise excitations are measured directly from the unmixed signals. The work in [14] proposes an iterative parameter estimation method that alternated between enhancing the signal and estimating the AR parameters, iteratively converging to an estimate of the clean signal parameters. This work also uses the colored noise model Kalman filter: using an augmented state matrix with an AR noise model allows the Kalman filter to model non-white measurement noises, greatly improving the performance in real life noise environments. The colored noise approach and its

associated parameter estimation method are extended in [41] where the Expectation-Maximization framework is used to estimate the clean signal parameters. The work in [42] jointly performs the signal and parameter estimation using dual Kalman filters.

The Kalman filter speech enhancement algorithm has also been extended from the time domain into transform domains. In subband Kalman filtering [43] [44], independent Kalman filters are run on band-pass filtered speech signals, then recombined to produce the enhanced fullband signal. The subband decomposition allows a high order fullband signal to be modeled using much lower order subband AR models. For example the 4 and 8 subband system of [43] uses 0^{th} order (white) models for speech and noise, while the fullband implementation of [14] uses AR(10) speech and AR(4) noise models. Subband Kalman filtering has been shown to offer much lower complexity and higher levels of noise attenuation than fullband Kalman filtering [43] [44]. The orthogonal band decomposition means that the signal does not follow the fullband speech production model so there is a risk of un-natural sounding artifacts; however the AR model in each subband smooths the subband signal trajectories, reducing musical noise [44].

2.3.5 Binaural and Multi-channel Speech Enhancement

Multi-microphone Noise Reduction

The availability of multiple spatially separated measurements opens up another possibility of separating speech and noise. Monaural systems rely on time, frequency and statistical differences between speech and noise; binaural systems with spatially separated microphones can exploit spatial differences as well. If the noise in the multiple channels is weakly correlated while the speech is correlated, the speech and noise estimates can be combined to obtain a single improved estimate.

In [45] the MMSE and log-MMSE spectral amplitude estimators are generalized to the multi-channel case. The estimator requires the transfer functions from the source to the microphones, or the inter-channel transfer functions. An adaptive estimation scheme for estimation of the inter-channel transfer functions is presented. In experiments using a tetrahedral microphone array with 29 cm microphone spacing, the system shows modest improvements in terms of SNR and spectral distortion compared to the single channel MMSE-STSA estimator. It is also reported that while the results from the single and multi-channel algorithms all contain musical noise, the musical noise level decreases as the number of channels is increased from 1 to 2 to 4; reflecting the lower variance estimate obtained with the multiple sensors. In [46] the Gaussian signal and noise model of [12] is extended to the multi-channel case. The multi-channel speech signals are assumed to be highly correlated jointly Gaussian random variables, while the noise is assumed Gaussian and uncorrelated between channels. In practice there will be some inter-channel noise correlation even for diffuse noise. An MMSE estimator is derived for frontal targets where the spectral amplitudes in all channels are assumed identical and an MAP estimator is derived for the direction-independent case, where a closed-form MMSE solution is intractable. Experiments using a 4-microphone array with 12 cm inter-microphone spacing show the multi-channel estimators offering higher noise reduction and lower speech distortion compared to the standard single channel MMSE-STSA estimator. In cafeteria babble noise the uncorrelated noise assumption breaks down, and the noise reduction improvement is decreased, but the speech quality remains higher than the single channel case. In [47] a multi-input single output Kalman filter enhancement scheme is proposed. The system uses a standard multi-channel Kalman filter, and makes the restrictive assumptions of knowledge of the transfer functions between the source

and noise signals to both microphones. All of these approaches are are general multi-channel and bilateral estimators, they are designed for microphone arrays do not exploit a-priori information about binaural listening situations.

In [48] a binaural speech enhancement system is presented based on multi-channel Wiener filtering. The left and right channel signals (possibly multi-channel) are stacked to create a multi-channel signal vector. The left and right signals are jointly enhanced using a least squares estimate of the multi-channel Wiener filter. The stacked noise correlation matrix is estimated during speech silence intervals so the noise is assumed to be quasi-stationary. A constraint is added to the least square solution to ensure preservation of the interaural time difference (ITD) cues. In [49] the system is extended to include the full interaural transfer function, thereby preserving ITD as well as interaural level difference (ILD) cues. Both implementations require estimates of the clean speech and noise correlation matrices, which are very difficult to obtain in practice. The noise correlation matrix is estimated during speech pauses, assuming quasi-stationarity of the noise and the presence of a perfect VAD. This estimate can be used to compensate the noisy speech correlation matrix to estimate the clean speech, but this can lead to musical noise and speech distortion in fluctuating noise.

Source-microphone transfer functions change frequently and rapidly with any change in the relative positions of the source and the microphones. This is especially problematic in a hearing aid application where the microphones are mounted on the user, so the paths are impacted by movements of both the source and the receiver. The variability of and difficulty in estimating these transfer functions precludes the hearing aid use of any enhancement schemes that require them.

Spatial Filtering

The human auditory system exploits the information provided by our binaural hearing system to perform spatial filtering, allowing a listener to focus on a target speaker while suppressing non-co-located interference sources. Interaural time differences and frequency dependent head-shadowing cues are used to provide a release from masking of up to 10 dB [9]. For hearing impaired listeners the benefits of these cues are diminished due to reduced time and frequency resolution. Multi-channel and binaural noise reduction systems create the possibility for a speech enhancement system to perform spatial filtering to restore some of the benefit.

In current hearing aids directional microphone arrays are used to perform this spatial filtering [23]. The directivity pattern of the array is chosen to orient forwards, enabling suppression of noise sources lateral to and behind the listener. This approach has the benefit of offering no distortion of sources in the target direction and has been shown to improve intelligibility [50]. Standard directional filtering is bilateral, with one array per ear, and requires synchronization between the ears to ensure the arrays are focused on the same direction. It also makes no distinction between desired targets and noise sources, attenuating any non-frontal sounds; this can be undesirable and even dangerous in, for example, a city street environment where a user would want to hear lateral sounds such as car horns.

Another approach to spatial filtering is to exploit known spatial characteristics of the noise field. Unlike diffuse noise, directional targets such as speech have a constant phase delay between the channels, so the magnitude-squared coherence (MSC) will approach unity across all frequencies, even if the target is non-frontal. A high MSC therefore indicates a directional target, like speech, which should be passed without attenuation; while a low coherence indicates no directional target, or a low SNR, so the signal at that frequency should be attenuated. This approach was first described

in [51] for reduction of room reverberation effects, and is also is the approach taken in [52], where functions of the coherence function are used as gain factors to filter the noisy speech. In [53] the approach was extended to coherent noise by estimating the noise cross-correlation during silence periods. The coherence function is attractive from signal processing perspective as it can be measured directly without knowledge of the desired signal, unlike the *a-priori* SNR. In practice reverberation reduces the coherence of clean speech, and realistic noises are not incoherent and on-line estimation of the noise cross-correlation shares the same difficulties as noise power spectrum estimation.

In [54] a diffuse noise field microphone array Wiener post-filter is derived. By invoking the known diffuse noise field coherence an expression for the clean speech PSD is obtained and used in a Wiener filter expression. This post-filter is shown to offer a significant improvement in noise reduction performance compared to a comparable post-filter assuming incoherent noise.

A different approach was taken in [55]. Analyzing the theoretical autocorrelations of the left and right channel signals using the HRTFs, it is noted that for coherent speech the signal in one channel is well predicted by the other channel, so the inter-channel prediction error is the result of incoherent noise. By manipulating the inter-channel prediction error an estimate of the diffuse noise field power spectrum can be obtained, and incorporating a modified free-field diffuse noise coherence function can compensate for the noise correlation at low frequencies.

In [56] a strategy selective binaural speech enhancement system is proposed and evaluated. The system switches between algorithms based on an estimate of the diffuseness of the environment. When mostly diffuse noise is detected coherence-based processing is applied. The magnitude of the coherence function is used directly as

a filter, without compensating for low frequency noise correlation, and adaptive estimation of the ILD is used to compensate for non-frontal targets. For directional noise from an interfering target, Wiener-type filtering is performed to attenuate the interference. Computation of the Wiener gains requires the interaural level differences for both the target and interference sources; these gains are estimated when only one of the sources is active and correlation processing is used to estimate which source is active. This type of noise reduction would break down in the case of multiple or moving interference sources. To improve the performance in the presence of multiple interferers, directional filtering is applied based on an assumed range of desired target locations. In experimental tests in cafeteria situations, where only the diffuse noise component was active, hearing impaired listeners subjectively preferred the strategy selective enhanced signal over the noisy signal, but no significant intelligibility improvement was found. In contrast, when a directional interference was present, users indicated preference for the noisy speech as often as the processed signal.

2.4 Particle Filtering

Particle filters use Monte Carlo simulation techniques to perform recursive Bayesian estimation of the state of a system described by a general state space model:

$$\mathbf{x}_n = \mathbf{f}_n(\mathbf{x}_{n-1}, \mathbf{d}_n) \tag{2.58a}$$

$$\mathbf{z}_n = \mathbf{g}_n(\mathbf{x}_n, \mathbf{v}_n). \tag{2.58b}$$

where $\mathbf{x}_n$ is the state, $\mathbf{z}_n$ is the measurement and (2.58a) and (2.58b) are referred to as the transition and measurement equations. The system and measurement noise $\mathbf{d}_n$ and $\mathbf{v}_n$ are assumed to be independent of $\mathbf{x}_n$, however they need not be Gaussian

distributed. In general we are interested in the filtered posterior distribution of the state vector $p(\mathbf{x}_n|\mathbf{z}_{1:n})$; this can be represented/approximated by a set of samples (particles) $\{\mathbf{x}_n^{(i)}\}_{i=1}^{N_p}$ in the state space, and their associated normalized probability weights $\{W_n^{(i)}\}_{i=1}^{N_p}$ as:

$$p(\mathbf{x}_n|\mathbf{z}_{1:n}) \approx \sum_{i=1}^{N_p} W_n^{(i)} \delta(\mathbf{x}_n - \mathbf{x}_n^{(i)}). \tag{2.59}$$

This pointwise representation of the posterior can then be used to obtain the MMSE posterior estimate of the clean signal:

$$\hat{\mathbf{x}}_n = \sum_{i=1}^{N_p} W_n^{(i)} \mathbf{x}_n^{(i)}. \tag{2.60}$$

2.4.1 Sequential Importance Sampling and Resampling

If we were able to draw samples from the true filtered distribution $\mathbf{x}_n^{(i)} \sim p(\mathbf{x}_n|\mathbf{z}_{1:n})$, we could use those samples along with equal weights $W_n^{(i)} = 1/N_p$ in (2.59) to represent the distribution. Unfortunately this distribution is usually not available to be sampled from, instead the particles are drawn from an importance distribution, $q(\mathbf{x}_n|\mathbf{x}_{0:n-1}, \mathbf{z}_{1:n})$ and the weights are computed as [57]:

$$W_n^{(i)} = W_{n-1}^{(i)} \frac{p(\mathbf{z}_n|\mathbf{x}_n^{(i)}) p(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)})}{q(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)}, \mathbf{z}_n)} \tag{2.61}$$

The only requirement of the sampling distribution is that the support of the true density must be a subset of the importance density support. However, the closer the shape of the importance density is to the true density, the more efficient the sampling; if the importance distribution is much broader than the true distribution many low-weight particles will be generated that contribute little to the overall density.

In order to prevent particle degeneracy, a re-sampling step is included to discard low-weight particles and propagate particles with high weights. Resampling algorithms generate a new set of samples by sampling with replacement from the original set, so that the distribution of the new samples is proportional to the normalized importance weights of the original samples. The resampled particles are assumed to be an N_p sample draw from the posterior discrete density represented by the original (non-resampled) set of particles, so the importance weights of the re-sampled particles are all equal. If re-sampling is carried out at every iteration, the term $W_{n-1}^{(i)}$ in (2.61) is a common factor and need not be included.

A common, though not optimal, choice for the importance distribution is the transitional prior: $q(\mathbf{x}_n|\mathbf{x}_{0:n-1}^{(i)}, \mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$ [57]. This choice is generally made for simplicity as the particle update step reduces to generating a sample from the process noise distribution and propagating the particle and the generated noise sample through the state update equation. Assuming resampling is carried out at each iteration, the weight update simplifies to:

$$W_n^{(i)} = p(\mathbf{y}_n|\mathbf{x}_n^{(i)}). \tag{2.62}$$

An recursive estimation algorithm, using sequential importance sampling of the transitional prior density and resampling at each interval was introduced in [58] under the name "bootstrap filter", and was applied with great success to a non-linear target tracking scenario. The bootstrap filter has since been recognized as a special case of the particle filter framework, and is generally referred to as the Sampling Importance Resampling (SIR) particle filter.

The transitional density used by the SIR particle filter does not make use of the measurement $\mathbf{z}_n$ in generating new particles making the algorithm vulnerable

to particle depletion when the likelihood is peaked and outliers are present. One approach to improving the sampling efficiency is to use an auxiliary SIR particle filter [59]. Auxiliary particle filters use the predictive likelihood to re-sample particles before the propagation step, resulting in more particles in high likelihood regions of the state space. Taking the measurement into account in the sampling step produces an adapted particle filter [59], which generally improves sampling efficiency.

2.4.2 Rao-Blackwellization

When the dimension of the state space is high, many particles are required to adequately model the state distribution and achieve good performance. Rao-Blackwellization is a technique that exploits the structure of the state-space model to reduce the dimension of the sampling space, improving sampling efficiency and achieving a lower variance estimate for a given number of particles [60]. Consider the case where the state vector can be partitioned into two sub-states:

$$\tilde{\mathbf{x}}_n = [\mathbf{x}_n, \theta_n]. \tag{2.63}$$

The chain rule of probability allows the joint posterior to be decomposed as:

$$p(\mathbf{x}_n, \theta_n | \mathbf{z}_{1:n}) = p(\mathbf{x}_n | \theta_n, \mathbf{z}_{1:n}) p(\theta_n | \mathbf{z}_{1:n}). \tag{2.64}$$

If $p(\mathbf{x}_n | \theta_n, \mathbf{z}_{1:n})$ can be computed analytically (eg. with a Kalman filter), then Monte Carlo inference need only be carried out to estimate $p(\theta_n | \mathbf{z}_{1:n})$, which is of a lower dimension than $p(\tilde{\mathbf{x}}_n | \mathbf{z}_{1:n})$, resulting in a lower variance estimate of the combined state vector $\tilde{\mathbf{x}}_n$ using fewer particles.

2.4.3 Particle filter speech enhancement

Provided with an appropriate generalized state-space model, particle filter algorithms can naturally account for dynamic behavior and non-Gaussian statistics of speech and noise; consequently a variety of particle filter approaches have been proposed for various speech processing applications.

A class of particle filter algorithms has been developed to compensate for background noise corrupting log spectral speech features in automatic speech recognition (ASR) systems. In [61] a Gaussian random walk is used to model the noise process and the results demonstrate that using the dynamic model, rather than assuming stationary noise, results in significant performance gains on recognition tasks. In [62] a first order AR model is used in place of the random walk. The AR model uses a full AR matrix which includes off-diagonal terms to capture correlation across both time and frequency. In [63] higher order AR noise models are considered, but models greater than AR(1) are not found to yield additional gains. In [64] the approach is further improved through the use of dynamic AR models; and an algorithm is presented to jointly compensate for noise and room reverberation, increasing performance over systems that treat the distortions independently.

Particle filter approaches have also been applied to the recovery of the clean speech waveform for human listeners. In [65] and [60] a time-varying auto-regressive (TVAR) model is assumed for the audio signal in the time domain and Rao-Blackwellized particle filter (RBPF) enhancement is applied. The model-based approach results in enhanced speech that is free from musical noise, with residual noise that retains the character of the original distortion [66].

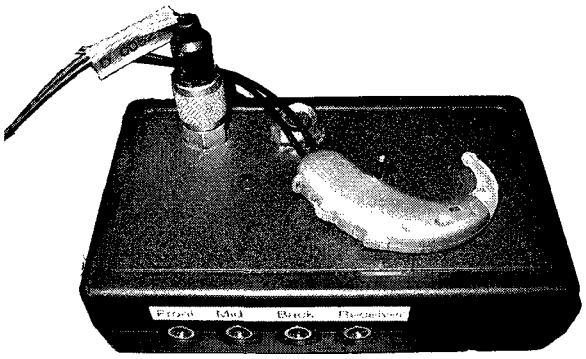
Chapter 3

System Setup and Evaluation

3.1 Speech and Noise Data

The speech and noise recordings were conducted by Siemens AG. Sentences from the TIMIT [67] database were played out over a CD player loudspeaker and binuaral recordings were captured from real hearing aids mounted on an acoustic research mannequin head. Fig. 3.1 depicts the left hearing aid connected to the data collection apparatus, the four stereo plugs on the front of the apparatus corresponding to the three binaural signals collected from the microphones and one binaural signal fed to the hearing aid loudspeakers. The recordings were sampled at 48 kHz, and contain six channels corresponding to the three microphones on each of the left and right hearing aids. The experiments that use single channel data use the recording from the first microphone on the left side; the binuaral experiments use the first right channel as well. All the speech files used in this work have been downsampled to the 16 kHz sampling rate of the original TIMIT files. The noise recordings contain real noise recorded in acoustic environments using the same apparatus as the speech recordings. The noises ranged from stationary white noise; to non-stationary pedestrian and traffic noise recorded in a street; to highly non-stationary, speech-shaped babble noise

Figure 3.1: Multi-channel hearing aid recording setup.



recorded in a university cafeteria. The speech and noise recordings were chosen to represent a broad range of common environments encountered by hearing aid users. The simulations are conducted in MATLAB, with the algorithms implemented using the MATLAB programming language; the parameters used in the reference algorithms are those given in the original sources unless otherwise stated.

3.2 Objective Evaluation Measures

The algorithms in this thesis are evaluated and compared to reference algorithms using objective performance measures. For the speech enhancement algorithms, the objective measures are selected to evaluate how the algorithms manage the trade-off between absolute noise reduction and speech intelligibility. For the feedback reduction algorithms, the measures are selected to evaluate how closely the adaptive filters model the true feedback path and, more importantly, how well they cancel the feedback signal without distorting the desired speech. In all cases informal listening tests are also used to ensure that the enhanced speech is natural sounding and free from glitches or disturbing artifacts. This is especially important for noise suppression algorithms, as musical noise artifacts are very disruptive to the listener, but do not

factor in to the objective tests.

3.2.1 Intelligibility Estimation

The coherence speech intelligibility index (CSII) [68] is used to evaluate speech intelligibility. The CSII is able to estimate intelligibility in the presence of non-linear distortion such as that imposed by a speech enhancement algorithm, and can therefore be used to evaluate the tradeoff between speech distortion and residual noise.

The speech intelligibility index (SII) [69] is an ANSI standard method to evaluate the intelligibility of speech in additive noise by estimating the amount of speech information that reaches a listener. The SNR is calculated in 1/3 octave spaced frequency bands, scaled to adjust for auditory threshold and masking effects, then weighted and summed to produce a numerical value between 0 and 1. An SII score of 1 indicates that all speech cues are available for the listener, while a score of 0 indicates that none are available. The CSII modifies the SII to account for non-linear speech distortion, specifically center and peak clipping. While the SII computes the SNR over the entire speech sample, the CSII divides the speech signal into overlapping Hamming-windowed-segments of 16-ms, to better account fluctuating noise conditions [70]. Rather than using the SNR, it estimates the signal to distortion ratio (SDR) using the MSC. For M frames of the clean signal spectrum $X(k)$ and the distorted signal spectrum $Y(k)$, the MSC is computed as:

$$|\eta(k)|^2 = \frac{\left|\sum_{m=0}^{M-1} X_m(k) Y_m^*(k)\right|^2}{\sum_{m=0}^{M-1} |X_m(k)|^2 \sum_{m=0}^{M-1} |Y_m(k)|^2}. \tag{3.1}$$

The MSC estimated speech power spectrum is $\hat{P}(k) = |\eta(k)|^2 S_{yy}(k)$ and the noise/distortion spectrum is $\hat{N}(k) = (1 - |\eta(k)|^2)\, S_{yy}(k)$ where $S_{yy}(k)$ is the is the estimated

PSD of the degraded signal. The SDR is then computed and weighted in the same way as the SNR in the SII computation. The frames are divided into low- mid- and high-level segments, and the weighted CSII values for the three types of frames are summed to yield a more accurate prediction of intelligibility scores.

3.2.2 Speech Quality Estimation

The ITU-T standard P.862.2 Wideband Perceptual Evaluation of Speech Quality (WPESQ) [71] is an objective measure designed to correlate with the subjective Mean Opinion Score (MOS). It analyzes the degradations imposed by a telecommunications network, such as distortion, noise and delay, and uses a perceptual model to estimate their what the impact would be on speech quality. The original PESQ was designed for operation in a telecommunications network, so the signals are first filtered to telephone bandwidth, 300 - 3400 Hz, WPESQ extends the model used in PESQ to include wideband frequencies, and operates at 16 kHz sampling rate. PESQ has been shown to offer good correlation with subjective evaluations of enhanced speech [72].

The log-likelihood ratio (LLR) measures the distance between the AR spectra of two signals as [27]:

$$d_{LLR}(\mathbf{a}_x, \mathbf{a}_y) = \log\left(\frac{\mathbf{a}_y^T \mathbf{R}_{xx} \mathbf{a}_y}{\mathbf{a}_x^T \mathbf{R}_{xx} \mathbf{a}_x}\right). \tag{3.2}$$

where $\mathbf{a}_x$ is the AR vector and R_{xx} is the toeplitz signal autocorrelation matrix. An AR(10) model is used to compute the LLR on 30 ms Hanning-windowed with 25% overlap. A smaller distance indicates a closer fit between the linear predictive coding (LPC) spectra, and therefore less distortion. The distance for each frame is limited to the range $[0, 2]$, then the overall distance is computed by taking the average distance over the lowest 95% of the frames to reduce the impact of outliers. A distance of 2 indicates severe distortion and a distance of 0 indicates no distortion. Compared

to the widely used segmental SNR, the LLR has been shown [72] to correlate much better with subjective overall speech quality and distortion. Since it measures the similarity of the AR spectra, the LLR is used to evaluate the parameter estimation performance of AR model-based algorithms.

3.2.3 Adaptive Filter Algorithm Evaluation

The two objective measures traditionally used to compared adaptive filter structures are output mean squared error (MSE) and system distance or misalignment. As the name implies, the MSE is the average squared difference between the output of the adaptive filter and the true output of the plant, and is generally expressed in dB:

$$MSE = 20\log_{10}(\mathbb{E}\{|d(n) - \hat{d}(n)|^2\}) \text{ dB} \tag{3.3}$$

In practice the expectation is computed by smoothing the computed difference.

System distance is the norm of the difference between the adaptive filter tap weight vector and the true plant impulse response. System distance provides an estimate of how deeply the system has converged by measuring the norm of the difference between the adaptive filter tap weight vector and the time-invariant Wiener solution. For a fullband filter, the system distance is given by:

$$\Delta[n] = 10\log_{10}\left(\frac{\|\mathbf{h} - \hat{\mathbf{h}}[n]\|^2}{\|\mathbf{h}\|^2}\right) \text{ dB} \tag{3.4}$$

Where $\mathbf{h}$ is the Wiener solution and $\hat{\mathbf{h}}[n]$ is the adaptive filter tap weight vector at time n. A subband adaptive filter does not converge exactly to the Wiener solution, rather a set of fully-converged subband filters, adapted by the error in each subband,

satisfies the condition [73]:

$$\sum_{m=0}^{M-1} \hat{\mathbf{h}}_m(z^D)\mathbf{H}_m(z) = \sum_{m=0}^{M-1} \mathbf{H}_m(z)\mathbf{h}(z).$$

where $H_m(z)$ is the analysis filter for the m^{th} subband. The equivalent subband system distance can therefore calculated as:

$$\tilde{\mathbf{h}} = \sum_{m=0}^{M-1} \mathbf{h} * \mathbf{H}_m \tag{3.5}$$

$$\tilde{\hat{\mathbf{h}}}[n] = \sum_{m=0}^{M-1} \mathbf{H}_m * \hat{\mathbf{h}}_m(n/D) \tag{3.6}$$

$$\Delta_{SB}[n] = 10\log_{10}\left(\frac{\|\tilde{\mathbf{h}} - \tilde{\hat{\mathbf{h}}}[n]\|^2}{\|\tilde{\mathbf{h}}\|^2}\right) \text{ dB} \tag{3.7}$$

In general system distance and MSE move in tandem: a small system distance indicates a small MSE and both metrics provide effective means for comparing the adaptation speed and tracking ability of different adaptive filtering algorithms. However, objective comparison of different adaptive filtering *structures* is complicated by fundamental differences in how the structures operate. System distance comparisons of fullband and subband adaptive filter structures are further complicated when an oversampled subband structure is used. Oversampled subband signals are narrowband, so each subband filter will quickly adapt to the high energy passband but experience very slow asymptotic convergence at the band-edges where there is little excitation [74]. In this case the MSE will rapidly decay close to a steady-state value but the asymptotic system distance convergence rate will be much slower. In addition, it is claimed in [75] that oversampled subband systems may treat the stopband regions of the subband signals as "don't care" regions, offering extra degrees of freedom to achieve a lower MSE but further degrading the system distance performance.

Chapter 4

Subband Acoustic Feedback Compensation

Hearing aid feedback suppression is typically performed on the fullband signal, before the subband decomposition block [15]. In this chapter we investigate the consequences of moving the feedback suppression block after the subband decomposition. Fig. 4.1 presents a block diagram of an adaptive subband feedback canceler. The signals are split into subbands where adaptive filtering is performed then the error signal is reconstructed using the synthesis filterbank.

Subband adaptive filtering has been used in acoustic echo cancellation to overcome problems of slow convergence for colored inputs and high complexity for large adaptive filters [76]. These issues are more compelling with wideband speech inputs due to the increased complexity at higher sampling rates, and the slow convergence caused by non-uniform frequency characteristics of wideband echo paths and speech. However, while acoustic echo paths can require thousands of taps to accurately model, acoustic hearing aid feedback paths are much shorter, raising the question of whether subbands structures are a viable option for feedback compensation. Also of interest is whether subband adaptive structures offer any other benefits besides complexity reduction or,

alternatively, if they introduce any new challenges.

In non-continuous adaptation (NCA) feedback compensation systems where, a probe signal $p[n]$ generated within the hearing aid is used to train the adaptive filter. In this type of controlled environment a white noise signal can be used to ensure that all modes of a fullband adaptive filter are excited, providing uniform convergence across all frequencies. Since the adaptation requires interruption of the regular hearing aid signal processing and the injection of a distracting probe signal, non-continuous adaptation systems require algorithms that can converge quickly the true feedback path, minimizing the interruption. In the ideal case, this adaptation would occur in quiet, when there is no incoming signal to disturb the adaptation, in practice the system must be robust to disturbing noise. In addition, the feedback path can change with time, so a good feedback compensation system requires an algorithm that not only offers fast initial convergence but good tracking capabilities as well.

In contrast to the NCA system, where the incoming signal is treated as disturbing noise, in a continuous adaptation (CA) feedback compensation system $p[n] = 0$ and the incoming signal must be used to adapt the feedback compensation filter. While the white noise probe signal of the NCA system offers controlled adaptation, the incoming signal used by the CA system is not controlled and is usually composed of speech; wideband speech energy varies across time and frequency leading to non-uniform convergence. Like NCA systems, CA systems must offer fast convergence and stable tracking of a changing feedback path; however unlike NCA systems, CA systems must also work to prevent the cancellation of incoming signal that results from its correlation with the feedback signal, as discussed in Chapter 2.

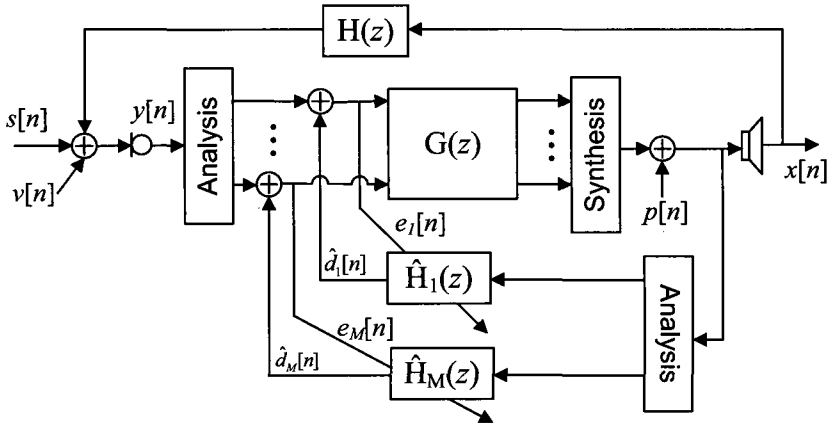


Figure 4.1: Subband adaptive feedback compensation structure with local error adaptation.

4.1 Complexity Analysis

The complexity reduction gained with subband adaptive filtering structures comes from the fact that the processing is carried out at a reduced rate, and each of the subband filters is much shorter than the fullband counterpart. For long adaptive filters these complexity savings outweigh the additional overhead required to divide the signals into subbands. In [76] it is shown that for a 16 band system with 256 tap polyphase analysis filterbank the overhead introduced by the two analysis and one synthesis filterbank operations is offset when the adaptive filter length reaches approximately 100 taps. In hearing aid feedback compensation the feedback paths are short, requiring dozens of taps rather than the thousands required for acoustic echo paths, so the ability to overcome the overhead and realize significant computational savings are questionable. However, as discussed in 2.1 subband processing is already used for noise reduction and multi-band amplification as a result subband analysis is already performed on the microphone signals and fullband synthesis on the enhanced signal, reducing the overhead of subband adaptive filtering to subband analysis of the

signal played out to the user. Furthermore, in [77] it is stated that with narrowband speech the audio quality of a critically sampled 4-band system was as good as fullband, but at lower complexity, although complexity figures are not given.

A common structure for filterbanks is the uniform modulated filterbank which splits the input into a set of subband signals each covering an equal fraction of the frequency spectrum. The Discrete Fourier Transform (DFT) kernel W_M^{-mn}, where $W_M = e^{j(2\pi m/M)}$, produces a bank of complex filters centered at $\omega_m = 2\pi m/M$, such that the $m = 0$ channel centered is at $\omega = 0$. The Generalized DFT (GDFT) extends DFT-modulated filterbanks to allow different band spacing, including placing the bin centers at $w_m = 2\pi m/M + \pi/M$. This is known as odd-channel stacking and it allows the real frequency range $\omega = [0, \pi]$ to be covered with a bank of $M/2$ evenly spaced filters, rather than the $M/2 + 1$ required by the even-stacked DFT. Both DFT and GDFT modulated filterbanks can be efficiently implemented by making use of the FFT. The analysis and synthesis filters used in this chapter are near perfect reconstruction GDFT modulated filters designed using the approach in [78]. The prototype filter has real valued coefficients and exhibits linear phase. In this work an $L_p = 64$ tap prototype filter with 45 dB stopband attenuation is used for an $M = 16$ band filterbank with a decimation factor of $D = 8$. The filters are linear phase, giving a total analysis-synthesis delay of 64 samples, which corresponds to 4 ms at 16 kHz sampling rate. In practice a higher stopband attenuation may be desired, requiring longer analysis and synthesis filters, in this case low-delay filterbank designs, such as those presented in [79], may be desirable.

For an M-band complex GDFT-modulated polyphase filterbank with a decimation factor of D and a length-L_p prototype filter, the number of real multiplications

required per fullband input sample is [80]:

$$C_f = \frac{1}{D}\left(4M\log_2 M + 6M + L_p\right) \tag{4.1}$$

For a real input, $M/2$ complex subband signals need to be processed. Standard NLMS requires $2N$ real multiplications to produce one fullband output sample. Assuming 4 real multiplies per complex multiply, subband NLMS with N/D taps per band requires $8N/D$ multiplies to produce a sample in each band at the decimated sampling rate. The total number of real multiplies required to update the adaptive filters for each fullband sample is $4\frac{NM}{D^2}$, and the total number of multiplies, including the added filtering step is:

$$C_{SB} = \frac{1}{D}\left(4M\log_2 M + 6M + L_p\right) + 4\frac{NM}{D^2} \tag{4.2}$$

For a fullband feedback path length of $N = 64$, an $M = 16$ band subband adaptive system with a decimation factor of $D = 8$ and an $L_p = 64$ tap prototype filter requires approximately 116 multiplies per sample, compared to 128 for the fullband implementation. However if, the decimation factor is increased to $D = 12$ and the prototype filter length is increased to $L_p = 128$ to keep the stopband attenuation at 45 dB, the number of multiplies for the subband configuration reduces to 69. With this computation it can be argued that if the filterbanks are implemented efficiently, the subband system can be at least as computationally efficient as the fullband, however the computational savings of a subband structure are not as compelling for the hearing aid feedback case as for echo cancellation.

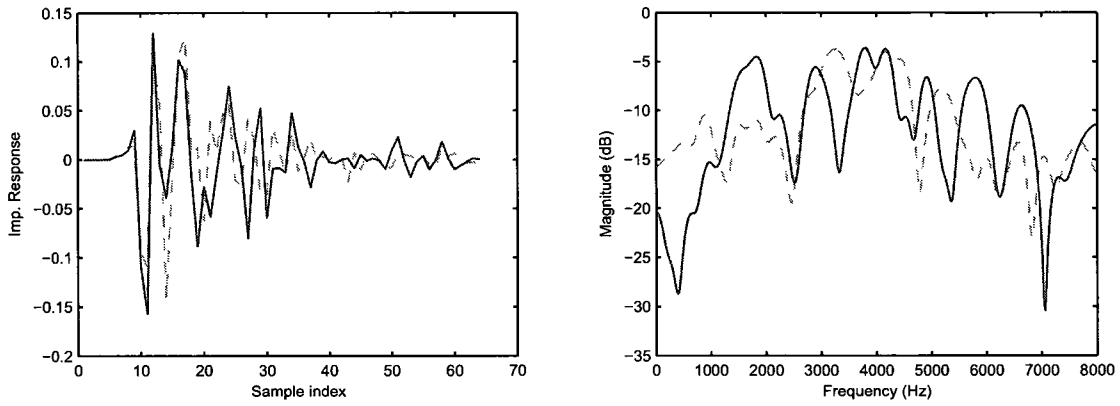


Figure 4.2: Impulse (left) and magnitude (right) response plots of two generated feedback paths.

4.1.1 Steady-State Performance Limitations

In order to compare fullband and subband structures, two acoustic feedback paths were generated with characteristics typical of real feedback paths: 64 active samples with an exponential decay in the time domain and a bandpass frequency domain characteristic between 1 – 4 kHz [6]. The impulse and magnitude frequency response plots of the two feedback paths are shown in Fig. 4.2.

The modeling performance of all adaptive filters is limited by background noise, undermodeling of the plant impulse response and non-linear distortion from sources such as the loudspeaker and microphone. In the case of subband adaptive filters there are additional limits posed by aliasing distortion and the need to model non-causal subband impulse responses that results from low-level aliasing noise [75]. According to [73] undermodeling of the plant and aliasing distortion caused by non-ideal analysis filterbanks are the greatest sources of excess MSE, and the MSE performance of the system is dominated by the larger of the two until the MSE is small enough that they both contribute. As discussed in [75], the need to model non-causal taps arises when a fullband echo path is approximated by a set of subband filters. Each

of the subband magnitude responses is a frequency windowed (bandpass filtered), and frequency expanded (decimated) version of the fullband magnitude response. In the time domain this frequency windowing corresponds to convolution with a modulated two-sided sinc(x) function, which results in non-causal subband impulse responses. The modified impulse response vector $\tilde{\mathbf{h}}$ in 3.5 is the causal representation of the fullband reconstruction of these non-causal subband impulse responses. Not modeling the non-causal taps can be seen as undermodeling the front of the impulse response rather than the tail, and the effect on excess MSE and system distance is the same. Some of the non-causal taps can be accounted for by delaying the desired signal relative to the reference signal, effectively increasing the flat delay of the plant. In systems where a flat delay already exists, the non-causal taps can be modeled adequately without added delay, so the overall effect on MSE may be minimal.

The different relationship between MSE and misalignment for fullband and subband systems can be seen in Fig. 4.3, which plots the average MSE over 20 trials for NCA adaptive feedback cancelers with white noise excitation and 50 dB and 25 dB measurement SNR. The subband and fullband filter lengths were set to enable modeling of the entire impulse response, and the step-size for both configurations was set to $\mu = 0.01$. The convergence differences between the two structures are evident at 50 dB SNR. While the fullband system experiences uniform convergence down to the noise floor, the subband system MSE has a rapid initial decay to around -30 dB, followed by a slower asymptotic convergence. The final MSE achieved by the subband structure is also limited, as the SNR effects mix with filterbank aliasing and non-causal tap limitations. The SNR due to in-band aliasing for the filterbank is 44 dB. In the more realistic 25 dB SNR case the subband system achieves the same rapid initial convergence, and both systems achieve the same steady-state MSE. Fig. 4.4 plots the system distance for the same two configurations. Despite the similar

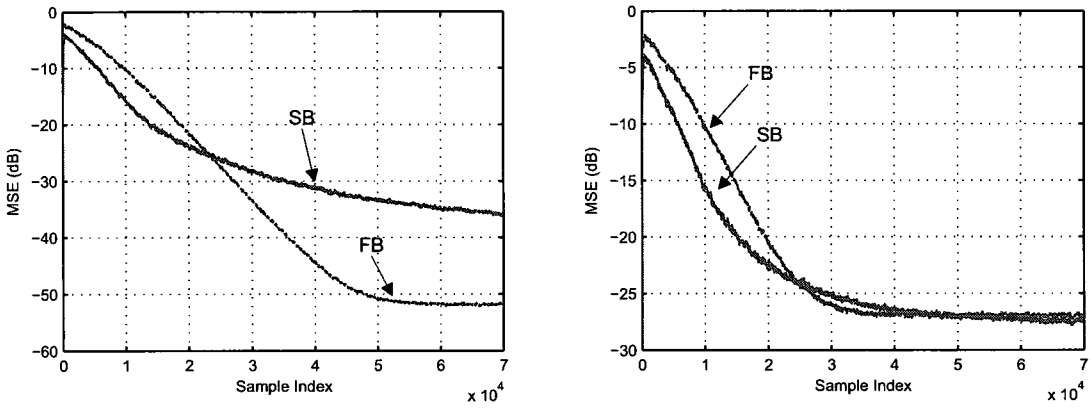


Figure 4.3: MSE plots for subband (solid) and fullband (dashed) non-continuous adaptation feedback cancelers with 50 dB (left) and 25 dB (right) SNR.

MSE performance in the 25 dB SNR case, the system distance differences between the two structures are significant, with the fullband system converging over 10 dB deeper than the subband. Furthermore, while the 50 dB SNR subband MSE is significantly lower than the 25 dB SNR case, the system distance curves reach the approximately the same level of -25 dB. These examples illustrate that under-excited, unconverged regions of each subband impulse response may contribute little to the fullband output MSE, but result in a significant excess equivalent fullband system distance. As a result system distance and MSE are not as correlated in the subband case as the fullband. For this reason MSE will be used more commonly than system distance to evaluate the performance of the feedback cancellation systems.

In summary, for a noise and distortion-free subband feedback canceler, undermodeling of both ends of the impulse response and aliasing distortion combine to limit the achievable MSE so that a fully-modeled fullband structure will always outperform an equivalent subband structure in the ideal case. In practice however background noise and signal distortion may restrict the MSE before these limitations become apparent.

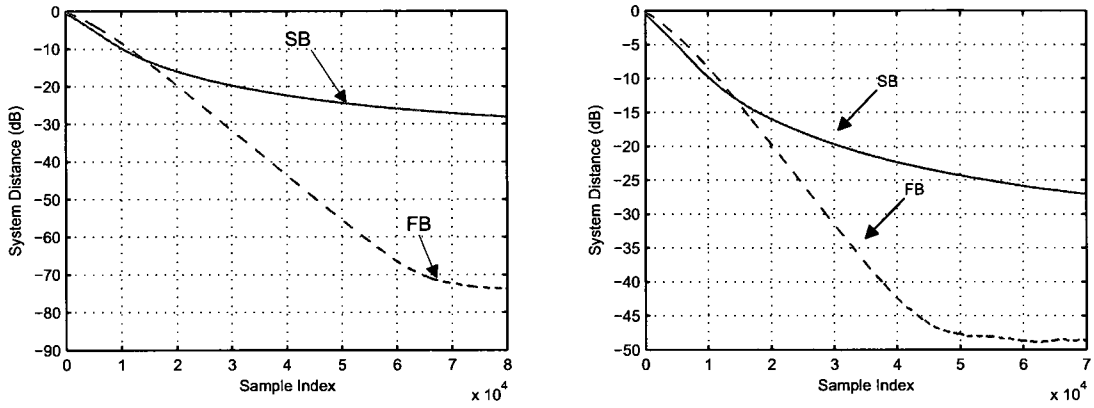


Figure 4.4: System distance plots for subband (solid) and fullband (dashed) non-continuous adaptation feedback cancelers with 50 dB (left) and 25 dB (right) SNR.

4.2 Simulation Results

A series of comparisons was carried out to investigate the performance of fullband and subband continuous and non-continuous feedback compensation structures under a variety of real-world conditions. A delay of 128 samples was added to the fullband hearing aid transfer functions to account for the 128 sample delay of the analysis and synthesis filtering of the subband structures. This delay decorrelates the input and reference signals for the CA systems, and does not have an impact on the NCA systems.

4.2.1 Changing Feedback Path Impulse Responses

Many comparisons of adaptive filtering algorithms focus on the speed of initial convergence, while less attention is given to tracking ability. Feedback paths can change with movements of the hearing aid caused by ear or jaw movements, or with changes in the arrangement of reflecting surfaces, such as hands or telephone receivers, near the outer ear [15]. Depending on the nature of the path change, some of the components

of the impulse response may remain the same despite the disruption. One source of feedback is mechanical coupling between the speaker and microphones caused by vibrations within the hearing aid. The location of the samples in the impulse response corresponding to these reflections will remain fixed regardless of external reflecting surfaces. Similarly, for a given hearing aid, the distance between the speaker and the microphones is fixed, consequently the delay associated with the direct path and the early reflections off of the inner ear will remain relatively constant. Since many of the reflecting surfaces in the ear and around the head are fixed, the general shape and decay rate of the impulse response tends to remains the same. Proportionate adaptation algorithms may be able to exploit this time-domain similarity to achieve faster tracking of changing impulse responses.

Just as feedback path changes leave some samples of the impulse response unchanged, the effect on the feedback path magnitude response can be limited to certain frequency regions. As seen in [15] different feedback paths have similar shapes in the frequency domain and frequency domains. Feedback paths typically have a bandpass characteristic, with the $1 - 4$ kHz regions experiencing the least attenuation [6]. A moving hand in the feedback path or a telephone receiver being placed in front of the ear may affect higher frequencies more than lower, because high frequencies are absorbed more easily and because the wavelength of the lower frequencies may be larger than the hand or obstructing object. For this type of feedback path change a subband adaptive filter might offer better tracking. While a fullband filter would have to adapt all tap weights in response to the disruption, a subband adaptive filter would only be required to adjust the weights of the filters in the affected bands. Even if the effects of the feedback path change are uniform across frequency, subband adaptive filters may have a tracking advantage as each of the subband adaptive filters is shorter than the fullband filter, so they should be faster to respond to and reconverge for all

types of path change.

To examine the impact of a changing feedback path on the feedback cancellation performance of adaptive filters under controlled conditions, fullband and subband versions of the NLMS and IPNLMS algorithms were compared using T samples of computer generated white Gaussian noise excitation with a simulated changing acoustic environment and 25 dB measurement SNR. After $T/4$ samples of initial convergence an abrupt feedback path change was simulated by changing the impulse response coefficients, then from $T/2$ to $3T/4$ samples a more realistic gradual change was simulated by linearly interpolating back to the first response, as in [81].

Non-Continuous adaptation

For the non-continuous adaptation structure the step-size for all algorithms was fixed at $\mu = 0.01$ in order to be able to observe the convergence rates, and $T = 1.5 \times 10^5$ samples were used for the simulation. The MSE performance, averaged over 20 trials, is shown in Fig. 4.5. The subband algorithms show the expected period of very rapid initial convergence, followed by a slower asymptotic convergence. However, during and after the feedback path changes, the subband algorithms offer lower worst-case MSE and faster reconvergence from the disturbance. It is interesting to note that while fullband IPNLMS offers faster initial convergence than subband NLMS, the reconvergence and tracking capabilities of subband NLMS are superior. The differences between the fullband and subband algorithms are most evident during the gradual path change: the worst case MSE of both subband algorithms is approximately 5 dB lower than the fullband algorithms. While fullband IPNLMS offers a significant initial convergence and tracking benefit, subband IPNLMS only offers a slight performance improvement over standard NLMS.

By exploiting the time and frequency domain characteristics of the feedback paths,

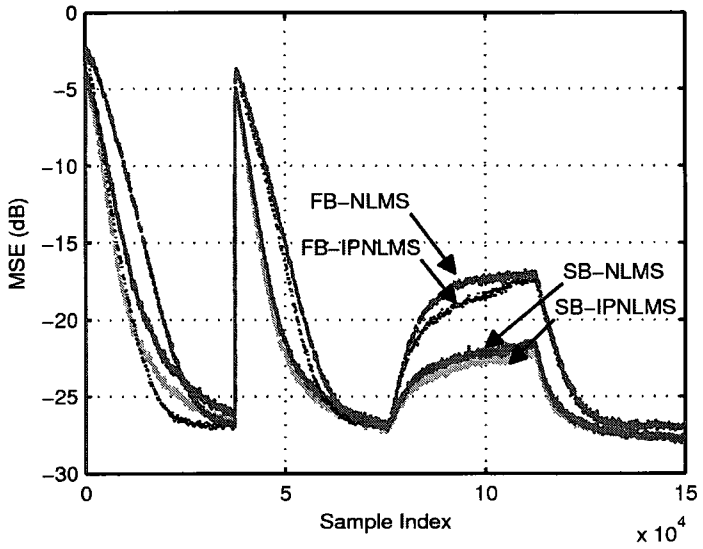


Figure 4.5: MSE performance of subband and fullband NLMS and IPNLMS non-continuous adaptation systems in synthetic changing environment.

the subband IPNLMS is able to offer the lowest consistent MSE in changing acoustic environments. However, unlike the fullband case the gain over standard subband NLMS is minimal and much of the tracking performance can be obtained with the lower complexity standard algorithm.

Continuous adaptation

With the NCA structure the processed microphone signal is replaced with the probe signal, so the feedback path is cut-off during adaptation; while in the CA structure the feedback path remains open, as the processed microphone signal is the signal used for adaptation. The initial adaptation and reconvergence periods of CA systems are therefore marked by periods of divergence while the growing feedback signal remains uncanceled. Also, while the MSE for the non-continuous case is upper bounded by the power of the probe signal, the MSE for the continuous adaptation structure is

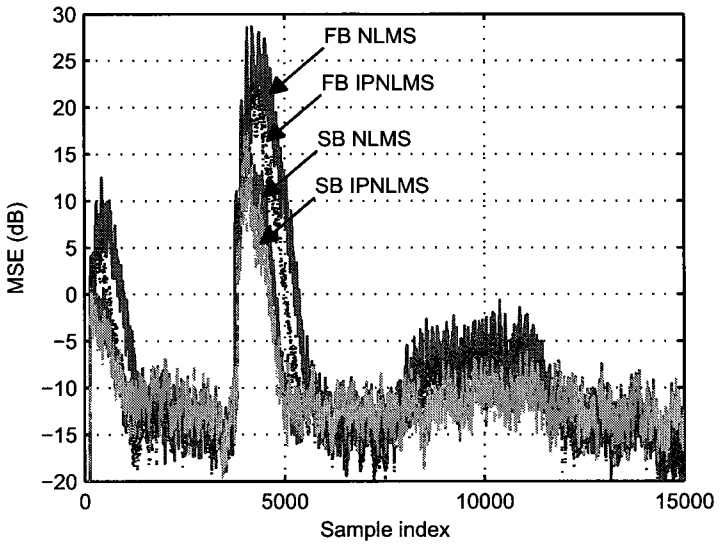


Figure 4.6: MSE performance of subband and fullband NLMS and IPNLMS continuous adaptation systems in synthetic changing environment with AWGN excitation.

unbounded. In order to prevent system instability during convergence, a larger step-size of $\mu = 0.125$ was used for the simulations, requiring fewer samples, $T = 1.5 \times 10^4$, to observe the response to changing paths.

Since the excitation signals are white and uncorrelated, the input signal cancelation that results from correlation is not present, and the results from the CA structures are similar to the NCA structures. Fig. 4.6 presents the MSE convergence curve for AWGN excitation, while Fig. 4.7 presents the spectrograms of the error signals. The MSE convergence curves demonstrate that the subband and fullband versions of both algorithms all diverge when the feedback path changes, although subband versions reconverge faster from the abrupt change and exhibit less divergence during the gradual change. Observing the spectrograms, we see that while the subband structure only diverges in isolated bands, corresponding to the highest peaks in the feedback path magnitude response, the fullband structure diverges across all frequencies.

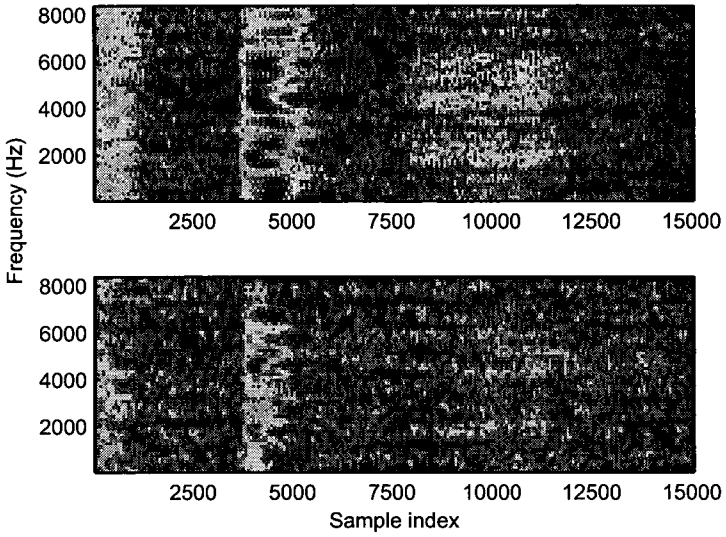


Figure 4.7: Spectrogram of error signal for fullband (top) and subband (bottom) continuous adaptation feedback compensation systems for changing feedback path with AWGN excitation.

4.2.2 Input Speech Signal

Non-continuous adaptation

Some NCA feedback cancellation systems generate and transmit the probe signal periodically during periods of talker silence, when there is no measurement noise signal to disturb the convergence [82]. The disadvantage to this approach is that the hearing aid processing is continually interrupted, even when the feedback path estimate is accurate, and instability can still occur if the feedback path changes significantly between updates. Another other option is to active the probe signal when divergence is detected, as in [25]. The challenge with this approach is that oscillation will not occur unless active speech is present, so the feedback canceler must adapt during what is known in echo cancellation literature as the "doubletalk" state.

Adaptation in doubletalk is problematic because the incoming speech signal acts as a high level disturbance causing the system to mis-converge. This problem is exacerbated in a NCA system where it is desirable to use a large step-size to minimize the interruption to the processing and offer fast re-convergence, as large step-sizes increase the rate of divergence. Mis-convergence in the presence of a high-level disturbance signal can be reduced by modifying the NLMS step-size normalization so that it is scaled by the sum of the signal and residual error powers [82]. For an N-tap filter the weight update step becomes:

$$\mathbf{w}[n] = \mathbf{w}[n-1] + \frac{\mu}{N(\sigma_x^2 + \sigma_e^2)} e[n]\mathbf{x}[n]. \tag{4.3}$$

While this approach slows initial convergence when the residual error is large, it also ensures that adaptation is slowed when the disturbing signal is large. The "error" signal at the output of the adaptive filter is composed to two parts: the uncanceled feedback signal and the measurement noise signal:

$$e_1[n] = d[n] - \mathbf{w}^H[n]\mathbf{x}[n] \tag{4.4}$$

$$e_2[n] = v[n] + s[n] \tag{4.5}$$

$$e[n] = e_1[n] + e_2[n] \tag{4.6}$$

When the disturbing signal is large $e_2[n]$ is large, and the normalization in (4.3) is increased, resulting in slower adaptation.

Fig. 4.8 plots the MSE curves for fullband and subband NLMS and IPNLMS in the presence of a disturbing speech signal at 0 dB and white measurement noise at 25 dB SNR. Since the disturbing speech signal is actually a desired signal, Fig. 4.8 only plots $e_1[n]$, the MSE due to the uncanceled feedback signal, while the filter

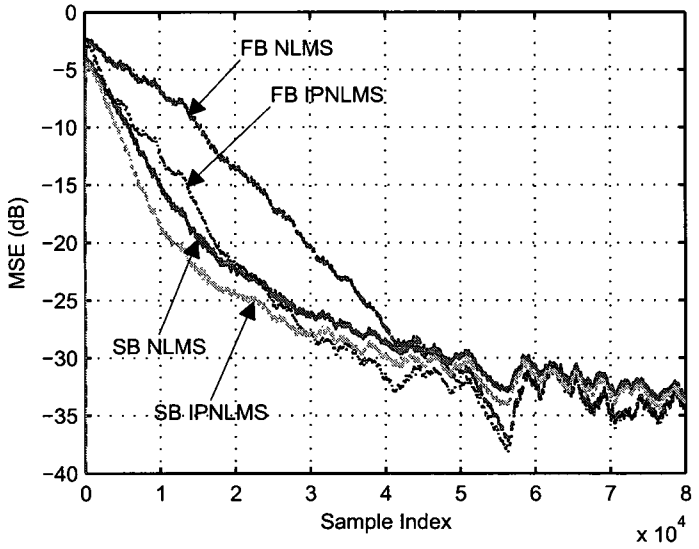


Figure 4.8: MSE convergence in the presence of disturbing speech.

adaptation was performed using $e[n]$, the full error signal. Despite the modified NLMS update in (4.3) and the small $\mu = 0.01$ step-size mild divergence can be seen during strong disturbing speech. Since the measurement noise $v[n]$ is not included in the MSE computation for Fig. 4.8, the steady-state performance limitations of the subband configuration manifest, resulting in a lower minimum MSE for the subband configuration. While the steady-state performance is lower, the initial convergence for the subband configurations is faster than both fullband configurations. The difference is especially evident with standard NLMS, which requires almost 1 second longer (16×10^3 more samples) than the other algorithms to achieve -20 dB MSE.

The changing acoustic environment experiment presented in Fig. 4.5 was repeated with the 0 dB disturbing speech signal, and the resulting MSE curves are presented in Fig. 4.9. When the only disturbing signal is white measurement noise, as in Fig. 4.5 the fullband IPNLMS offers faster initial convergence than subband NLMS and is just as fast as subband IPNLMS. However, when a disturbing speech signal is added, the

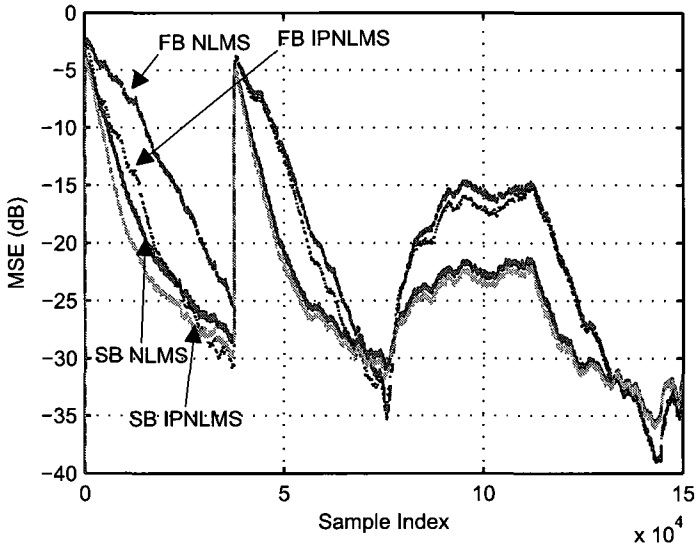


Figure 4.9: MSE convergence with disturbing speech in a changing acoustic environment.

initial convergence of both subband algorithms is faster than the subband IPNLMS. This may be due to the variable spectral content of the speech signal, which results in some of the subband filters having high SNR glimpses even when the fullband SNR is low, enabling those bands to converge more rapidly.

In addition to faster convergence, the subband configuration also offers more flexibility in dealing with disturbing speech. Since filter divergence occurs when the disturbing signal power is high, and the power of the speech signal decays with frequency, it is possible to use a smaller step-size in the low SNR bands with high incoming speech power, and a larger step-size in the higher SNR bands. Fig. 4.10 demonstrates the potential of this strategy. MSE curves are presented for the fullband algorithms with step-sizes of $\mu = 0.005$ and $\mu = 0.025$, and for the subband algorithms with logarithmically increasing step-sizes in the range $[0.005, 0.025]$. The fullband algorithms with the large step-size converge rapidly, but also exhibit MSE fluctuations of up to 15 dB as the interfering speech signal drives the divergence;

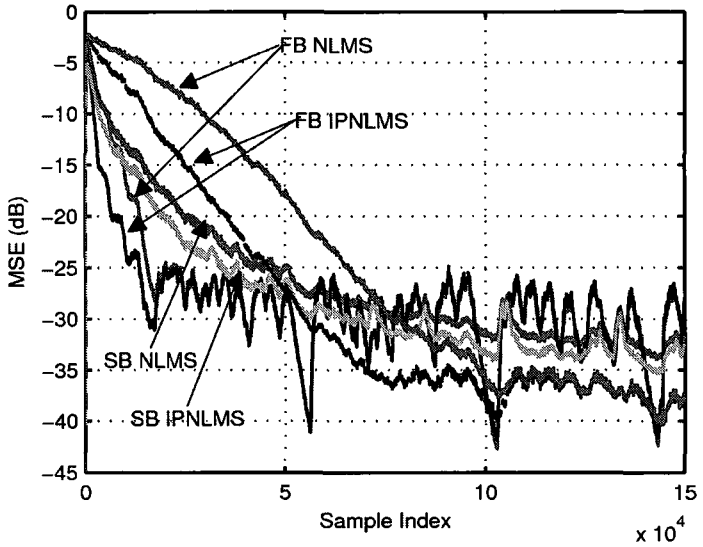


Figure 4.10: MSE convergence with disturbing speech, fullband step-sizes of $\mu = 0.005$ and 0.025, subband step-sizes logarithmically increasing with frequency between 0.005 and 0.025.

the fullband algorithms with small step-size exhibit greatly reduced MSE fluctuations, but the rate of convergence is much slower. The subband algorithms with the logarithmically-spaced step-sizes offer a compromise: their initial convergence rate approaches that of the large step-size fullband algorithms, but the divergence during strong speech segments is minimal. A slight divergence can be seen around sample 10×10^4, which is an area where the speech signal contains an onset with high frequency content, driving divergence in the high frequency bands with the larger step-sizes. However, in contrast to the fullband case, this divergence is localized to the bands where the disturbing signal is strong.

Continuous adaptation

While the results from section 4.2.1 demonstrate that the subband system can offer less divergence during a feedback path change when the input signal is uncorrelated

white noise, this situation is not likely to be encountered in practice as the primary input signal for a hearing aid is speech. Also, while the input speech signal disturbs the adaptation of NCA systems, in CA systems it forms the signal that must be used to adapt the feedback compensation filter. This use of the speech signal poses several challenges for a CA feedback compensation systems: its spectral coloration leads to slower and less-uniform convergence than a white input, and the auto-correlation of the speech signal leads to a cross-correlation between the desired and reference signals provided to the adaptive filter, causing system divergence. Adding a decorrelating delay to the hearing aid signal path can reduce the correlation for some speech inputs however voiced segments contain pitch-excited harmonic content with long-term correlation relationships. Since these harmonic signals are narrowband, the decorrelation of the subband transform, combined with the the ability to tailor the adaptation speed on a per-subband basis mean that a subband feedback compensation structure may offer benefits in a CA system with speech inputs.

In contrast to the NCA results and the white noise CA results, and despite the difficulties in comparing fullband and subband system distance outlined in section 4.1.1, system distance is used in place of MSE convergence to evaluate performance. The focus of the comparison will not be the steady-state system distance, but rather the changes in system distance. Rapid changes indicate divergence caused by adapting to the input signal rather than the feedback path. For hearing aid systems preservation of the incoming speech signal is critically important, and while a low MSE can be achieved while canceling the input speech signal, this type of divergence would be reflected in the system distance changes.

Figures 4.12 and 4.13 present the system distance convergence for fullband and subband CA feedback compensation systems using NLMS and IPNLMS adaptive

filters with speech inputs and white measurement noise at 25 dB SNR. As expected, in both cases the subband systems offer faster initial convergence, but higher steady-state system distance. The difference in convergence rates between NLMS and IPNLMS is notable for $\mu = 0.01$, but minimal for $\mu = 0.125$, especially in the subband configuration. For $\mu = 0.125$, all systems display rapid increases in system distance, indicating divergence as the filters adapt to cancel the input signal due to the cross-correlation between the input and the reference signal. The divergence periods correspond to the strongly voiced segments in the input that can be seen in the spectrogram of Fig. 4.11. The system distance fluctuations of the subband configurations are lower, as the divergence is limited to the bands where harmonic content is the strongest. With $\mu = 0.01$, the convergence in all cases is much slower but is more uniform and the systems do not exhibit divergence during the voiced speech segments; the only exception is the fast-tracking fullband IPNLMS, which exhibits mild divergence. As these trials indicate, the choice of step-size is a trade-off between large step-sizes which provide fast convergence, making the system robust to feedback oscillations, and small step-sizes that lead to slower convergence and less distortion of the input speech. Figure 4.14 demonstrates that the subband configurations can help mitigate this trade-off. The convergence curves for subband systems with step-sizes of $\mu = 0.125$ and $\mu = 0.01$ are repeated, along with the results from a configuration that tailors the step-sizes in each band to the signal content. A step-size of $\mu = 0.01$ was used in the two lowest frequency bands to ensure stable uniform convergence in the regions containing strong harmonic content, while a step-size of $\mu = 0.125$ was used in the three highest bands, as they contain mostly noise-like content with less correlation and can employ larger step-sizes for faster convergence. The remaining three middle bands contain a mix of lower power harmonic and noise-like content, and used a trade-off step-size of $\mu = 0.05$. The resulting convergence is much faster than

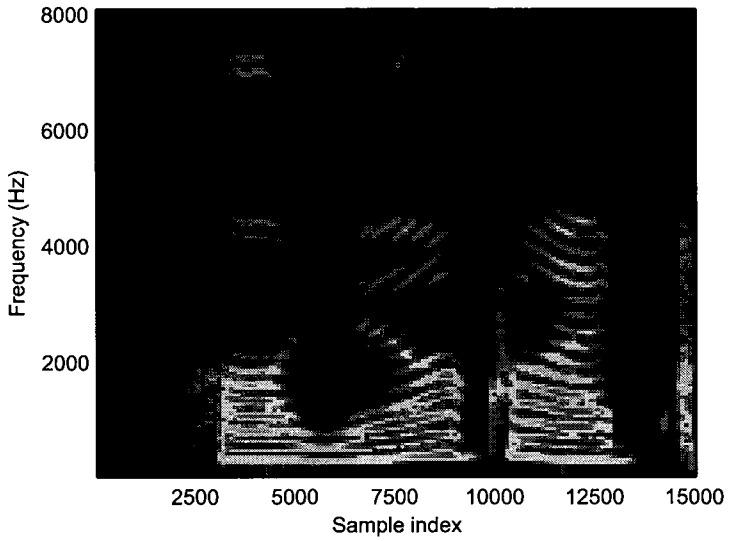


Figure 4.11: Spectrogram of speech segment used for feedback canceler trials.

the system using $\mu = 0.01$, but does not exhibit the divergence and signal distortion displayed with $\mu = 0.125$.

Figures 4.15 and 4.16 present the results of the changing acoustic environment simulations using speech instead of WGN input. The differences between the subband and fullband structures are even more pronounced in this case, attributed to the subband adaptive filter's superior convergence for colored inputs such as speech. The spectrograms indicate that, as with the NCA case, the divergence of the subband system is localized to the frequency bands where the feedback path attenuation is the lowest, while the fullband system diverges across all frequencies.

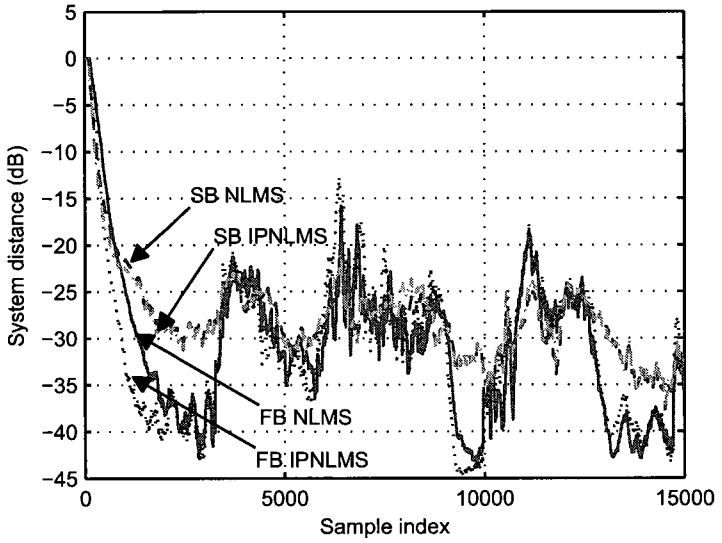


Figure 4.12: System distance convergence for fullband and subband continuous adaptation feedback cancelers with speech inputs, adaptation step-size $\mu = 0.125$.

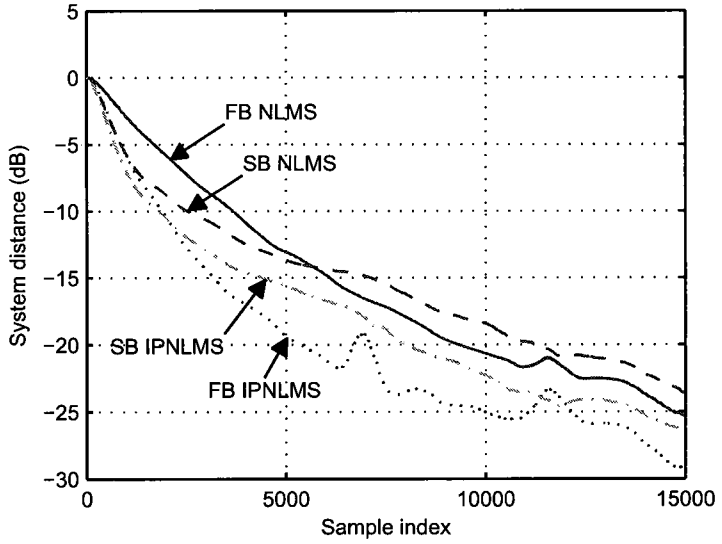


Figure 4.13: System distance convergence for fullband and subband continuous adaptation feedback cancelers with speech inputs, adaptation step-size $\mu = 0.01$.

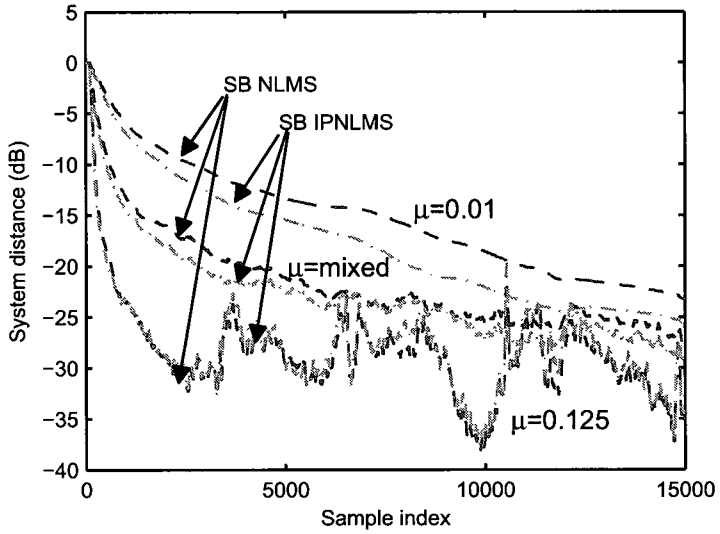


Figure 4.14: Subband system convergence for speech inputs per-band step-size assignment.

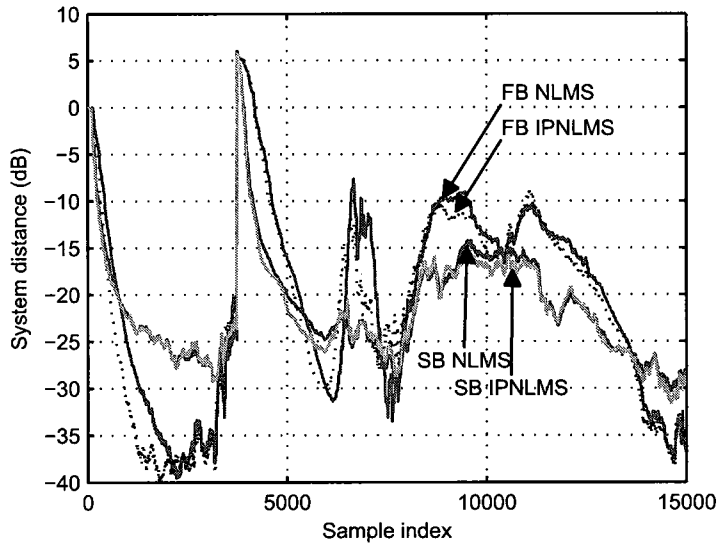


Figure 4.15: System distance performance of subband and fullband NLMS and IPNLMS continuous adaptation systems in synthetic changing environment with speech excitation.

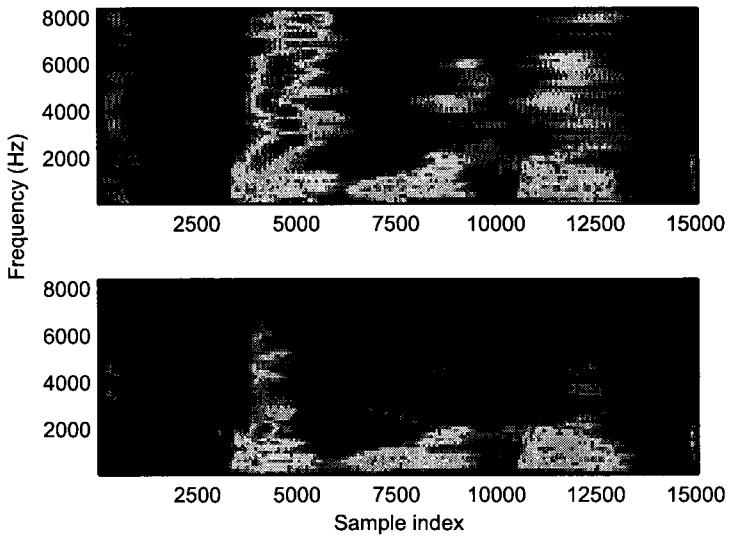


Figure 4.16: Spectrogram of error signal for fullband (top) and subband (bottom) continuous adaptation feedback compensation systems for changing feedback path with speech excitation.

4.3 Summary

In this chapter subband adaptive filtering structures were investigated for potential application in hearing aid acoustic feedback compensation. In contrast to the traditional subband adaptive filtering application, acoustic echo cancellation, where the impulse response being modeled is very long and subband structures are used to reduce the computational burden, hearing aid acoustic feedback paths are short, so the computational advantage is significantly reduced. Furthermore, under low measurement noise conditions the steady-state performance of subband adaptive filters is lower than the equivalent fullband system due to the effects of aliasing and imperfect filterbanks. However, other advantages of the subband configuration were revealed through simulation of fullband and subband continuous and non-continuous feedback compensation systems. It was demonstrated that, by only adapting the

filters in affected frequency bands, subband structures can offer faster reconvergence and tracking of changing acoustic feedback paths. In addition, the flexibility offered by treating different frequency regions separately allows subband systems to offer faster and more stable convergence when presented with wideband speech signals. System robustness was also improved, as divergence of the adaptive filter coefficients caused by feedback path changes or input signal correlation was confined to individual frequency bands. Under all conditions the fullband IPNLMS offered a significant initial convergence and tracking advantage compared to standard fullband NLMS. In the subband case IPNLMS maintained an advantage, although the benefit was significantly reduced, likely due to the very short subband adaptive filters.

Chapter 5

Particle Filter Enhancement of Speech Discrete Cosine Transform Coefficients

As discussed in Chapter 2 traditional speech enhancement methods such as spectral subtraction [11] and Bayesian spectral amplitude estimators, such as the minimum-mean-square error short-time spectral amplitude (MMSE-STSA) estimator, [12], use measurements of the transform coefficients of the noisy signal to obtain an estimate of the clean speech transform coefficients. While the transform is typically achieved with the short-time Fourier transform (STFT), other transforms can be used such as the discrete cosine transform (DCT) [32]. A main shortcoming of these approaches is that they do not take into account any correlation present in the speech signal: each frame is estimated independently of the others leading to musical noise when the noise spectrum is successively over- and under-estimated.

State-space model-based speech enhancement methods such as Kalman filter speech enhancement [13], model the clean signal waveform as an autoregressive (AR) process with Gaussian excitation, permitting the use of the Kalman filter equations to obtain the linear minimum mean-square error (MMSE) estimate of the clean speech waveform. The parametric model used by this approach provides a low variance

spectrum estimate [44], resulting in natural sounding residual noise that is free from musical noise. The main challenge in Kalman filter speech enhancement is estimating the clean signal parameters from the noisy measurement. One promising estimation method involves the use of sequential Monte-Carlo methods, particle filtering. Instead of estimating the speech parameters directly from the noisy signal, Rao-Blackwellized Particle Filter (RBPF) speech enhancement algorithms [65] [60] assume a dynamic model for the parameter evolution and the enhanced speech signal is taken as the ensemble average of the output of several realizations of the parameter process.

In this chapter we propose using the RBPF algorithm to enhance the speech DCT coefficients rather than the time domain speech waveform. By applying dynamic models to the frequency transform coefficients the proposed system attempts to offer frequency-dependent processing of wideband speech while preserving the naturalness of the output speech and residual noise. An empirical upper bound for the performance of the DCT and fullband domain particle filters is determined using a Kalman filter provided with the speech and noise parameters measured from the original signals. A series of experiments is then carried out to evaluate the performance of fullband and DCT-domain particle filters under realistic conditions when the true parameters are not supplied. The DCT-RBPF is compared, using perceptually motivated objective measures and informal listening tests, against a set of reference algorithms including the fullband particle filter. A hybrid DCT-domain RBPF/Wiener filter is also proposed, wherein a simple Wiener filter is substituted for the RBPF in the perceptually less relevant bins. The DCT-RBPF is found to offer improved noise suppression and a significant complexity reduction compared to the fullband RBPF. While the hybrid structure offers further computational savings, there is a reduction in the resulting enhanced speech quality.

5.1 Rao-Blackwellized Particle Filter Speech Enhancement

As introduced in Chapter 2, in the RBPF speech enhancement algorithms of [65] [60] the speech signal at time n, x_n is assumed to be corrupted by additive measurement noise v_n giving the measurement signal z_n:

$$z_n = x_n + v_n \tag{5.1}$$

The clean speech signal is assumed to follow a p^{th} order time-varying autoregressive (TVAR) model:

$$x_n = \sum_{i=1}^{p} a_n(i) x_{n-i} + d_n \tag{5.2}$$

$$= \mathbf{a}_n^T \mathbf{x}_n + d_n \tag{5.3}$$

where:

$$\mathbf{x}_n = [x_{n-1} \dots x_{n-p}]^T \tag{5.4}$$

is the signal vector,

$$\mathbf{a}_n = [a_n(1) \dots a_n(p)]^T \tag{5.5}$$

is the vector of auto-regression coefficients at time n, and d_n is the excitation sequence. Note that this differs from the standard autoregressive model assumed by the standard Kalman filter algorithm of [13] where the AR parameters are assumed fixed over the measurement interval and are only updated periodically.

Defining the state-transition and measurement matrices $\mathbf{F}_n$ and $\mathbf{G}$, (5.2) can be

described by the following linear state-space model:

$$\mathbf{F}_n = \begin{bmatrix} & \mathbf{a}_n^T & \\ \mathbf{I}_{p-1} & & \mathbf{0}_{p-1\times 1} \end{bmatrix} \tag{5.6}$$

$$\mathbf{G} = \begin{bmatrix} 1 \\ \\ \mathbf{0}_{p-1\times 1} \end{bmatrix} \tag{5.7}$$

$$\mathbf{x}_n = \mathbf{F}_n \mathbf{x}_{n-1} + \begin{bmatrix} d_n \\ \\ \mathbf{0}_{p-1\times 1} \end{bmatrix} \tag{5.8}$$

$$z_n = \mathbf{G}\mathbf{x}_n + v_n \tag{5.9}$$

where (5.8) and (5.9) are the state-space transition and measurement equations respectively.

If all of the TVAR parameters are known, and $d_n \sim \mathcal{N}(0, \sigma_d^2)$ and $v_n \sim \mathcal{N}(0, \sigma_v^2)$ are uncorrelated zero-mean white Gaussian noise processes, then an MMSE optimal estimate of x_n can be obtained using a Kalman filter. In the RBPF speech enhancement algorithm of [65], the TVAR parameters themselves are assumed to evolve stochastically. A constrained Gaussian random walk model is used for the AR coefficient vector, where the roots of the selected vector are forced to lie inside the unit circle (a sufficient but not necessary condition for stability of the TVAR model):

$$p(\mathbf{a}_n|\mathbf{a}_{n-1}) \propto \begin{cases} \mathcal{N}(\mathbf{a}_n|\mathbf{a}_{n-1}, \Delta_a^2\mathbf{I}) & \text{If stable} \\ 0 & \text{Otherwise,} \end{cases} \tag{5.10}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ denotes a Gaussian PDF with mean μ and (co-)variance σ^2 evaluated at x. The speech and noise variances are assumed to evolve according to a random walk in the log-domain, to ensure that the variances remain positive:

$$p(\log(\sigma_{d_n}^2)|\log(\sigma_{d_{n-1}}^2)) = \mathcal{N}(\log(\sigma_{d_n}^2), \delta_d^2) \tag{5.11}$$

$$p(\log(\sigma_{v_n}^2)|\log(\sigma_{v_{n-1}}^2)) = \mathcal{N}(\log(\sigma_{v_n}^2), \delta_v^2) \tag{5.12}$$

The hyper-parameters $(\Delta_a^2, \delta_d^2, \delta_v^2)$ determine the variances of the random walks taken by the TVAR parameters.

Defining the vector of TVAR parameters:

$$\theta_{\mathbf{n}} = [\mathbf{a}_n^T,\ \sigma_{d_n}^2,\ \sigma_{v_n}^2]^T \tag{5.13}$$

we have an augmented state-vector $\tilde{\mathbf{x}}_n = [\mathbf{x}_n^T,\ \theta_n^T]^T$. The transition equations describing the evolution of this augmented state vector are stochastic and not linear, however, conditioned on θ_n, $\mathbf{x}_n$ evolves according to a linear Gaussian state space model as described by (5.8) and (5.9), and the MMSE optimal estimate of $\mathbf{x}_n$ can be obtained using the Kalman filter equations. The augmented state vector can therefore be partitioned and Rao-Blackwellization can be applied.

An RBPF is therefore two interacting algorithms: at each time step the SIR particle filter generates N candidate TVAR parameter vectors $\{\theta_{\mathbf{n}}{}^{(i)}\}_{i=1}^N$. These parameters are supplied to a bank of Kalman filters, which perform the signal enhancement and update the error covariance matrix which is used to compute the conditional likelihood of the measurement signal given the parameter set. The candidate TVAR vectors and their corresponding Kalman mean and covariance are then re-sampled for propagation to the next iteration. The RBPF approach circumvents the problems

associated with estimating the clean speech model from the noisy signal by simulating and evaluating the ensemble of model parameters. In addition, with the RBPF approach the parameters are not estimated on a frame-by-frame basis and can instead be allowed to vary continuously. This may allow the RBPF approach to better model the time-variations of the human vocal-tract filter.

The RBPF speech enhancement algorithm is summarized in Algorithm 1. Systematic re-sampling is carried out at each sampling interval to prevent particle degeneracy.

5.2 DCT-domain RBPF

RBPF speech enhancement has been shown to offer effective low-distortion noise reduction with natural sounding residual noise [66], however it is very computationally intensive as a Kalman filter iteration must be performed for each candidate TVAR vector at each time-step. Furthermore, while the inter-speech noise attenuation is very strong, fullband RBPF enhanced speech contains relatively high intra-speech residual noise. In this section we show that the intra-speech noise can be attributed to insufficient attenuation in spectral troughs that arises when modeling a wideband speech signal with a moderate order all-pole model. To address the complexity and noise issues we proposed applying the RBPF algorithm to the speech DCT coefficients, rather than directly to the speech waveform. Modeling the evolution of the DCT coefficients using low-order AR models produces a very high-order equivalent fullband model at significantly reduced computational cost. Subband Kalman filtering has been shown to offer much lower complexity and higher levels of noise attenuation than fullband Kalman filtering without introducing un-natural artifacts [43, 44]. As RBPF speech enhancement can be seen as an extension of Kalman filtering, the benefits of subband processing are expected to extend in the particle filter case. This

Algorithm 1 Rao-Blackwellized Particle Filtering [65].

1: **for** $n \geq 0$ **do**
2: **for** $i = 1, \ldots, N$ **do**
3: Draw candidate TVAR parameters $\theta_n^{(i)} \sim p(\theta_n | \theta_{n-1}^{(i)})$.
4: Obtain $\mathbf{F}_n^{(i)}, \sigma_{d_n}^{2,(i)}, \sigma_{v_n}^{2,(i)}$ from $\theta_n^{(i)}$.
5: Use the TVAR parameters to update the Kalman statistics:

$$\mathbf{x}_{n|n-1}^{(i)} = \mathbf{F}_n^{(i)} \mathbf{x}_n^{(i)}$$
$$\mathbf{P}_{n|n-1}^{(i)} = \mathbf{F}_n^{(i)} \mathbf{P}_{n-1|n-1}^{(i)} \mathbf{F}_n^{T,(i)} + \sigma_{d_n}^{2,(i)} \mathbf{I}$$
$$\mathbf{T}_n^{(i)} = \mathbf{G} \mathbf{P}_{n|n-1}^{(i)} \mathbf{G}^T + \sigma_{v_n}^{2,(i)} \mathbf{I}$$
$$\mathbf{K} = \mathbf{P}_{n|n-1} \mathbf{G}^T \mathbf{T}_n^{-1,(i)}$$
$$\mathbf{x}_{n|n}^{(i)} = \mathbf{x}_n^{(i)} + \mathbf{K}(z_n - \mathbf{G}\mathbf{x}_{n|n-1}^{(i)})$$
$$\mathbf{P}_{n|n}^{(i)} = (\mathbf{I} - \mathbf{K}\mathbf{G}) \mathbf{P}_{n|n-1}^{(i)}$$

6: Compute the un-normalized importance weights:

$$\tilde{W}_n^{(i)} = \mathcal{N}(z_n | \mathbf{G}\mathbf{x}_{n|n-1}^{(i)}, \mathbf{T}_n^{(i)})$$

7: **end for**
8: Normalize the importance weights.

$$\{W_n\}_1^N \leftarrow \frac{\{\tilde{W}_n\}_1^N}{\sum_{i=1}^N \tilde{W}_n^{(i)}}$$

9: Compute Bayesian posterior estimate.

$$\hat{x}_n \leftarrow \sum_{i=1}^N W_n^{(i)} x_n^{(i)}$$

10: Resample: $\{\mathbf{x}^{(i)}, \mathbf{P}^{(i)}, \theta^{(i)}\}$
11: **end for**

is supported by the observations in [83] that indicate RBPF speech enhancement applied in a non-uniform filterbank configuration can achieve lower segmental SNR scores than an unspecified fullband configuration.

5.2.1 Subband AR Modeling

The RBPF algorithm models speech as a TVAR process, with the TVAR parameters estimated using a particle filter. As the model order increases, the size of the state-space modeled by the particles increases, requiring more particles to achieve good performance [84]. In order to manage complexity and performance, a low-order model is therefore desirable. However, very high order pole-only models are required to model the pitch structure of speech [44]. Furthermore, while low-order AR modeling of speech produces a good fit in spectral peaks, it can over-estimate the signal power in spectral troughs. In the context of Kalman filtering speech enhancement, this over-estimation leads to intra-speech residual noise as the noise between spectral peaks is not sufficiently attenuated. If the residual noise is sufficiently far from a formant peak, it will not be masked by the formant and will be perceptually noticeable. This problem is especially prominent for wideband speech where the spectral dynamic range within a frame is higher, and there can be high and low frequency energy in the same frame.

An alternative to using a single high-order AR process is to perform the enhancement in a transform domain, using one low-order model per transform bin. It has been shown [85], that with an ideal filterbank subband linear prediction can achieve lower prediction error variance than fullband order-p linear prediction; this was demonstrated to hold when each band uses a prediction order p, and even when the sum of the prediction orders across all bands is p.

To compare the performance of low-order fullband and subband linear pole-only

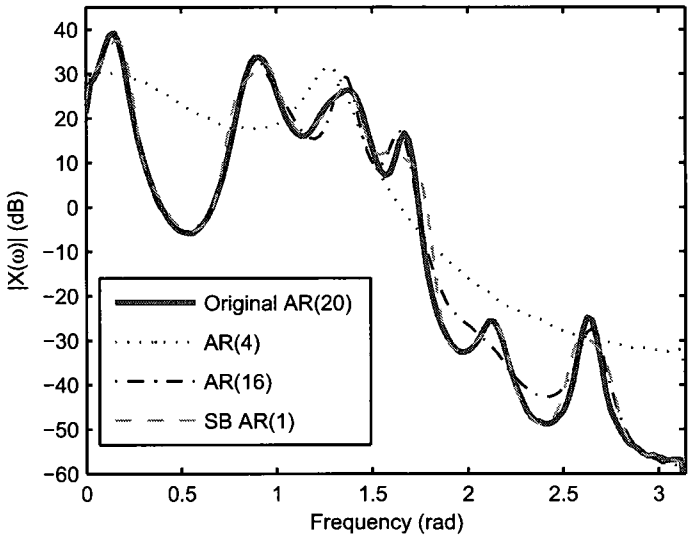


Figure 5.1: Spectra obtained from AR modeling of an AR(20) process with a fullband AR(4) model, a fullband AR(16) model and a 16 subband configuration with an AR(1) model in each band.

modeling of complex processes, an AR(20) process was generated by passing white noise through a filter measured from a segment of voiced speech, and an ARMA process was generated by filtering white noise through a system with zeros at $\pm 0.25\pi$ and poles at $\pm 0.1\pi$ and $\pm 0.8\pi$. Fullband AR(4) and AR(16) models were fitted to the signals, and an AR(1) model was fitted to each band of a 16-band cosine modulated filterbank. The spectral estimates are shown in Fig. 5.1 for the AR(20) process and Fig. 5.2 for the ARMA process. The results from the AR(4) estimates demonstrate that when low-order fullband AR modeling is used to estimate a higher order process, the smooth fitting of a curve between the poles causes the signal energy between the poles to be significantly over-estimated. While the AR(16) model is better than the AR(4) model at attenuating the spectral troughs, the subband estimate still provides the closest fit.

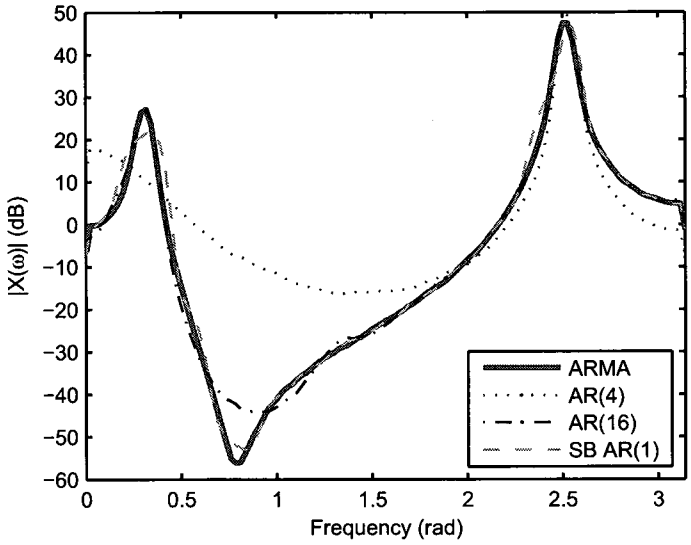


Figure 5.2: Spectra obtained from AR modeling of an ARMA process with a fullband AR(4) model, a fullband AR(16) model and a 16 subband configuration with an AR(1) model in each band.

5.2.2 Algorithm Description

We propose to use the DCT for the subband decomposition, as it is a real frequency transform that can be efficiently implemented using an FFT-like algorithm. For an M-point signal x_n, the standard forward and inverse DCT equations are:

$$X_n(m) = \frac{2c(m)}{M} \sum_{n=0}^{M-1} x_n \cos\left[\frac{m\pi}{2M}(2n+1)\right], \qquad m = 0, 1, \ldots, M-1 \tag{5.14}$$

$$x_n = \sum_{m=0}^{M-1} c(m) X(m) \cos\left[\frac{m\pi}{2M}(2n+1)\right], \qquad n = 0, 1, \ldots M-1 \tag{5.15}$$

where

$$c(m) = \begin{cases} 1/\sqrt{2} & m = 0 \\ 1 & m = 1, 2, \ldots, M-1 \end{cases} \tag{5.16}$$

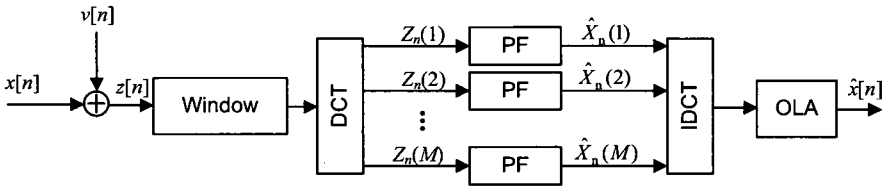


Figure 5.3: Block diagram of DCT domain particle filtering speech enhancement.

The DCT is well suited to speech processing as it has been shown to provide good decorrelation and energy compaction of speech, while DCT transformed noise is approximately white [32]. Fig. 5.3 presents a block diagram of the proposed DCT-based particle filtering system. The speech signal is divided into overlapping frames, windowed, then transformed to the DCT domain. An RBPF using a low-order TVAR speech and a white noise model is used to model each DCT coefficient. Since the white noise power can differ in each DCT bin, the proposed system can be used to enhance signals corrupted by colored as well as white noise. The enhanced signal is transformed back to the time domain using the inverse DCT (IDCT), then reconstructed by overlapping and adding successive frames (OLA).

Algorithm 2 summarizes the DCT-RBPF algorithm using an M-bin DCT with a frame overlap of $M/2$. The notation $\mathbf{0}_{1\times M/2}$ indicates a row vector of $M/2$ zeros.

Reduced Complexity Hybrid Particle-Wiener Filter

In addition to the performance and complexity advantages, the DCT configuration enables different processing approaches to be applied to different coefficients. The number of particles, the order of the speech and noise AR models, the structure of the speech and noise models and even the algorithm employed can all be made to be frequency dependent. To demonstrate the potential of this type of processing, we propose combining the DCT-RBPF with the DCT Wiener filter of [32]; the RBPF is

Algorithm 2 DCT-RBPF.

1: **for** $n = lM/2,\ l \in \mathbb{Z}$ **do**
2: Frame noisy signal and multiply with windowing function $\mathbf{w}$:

$$\tilde{\mathbf{z}}_n = [z_n, z_{n-1}, \ldots, z_{n-M+1}]^T$$
$$\mathbf{z}_n = \mathbf{w}^T \tilde{\mathbf{z}}_n$$

3: Compute forward DCT of $\mathbf{z}_n$ to obtain $\mathbf{Z}_n$, using (5.14).
4: **for** $m = 1, \ldots, M$ **do**
5: Perform RBPF computations in Algorithm 1 using $Z_n(m)$ as input to obtain estimate of clean DCT coefficient $Y_n(m)$.
6: **end for**
7: Compute inverse DCT using (5.15) to obtain estimate of clean signal frame y_n.
8: Overlap and add to obtain clean signal estimate:

$$\hat{\mathbf{x}}_{n+M/2} = \left[\hat{\mathbf{x}}_{n-M/2}^T \mathbf{0}_{1\times M/2}\right]^T + \mathbf{y}_n$$

9: **end for**

used to model the perceptually important DCT coefficients, while the simpler Wiener filter is used to compensate the remainder..

The Wiener filter estimate of the k^{th} DCT coefficient at time n, $X_n(k)$ is given as:

$$\hat{X}_n(k) = H_n(k) Y_n(k) \tag{5.17}$$
$$= \frac{\xi_n(k)}{1+\xi_n(k)} Y_n(k). \tag{5.18}$$

where:

$$\xi_n(k) = \frac{\mathbb{E}\{|X_n(k)|^2\}}{\mathbb{E}\{|V_n(k)|^2\}}. \tag{5.19}$$

is the *a-priori* SNR. As this value cannot be evaluated directly without knowledge of the clean signal, it is commonly estimated using the decision-directed approach

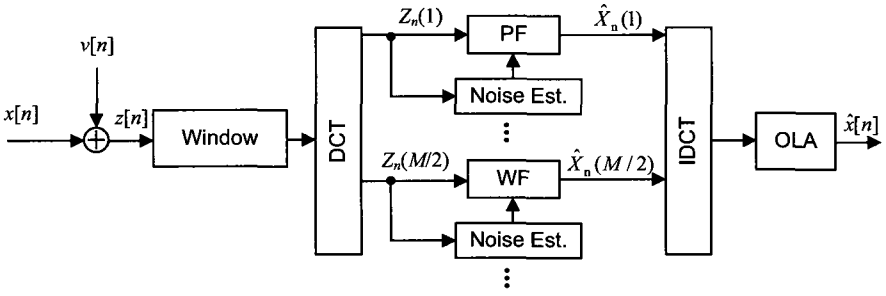


Figure 5.4: Block diagram of hybrid DCT-domain particle and Wiener filter. PF denotes a particle filter, WF denotes a Wiener filter.

of [12]:

$$\xi_n(k) \approx \alpha \frac{|\hat{X}_{n-1}(k)|^2}{\mathbb{E}\{|V_n(k)|^2\}} + (1-\alpha)\max\left[\frac{|Y_n(k)|^2}{\mathbb{E}\{|V_n(k)|^2\}} - 1, 0\right]. \tag{5.20}$$

where α is a smoothing parameter $0 \leq \alpha \leq 1$, and the max function ensures the SNR remains non-negative.

Figure 5.4 presents a block diagram of of the proposed configuration. For wide-band (16 kHz sampling rate) speech input the lower half of the DCT coefficients correspond to telephone bandwidth speech and are the most perceptually important. These coefficients are processed using the higher performing but computationally intensive particle filter (denoted PF), while the higher frequency bins are processed with a Wiener filter according to (5.18). The RBPF and Wiener filter algorithms are compatible as both employ external noise estimation algorithms to obtain an estimate of the noise power in each DCT bin. While some versions of the RBPF can estimate the noise on-line, this is not recommended for stability reasons, as will be discussed in section 5.3.2.

5.2.3 Complexity

As with standard Kalman filtering, the complexity reduction benefits of transform domain processing are numerous in the particle filtering case. While the DCT configuration requires multiple particle filters, each filter is operating at a reduced rate (time-decimated) and each DCT coefficient series is modeled using a low-order AR process.

For an AR speech model and a white noise model, a standard RBPF requires: particle generation; Kalman filter iteration; and weight computation steps. The weight computation is $\mathcal{O}(1)$, and is the same regardless of the AR vector size. In contrast, the particle generation and Kalman filter update steps are dependent on the AR model order p. The particle generation requires generating a random walk disturbance for each element of the AR vector, this step is conservatively estimated to be $\mathcal{O}(p)$. However since stable AR vectors are generated by sampling until a stable draw is achieved, and the stability of the entire AR vector is affected by every coefficient, the number of vectors that are rejected as unstable will increase as the AR model order grows. The Kalman filter speech enhancement complexity is $\mathcal{O}(p)$ if fast Kalman filter techniques are applied [13]. The overall complexity of an N-particle fullband RBPF is approximately $\mathcal{O}(pN)$ computations per unit time.

An M-bin DCT-domain RBPF with 50% overlap between frames requires M RBPFs operating at the decimated rate of $M/2$ times the fullband rate. Additional computations are required for the forward and inverse transforms. For a q^{th} order AR speech model and a white noise model for each DCT coefficient, the complexity of an N-particle M-bin DCT-RBPF is approximately $\frac{M}{M/2}(\mathcal{O}(qN)+\mathcal{O}(M\log(M))) = \mathcal{O}(2qN)+\mathcal{O}(M\log(M)))$ computations per unit time. For the hybrid configuration the complexity of the Wiener filter operation is $\mathcal{O}(1)$, and the computational savings are dependent on the number of DCT coefficients processed using each algorithm.

Fig. 5.5 compares the execution time required to enhance 2 seconds of speech sampled at 16 kHz for a fullband RBPF using an AR(10) model to a DCT-RBPF using an AR(2) model in each DCT bin. An $M = 128$ bin DCT was used with a 50% overlap. Both structures used the same MATLAB C-Mex file to perform the particle filtering, and the timing for the DCT-RBPF includes the DCT overhead. For the hybrid configuration the upper half of the DCT coefficients were processed with the Wiener filter. Even if the same number of particles are used for fullband and DCT structures, the DCT overhead is quickly overcome by the complexity reduction achieved by using smaller AR models. Furthermore, the smaller model order means that fewer particles can be used in each DCT bin than the fullband, as lower dimension state spaces require fewer particles [84]. The execution time of the hybrid configuration is approximately half that of the DCT-RBPF owing to the computational simplicity of the Wiener filter. It should also be noted that the complexity of both the fullband and DCT-RBPF structures could be further reduced by using the modified RBPF [86], which only requires computing one Kalman gain for all particles.

5.3 Evaluation

Two sets of experiments are used to evaluate the performance of the proposed DCT-RBPF. In the first set of trials the best-case performance of the DCT-RBPF is compared to that of the standard fullband RBPF by providing both algorithms with TVAR parameters measured directly from the unmixed speech and noise signals. In second set of trials the RBPF algorithms operating without the true TVAR parameters are compared to reference algorithms from the various speech enhancement families.

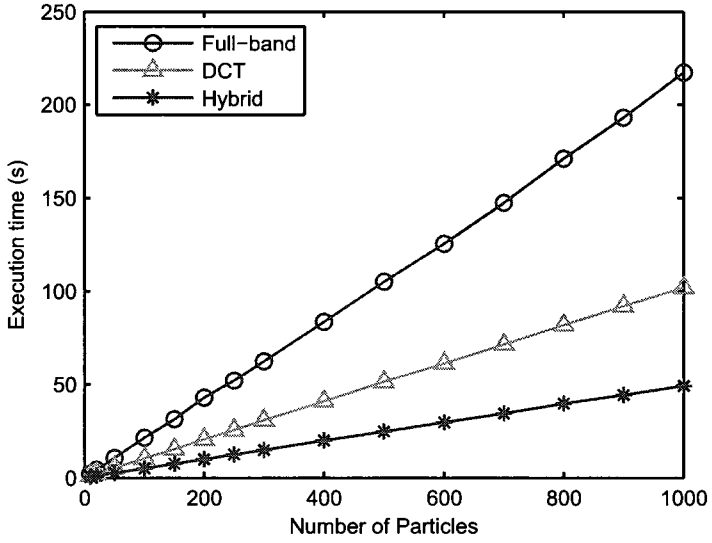


Figure 5.5: Execution time comparison for fullband (FB) and DCT-RBPFs processing two seconds of speech sampled at 16 kHz.

5.3.1 Simulation Parameters

The following parameters are used for the speech enhancement algorithms in the comparison:

1. Fullband RBPF [65] using an AR(10) speech model and white Gaussian noise model. The AR vector and speech and noise variances evolve according to a Gaussian random walk. The augmented state vector colored noise RBPF was also considered but did not offer a significant advantage over the white noise model in the tested conditions.

2. DCT domain Wiener filter (DCT-WF) [32], modified to use decision-directed SNR estimation [12] with $\alpha = 0.98$. A 128-point Von Hann window and a 50% overlap were used to frame the speech signal.

3. MMSE-LSA [33] using decision-directed SNR estimation with $\alpha = 0.98$ and a

512-point Von Hann window with 50% overlap for framing..

4. Signal subspace approach [40][1] using sine-taper covariance estimation and 32 ms Hamming windowed frames.

5. Proposed DCT-RBPF using an AR(2) speech model, and a white Gaussian noise model in each frequency bin. DCT parameters the same as the reference Wiener filter.

6. Proposed hybrid DCT-RBPF/DCT-Wiener filter. Particle filter with the same parameters as the DCT-RBPF operating in the lower 64 bins and a Wiener filter with the same parameters as (2) operating in the upper 64 bins.

The particle filters used 100 particles per AR vector element, giving 1000 particles for the fullband and 200 particles per filter for the DCT configuration.

The structures were evaluated under the following set of speech and noise test scenarios.

1. Dry speech with synthetic additive white Gaussian noise (AWGN) .

2. Dry speech with street noise. The street noise is mostly from passing traffic and was recorded at an intersection with the listener facing away from the street.

3. Speech recorded in a cafeteria, with babble recorded during a meal at the same university cafeteria. The speech-shaped power distribution and time-varying nature of babble noise make it a challenging noise to reduce.

The speech recording was sampled at 16 kHz and contained four sentences by different speakers from the TIMIT database, two spoken by males and two by females, with a total duration of 14 seconds. The noise conditions and environments

[1]Implementation modified from [27].

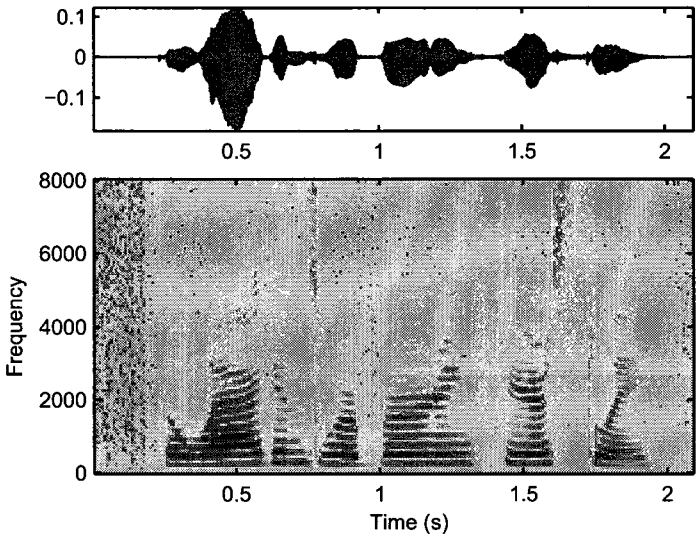


Figure 5.6: Waveform (top) and spectrogram (bottom) of segment of clean speech.

were chosen to provide a broad range of realistic test situations. Each speech and noise file was simulated at overall SNRs of 5, 10, 15 and 20 dB, giving a total of 16 test conditions. For the particle filter algorithms, each combination of algorithm and speech-file was run 20 times, and the performance results were averaged. For consistency all of the algorithms use an ideal voice-activity detector (VAD) to update the noise estimate during speech pauses, although on-line estimation approaches could also be employed. Time and time-frequency domain plots of the clean signal and the noisy signals at 10 dB SNR are presented in Figs. 5.6 – 5.9.

5.3.2 DCT/Fullband Comparison

Empirical Performance Bounds

RBPF speech enhancement is tightly linked to Kalman filter enhancement, the output of an RBPF is the weighted sum of the outputs of a bank of Kalman filters. The performance of RBPFs is therefore upper-bounded by a Kalman filter using the true

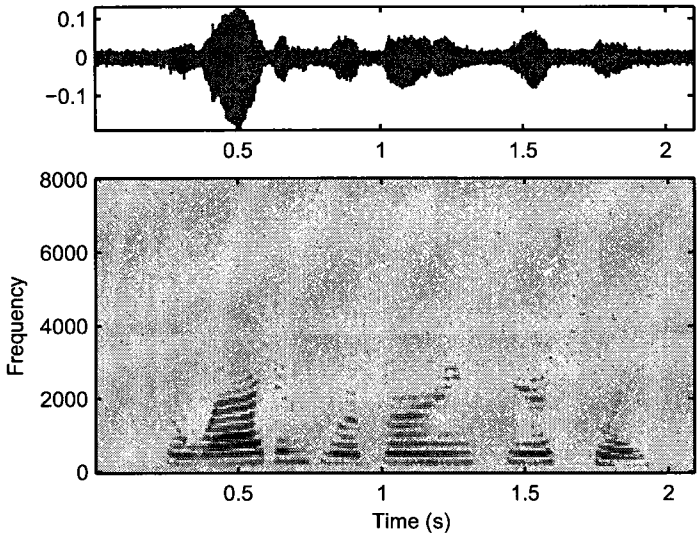


Figure 5.7: Waveform (top) and spectrogram (bottom) of segment of speech corrupted by AWGN at 10 dB SNR.

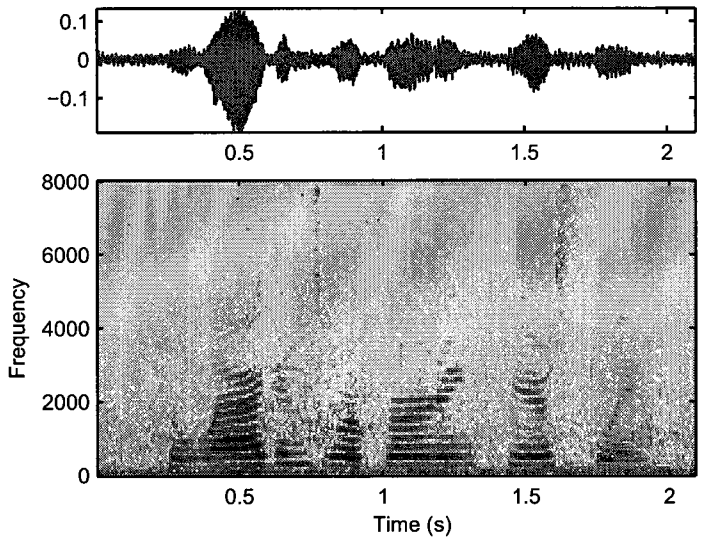


Figure 5.8: Waveform (top) and spectrogram (bottom) of segment of speech corrupted by street noise at 10 dB SNR.

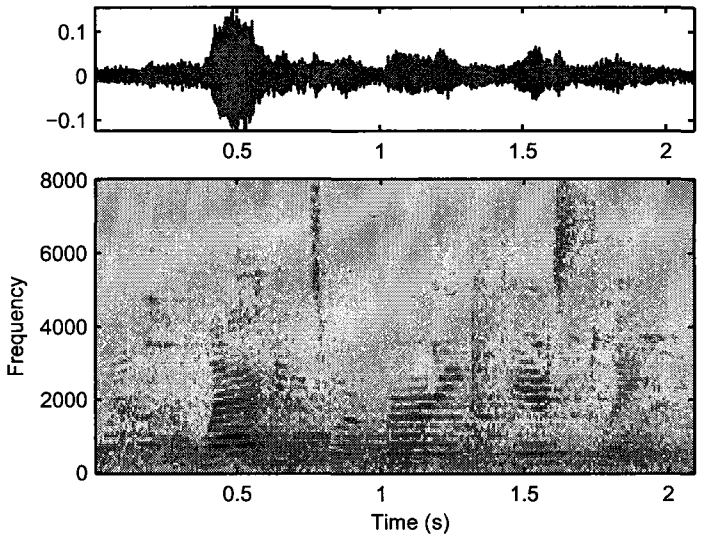


Figure 5.9: Waveform (top) and spectrogram (bottom) of segment of speech corrupted by cafeteria babble at 10 dB SNR.

parameters. The empirical upper-bounds for the fullband and DCT RBPFs were investigated by supplying a Kalman filter algorithm with measured AR parameters and noise variances at every sample iteration. A window size of 8 samples was used to estimate the AR parameters of each DCT component, while 64 samples were used in the fullband case. The analysis window length was chosen by experimentation and represents a trade-off between capturing the time-varying characteristics, especially of the excitation signal, and ensuring the signal was stationary over the analysis frame. This trade-off prevents us from obtaining the true ideal parameters, but they serve to provide a reasonable upper-bound on the performance of the particle filter algorithms. The noise variance at time n was the true squared noise signal in either the time or DCT domains. Then, to estimate the upper-bound on fullband and DCT-RBPF parameter estimation, the true noise excitation was supplied while the speech TVAR parameters were estimated by the particle filters.

The upper-bound results from all noise conditions are presented in tables 5.1 –

5.3. In all cases the performance of the DCT-domain algorithms is superior to that of the fullband structures, achieving higher levels of noise reduction without imposing additional distortion. This is even true in the AWGN case, where the fullband noise model assumptions are met. Fig. 5.10 and Fig. 5.11 compare waveforms and spectrograms of fullband and DCT domain ideal Kalman filtering of the 10 dB SNR AWGN corrupted speech segments. Relatively high levels of intra-speech residual noise are clearly visible in the spectrogram of the fullband enhanced signal, while in the DCT enhanced signal the residual noise is confined to smaller time-frequency blocks. Along with the high levels of noise reduction, the DCT Kalman filter is still able to preserve most of the low-energy speech components. Also note that for both the fullband and DCT cases there is little degradation from the ideal Kalman filter to the measured noise RBPF, indicating that the particle filters are effective at tracking the speech TVAR parameters. As would be expected the deviation from the measured parameter results is greatest in low-SNR conditions.

Table 5.1: Performance results for AWGN, measured noise parameters.

	WPESQ Scores			
Algorithm	5 dB	10 dB	15 dB	20 dB
Noisy	1.06	1.14	1.35	1.78
Ideal FB	1.36	1.56	1.86	2.26
True Noise FB	1.17	1.35	1.67	2.13
Ideal DCT	2.49	2.84	3.22	3.61
True Noise DCT	2.11	2.50	2.92	3.34
	CSII Scores			
Noisy	0.44	0.72	0.88	0.95
Ideal FB	0.76	0.88	0.95	0.97
True Noise FB	0.79	0.90	0.95	0.98
Ideal DCT	0.82	0.92	0.96	0.98
True Noise DCT	0.76	0.88	0.95	0.98
	LLR Distance			
Noisy	1.69	1.59	1.43	1.20
Ideal FB	1.04	0.95	0.83	0.68
True Noise FB	1.42	1.28	1.10	0.89
Ideal DCT	0.42	0.36	0.30	0.26
True Noise DCT	0.60	0.46	0.35	0.26

Table 5.2: Performance results for street noise, measured noise parameters.

	WPESQ Scores			
Algorithm	5 dB	10 dB	15 dB	20 dB
Noisy	1.18	1.45	1.96	2.51
Ideal FB	1.45	1.74	2.12	2.57
True Noise FB	1.24	1.55	2.05	2.62
Ideal DCT	2.54	2.90	3.23	3.61
True Noise DCT	2.29	2.61	2.95	3.33
	CSII Scores			
Noisy	0.60	0.84	0.94	0.97
Ideal FB	0.68	0.85	0.94	0.97
True Noise FB	0.65	0.85	0.95	0.98
Ideal DCT	0.90	0.96	0.98	0.98
True Noise DCT	0.86	0.94	0.97	0.98
	LLR Distance			
Noisy	0.98	0.75	0.55	0.40
Ideal FB	0.74	0.58	0.44	0.32
True Noise FB	0.90	0.66	0.46	0.29
Ideal DCT	0.27	0.19	0.12	0.08
True Noise DCT	0.37	0.25	0.17	0.11

Table 5.3: Performance results for cafeteria babble noise, measured noise parameters.

	WPESQ Scores			
Algorithm	5 dB	10 dB	15 dB	20 dB
Noisy	1.20	1.57	2.19	2.99
Ideal FB	1.44	1.85	2.46	3.21
True Noise FB	1.30	1.69	2.32	3.09
Ideal DCT	2.67	3.12	3.58	3.96
True Noise DCT	2.28	2.78	3.34	3.87
	CSII Scores			
Noisy	0.38	0.67	0.87	0.96
Ideal FB	0.61	0.79	0.91	0.97
True Noise FB	0.56	0.77	0.91	0.97
Ideal DCT	0.80	0.91	0.97	0.98
True Noise DCT	0.73	0.87	0.95	0.98
	LLR Distance			
Noisy	0.49	0.33	0.20	0.11
Ideal FB	0.41	0.28	0.18	0.10
True Noise FB	0.56	0.38	0.23	0.13
Ideal DCT	0.11	0.08	0.05	0.03
True Noise DCT	0.11	0.08	0.05	0.03

5.3.3 Comparative Study

In order to evaluate performance in realistic situations, the particle filters were not supplied with the true speech and noise parameters for the comparative study. The particle filter framework enables the noise to be estimated within the algorithm along with the speech parameters, or externally using traditional noise estimation techniques. Three noise estimation approaches were considered. In the first configuration the noise power was updated by recursive averaging during during silence intervals. The silence intervals were manually labeled to simulate a perfect VAD. In the second configuration the particle filters were provided with an initial estimate of the measurement noise power, then the log noise power was allowed to evolve according to a random walk, as in [60]. This approach can be used to estimate both stationary and time-varying noise powers, with the variance of the random walk modeling the variability of the noise power. Note that even if the noise is assumed stationary, a non-zero variance should be used for the random walk to preserve particle diversity. In this case the variance of the log noise power was a fixed hyper-parameter set to $\delta_v = 1 \times 10^{-4}$. The third configuration was a combination of the first two: the noise power was updated from the measurement during silence intervals, and was tracked by the particle filter during the speech utterances.

In the DCT-domain there was little distinguishing the three configurations while there was more of a difference in the fullband case. For the babble noise input, the structure using particle filter noise estimation out-performed the external noise estimation configurations, indicating that the random walk noise model is a good fit for the babble talk, being able to capture the noise power fluctuations. In contrast when the measurement noise was constant AWGN the external noise estimation gave better results than using the particle filter method with a random walk noise model. Reducing the variance of the noise power random walk improved the performance for

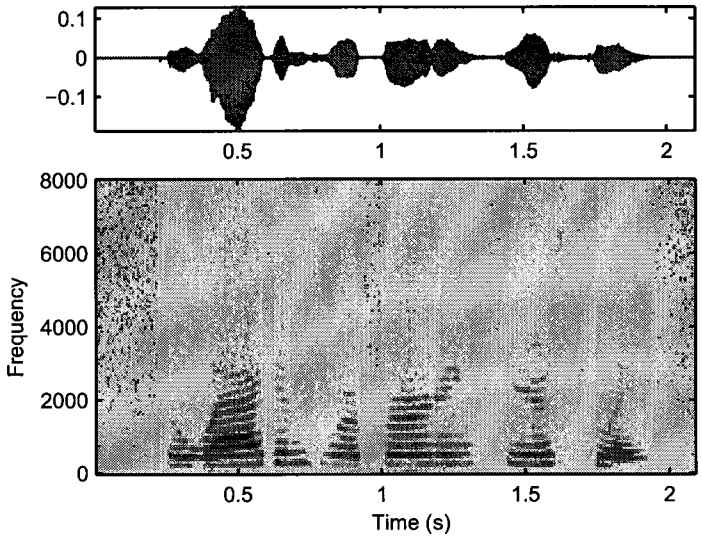


Figure 5.10: Waveform (top) and spectrogram (bottom) of 10 dB SNR AWGN corrupted speech enhanced by fullband ideal Kalman filter, showing intra-speech residual noise.

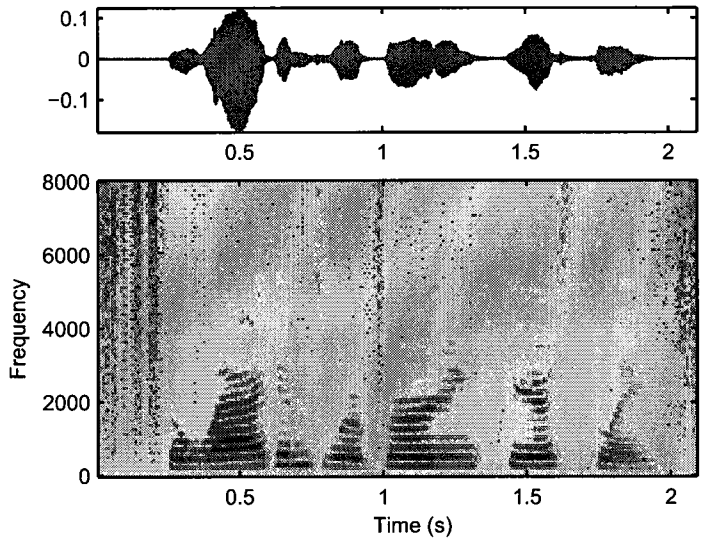


Figure 5.11: Waveform (top) and spectrogram (bottom) of 10 dB SNR AWGN corrupted speech enhanced by DCT-domain ideal Kalman filter.

AWGN at the expense of the babble noise case. This implies that the performance of the RBPF algorithms could be further improved with better noise estimation techniques or a better model of noise evolution.

Despite its strong performance for the changing noise conditions, it was observed that the particle-filter-only noise estimation with a random walk on the noise power should not be used unsupervised in on-line speech enhancement, as mis-matches between the assumed model and the actual noise power evolution can lead to complete removal of the speech signal. If the speech excitation and noise power elements of the particle filter are drawn without any external guidance, particles from the random walk model can be generated with a small speech excitation power and a noise power equal to the power of the noisy speech signal. These particles provide a good model fit and will receive high importance weights. As a result the speech and noise signal will be treated as noise-only, and the entire noisy speech signal will be canceled. These particles will be resampled and propagated, and if the variance of the noise power random walk is comparable to the variance of the speech and noise power, the particle filter can continue to track and cancel the noisy speech signal. The problem is most acute when a small number of particles is used, although the same phenomenon can also occur with more particles are used but the SNR is low and the noise is more stationary than is assumed by the model. This phenomenon was observed in simulations with the fullband RBPF with 150 particles and a random walk model for the noise power evolution attempting to remove stationary AWGN measurement noise. Reducing the variance of the noise power random walk reduced the amount of speech cancellation, but it also reduced the performance in changing noise conditions. In order to avoid this phenomenon some form of external noise estimation should be employed, at least periodically, to keep the particle filter on track. For this reason the combined particle filter and external noise estimation approach is used for the

particle filter results in the comparative study.

The results of the comparative study are presented in tables 5.4 – 5.6. Observing the results we find that the difference between the best-case and actual performance tends to be greater at low SNRs and for the cafeteria babble. The DCT-domain particle filter experiences a larger degradation in moving from true to estimated noise parameters. Despite this fact it still far outperforms the much more complex fullband RBPF, for all three performance measures. In general the performance of the hybrid DCT Wiener/RBPF configuration is somewhere between the two "parent" algorithms, and the residual noise remained natural sounding for the hybrid configuration. The strong bin decorrelation properties of the DCT enables the bins to be processed independently without causing distortion, so the mixture of the two different algorithms does not result in any additional artifacts.

Of the reference algorithms the MMSE-LSA and signal subspace approach were the highest performing. In general the MMSE-LSA has higher WPESQ scores than the signal subspace approach, but lower CSII scores. This is consistent with the subjective tests in [27] that showed MMSE-LSA has better sound "quality", while signal subspace offers less signal distortion and higher intelligibility. The DCT-RBPF on the other hand balances the trade-off between these two objectives, matching the quality of MMSE-LSA and the intelligibility scores of signal subspace, offering a modest but consistent improvement in objective scores over all conditions. It is notable that the DCT-RBPF has the lowest LLR distances of all algorithms, and the fullband RBPF has the second lowest distances for most noise conditions, demonstrating the ability of the RBPF algorithm to model the time-varying speech parameters. It should also be noted that the objective performance differences among the algorithms were consistent when the male and female sentences were evaluated separately or grouped together.

The objective results are also supported by informal listening tests. At low SNR conditions the fullband RBPF has relatively high levels of intra-speech residual noise while the DCT-RBPF does not. This is in agreement with the empirical upper-bound results which showed that intra-speech residual noise is present for fullband Kalman filter-based algorithms even under ideal conditions. The residual noise for the fullband RBPF maintains the character of the original noise signal, with its power modulated by the envelope of the speech signal. For the DCT-RBPF this modulation occurs with the envelope of each DCT coefficient, so the time-frequency character of the residual noise is modified. However, unlike the signal subspace approach, the DCT-RBPF enhanced speech does not exhibit musical noise artifacts. The MMSE-LSA offers the most natural sounding residual noise, but it also imposes the most speech smoothing.

5.4 Summary

In this chapter a speech enhancement algorithm using RBPF modeling of speech DCT coefficients was proposed and evaluated. Processing the speech signal at a decimated frame rate, and employing low-order TVAR models with white measurement noise significantly reduces the complexity of the proposed DCT-RBPF algorithm compared to the standard fullband RBPF. In experiments using measured speech parameters, the DCT-RBPF was shown to be more capable of modeling the wideband speech spectrum, offering superior best-case noise reduction. Performing the enhancement in a transform domain also enables frequency-dependent processing strategies. To demonstrate this approach a hybrid DCT-domain structure was proposed that uses the RBPF in the perceptually important lower bins, and a simple Wiener filter in the upper bins, providing a performance/complexity trade-off. An interesting extension to this work would be to incorporate additional psychoacoustic information. As in [87],

Table 5.4: Performance results for AWGN.

	WPESQ Scores			
Algorithm	5 dB	10 dB	15 dB	20 dB
Noisy	1.06	1.14	1.35	1.78
MMSE-LSA	1.55	1.95	2.34	2.70
Subspace	1.56	1.92	2.47	3.10
FB-RBPF	1.09	1.20	1.46	1.92
DCT-WF	1.43	1.60	1.86	2.26
DCT-RBPF	1.63	1.99	2.47	2.97
Hybrid DCT	1.65	1.97	2.29	2.60
	CSII Scores			
Noisy	0.44	0.72	0.88	0.95
MMSE-LSA	0.59	0.76	0.87	0.93
Subspace	0.68	0.85	0.94	0.97
FB-RBPF	0.47	0.75	0.90	0.96
DCT-WF	0.39	0.60	0.77	0.89
DCT-RBPF	0.63	0.83	0.93	0.97
Hybrid DCT	0.56	0.77	0.88	0.93
	LLR Distance			
Noisy	1.69	1.59	1.43	1.20
MMSE-LSA	1.48	1.31	1.16	1.05
Subspace	1.62	1.44	1.26	1.07
FB-RBPF	1.64	1.51	1.34	1.10
DCT-WF	1.48	1.35	1.23	1.09
DCT-RBPF	1.25	1.07	0.89	0.72
Hybrid DCT	1.32	1.18	1.06	0.92

Table 5.5: Performance results for street noise.

	WPESQ Scores			
Algorithm	5 dB	10 dB	15 dB	20 dB
Noisy	1.18	1.45	1.96	2.51
MMSE-LSA	1.95	2.29	2.63	2.90
Subspace	1.50	1.87	2.35	2.85
FB-RBPF	1.22	1.51	1.95	2.50
DCT-WF	1.53	1.75	2.05	2.38
DCT-RBPF	1.77	2.16	2.62	3.09
Hybrid DCT	1.69	2.01	2.38	2.67
	CSII Scores			
Noisy	0.60	0.84	0.94	0.97
MMSE-LSA	0.73	0.85	0.93	0.96
Subspace	0.78	0.91	0.97	0.98
FB-RBPF	0.60	0.82	0.94	0.97
DCT-WF	0.56	0.73	0.86	0.93
DCT-RBPF	0.75	0.90	0.96	0.98
Hybrid DCT	0.61	0.79	0.90	0.94
	LLR Distance			
Noisy	0.98	0.75	0.55	0.40
MMSE-LSA	0.98	0.93	0.90	0.88
Subspace	1.25	1.13	0.74	0.61
FB-RBPF	1.06	0.84	0.64	0.47
DCT WF	1.17	1.04	0.94	0.85
DCT-RBPF	0.78	0.68	0.63	0.61
Hybrid DCT	1.02	0.94	0.88	0.85

Table 5.6: Performance results for cafeteria babble noise.

	WPESQ Scores			
Algorithm	5 dB	10 dB	15 dB	20 dB
Noisy	1.20	1.57	2.19	2.99
MMSE-LSA	1.61	2.09	2.60	3.04
Subspace	1.19	1.50	2.02	2.69
FB-RBPF	1.44	1.85	2.46	3.21
DCT-WF	1.37	1.83	2.49	3.17
DCT-RBPF	1.42	1.98	2.65	3.29
Hybrid DCT	1.35	1.78	2.26	2.71
	CSII Scores			
Noisy	0.38	0.67	0.87	0.96
MMSE-LSA	0.51	0.71	0.85	0.93
Subspace	0.50	0.74	0.89	0.96
FB-RBPF	0.34	0.55	0.69	0.78
DCT-WF	0.32	0.50	0.68	0.84
DCT-RBPF	0.51	0.75	0.91	0.97
Hybrid DCT	0.41	0.65	0.83	0.93
	LLR Distance			
Noisy	0.49	0.33	0.20	0.11
MMSE-LSA	0.56	0.51	0.50	0.50
Subspace	1.31	0.88	0.60	0.45
FB-RBPF	0.50	0.37	0.28	0.21
DCT-WF	1.00	0.79	0.60	0.48
DCT-RBPF	0.41	0.29	0.27	0.28
Hybrid DCT	0.76	0.62	0.52	0.46

operating in the transform domain would facilitate the computation and application of a perceptual post-filter that takes into account noise masking thresholds.

In a comparative study using real recorded wideband speech and noise signals, the DCT-RBPF achieved higher scores on objective evaluations designed to estimate speech quality and intelligibility. Notably, despite its reduced complexity, the DCT-RBPF provides a significant improvement over the fullband RBPF, displaying lower residual noise levels without increased speech distortion.

Chapter 6

Particle Filter Enhancement of Speech Spectral Amplitudes

The DCT-RPBF algorithm presented in Chapter 5 achieves transform model-based speech enhancement by adding a frequency division to a model-based speech enhancement algorithm. This frequency division is shown to increase model flexibility and improve the enhancement performance for wideband speech. In this chapter, the transform model-based enhancement is achived by adding a dynamic model into spectral amplitude speech enhancement to reduce the need for heuristic smoothing that can distort speech onsets and remove transient speech components.

As discussed in Chapter 2, statistically motivated short time spectral amplitude (STSA) speech enhancement approaches treat the recovery of the clean speech spectral amplitudes as a Bayesian estimation problem. As phase is relatively unimportant to perceived speech quality [29], it is common to restore the spectral amplitude and use the noisy phase to reconstruct the enhanced signal. Spectral amplitude speech enhancement offers high levels of noise reduction and the frequency-dependent frame-based processing makes it desirable for wideband speech processing. Closed form minimum mean-square error (MMSE) Bayesian spectral amplitude estimators have

been derived under different assumptions for speech and noise including the standard Gaussian prior [12] [33], as well as super-Gaussian priors [88] [36]. All of these existing spectral amplitude speech enhancement algorithms assume the spectral amplitudes in successive frames are statistically independent. As discussed in [34] this inter-frame independence assumption breaks down when short and/or overlapping frames are used for the spectral analysis, and it also conflicts with the decision-directed method [12] which is typically used to estimate the *a-priori* SNR. The smoothing parameter α in the decision-directed estimator is selected heuristically to balance the tradeoff between under-smoothing, which leads to SNR fluctuations and musical noise, and over-smoothing which can remove low-amplitude transient speech components and distort speech onsets [8].

The non-linear recursive averaging used by the decision-directed approach implicitly assumes a dependence between successive frames while the algorithm derivation assumes they are independent conditioned on the spectral power. This contradiction is discussed in [34], where an attempt is made to rectify the contradiction by deriving estimators for the *a-posteriori* SNR that take time-correlation into account. The presented results show that, when used in conjunction with standard spectral amplitude enhancement algorithms, the performance of the causal estimator is equivalent to that of the decision-directed estimator. The success of the decision-directed estimator demonstrates that in order to reduce distortion of speech cues, speech enhancement algorithms should account for the high correlation of spectral amplitudes in consecutive frames.

While inter-frame dependence is not typically exploited in STSA estimation, dynamic model based algorithms have been applied to enhancement of the complex DFT coefficients. In [89] the real and imaginary parts of the clean signal DFT coefficients are assumed independent and are modeled as Gaussian-excited auto-regressive

(AR) processes. The AR parameters are estimated from the past enhanced signal, and Kalman filtering is applied to recover the clean signal coefficients. In [90], rather than separately estimating the real and imaginary parts, the clean signal DFT coefficients are modeled as a complex AR process and are recovered using a two-step propagate-update recursion. In the update step, the prediction error is estimated by denoising the difference between the measurement and the propagated DFT coefficients. Three update rules are considered, corresponding to different assumed prediction error distributions, the rules derived from the super-Gaussian estimators in [88] and [36] are found to offer better performance than the standard Gaussian Kalman filter equations. As in [89] the AR parameters are estimated from the past enhanced signal. The presented results show that incorporating the dynamic model improves performance compared to traditional statistical estimators. Unfortunately removing the inter-frame independence assumption leads to spectral amplitude estimators that cannot be solved in closed form.

To address this issue we introduce a particle filter approach for short-time spectral amplitude speech enhancement. The particle filter framework enables modelling the time dependence of the spectral amplitudes; rather than estimating the parameters of the instantaneous distribution of the spectral amplitudes, we parameterize a model of the evolution of the spectral amplitudes. The distribution of the speech spectral amplitudes evolves online, allowing us to account for the inter-frame correlation and the non-Gaussian speech statistics observed in [88] and [36]. Two variants of the standard algorithm are also presented: one that uses an interacting multiple model approach to account for transitions between active speech and silence intervals, and one that allows for phase differences between the clean speech and noise complex Fourier transform coefficients. All of the particle sampling distributions are constrained to take the measurement into account, improving sampling efficiency.

6.1 Spectral Amplitude Particle Filtering

In this section we develop the importance sampling and weighting steps required to implement a particle filter algorithm for the restoration of speech short-time spectral amplitudes. The basic algorithm acts as a framework, and imposes several limiting assumptions which are eased in the subsequent algorithm variations presented in Section 6.2; the impact of these assumptions on algorithm performance is examined in Section 6.3. The proposed approach shares some similarities with the existing particle filter solutions discussed in Chapter 2, but differs by operating directly in the spectral amplitude domain, modeling the evolution of the clean speech DFT magnitude series.

Conceptually the proposed algorithm can described as follows: candidate spectral amplitudes are generated using a dynamic model of spectral amplitude evolution with online estimation of the model parameters; the likelihood of each candidate is computed using the background noise density, and those likelihood weights are then used to compute the MMSE estimate of the clean signal amplitude as a weighted sum of the candidate amplitudes; and resampling is performed to select the most likely candidates to be propagated to the next frame. Figure 6.1 presents a block diagram comparison of standard spectral amplitude estimation, where the parameters of the static spectral amplitude distribution are estimated directly from the noisy signal, and the proposed particle filter approach, where the parameters of the spectral amplitude dynamic *process* are estimated.

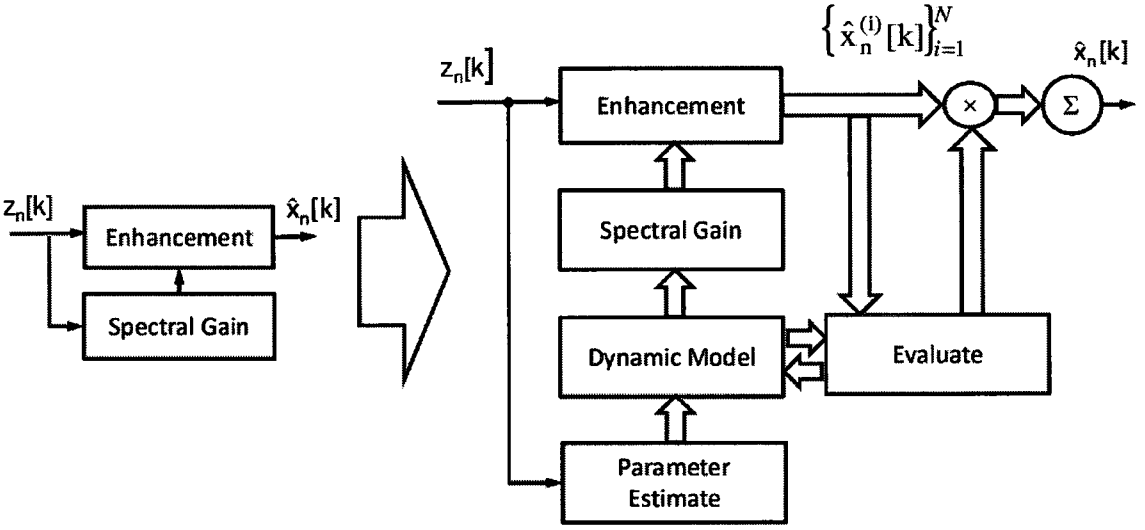


Figure 6.1: Block diagram comparison of traditional spectral amplitude estimation (left) and particle filter based spectral estimation (right).

6.1.1 Measurement Model

Let $x_n(k)$ be a shorthand notation to denote the magnitude of the k^{th} coefficient of a length-N DFT of the clean signal $x[n]$, windowed with a function $w[t]$.

$$x_n(k) \triangleq |X_n(k)| = \left| \sum_{t=n-N}^{n} w[t]x[n-t]e^{-\frac{j2t\pi}{N}k} \right| \tag{6.1}$$

The DFT magnitudes in each frequency bin are assumed to be mutually independent, therefore each bin is processed independently and the frequency index is dropped in the following. This assumption is made for computational convenience, and only holds as the frame length approaches infinity [12], therefore the dynamic models employed may not exploit all of the correlation information present in the signal. Using the noisy signal phase to reconstruct the enhanced signal assumes the speech and noise signal DFT coefficients are in phase so their magnitudes are additive, giving the

measurement equation:

$$z_n = x_n + v_n \tag{6.2}$$

This basic linear measurement model is restrictive, however the simplified description facilitates the development of the importance sampling and weighting steps used in the algorithms that follow. As will be discussed in Section 6.2.1, assuming the signal and noise to be in phase reduces the achievable enhancement performance especially in low SNR environments.

Assuming the real and imaginary parts of the background noise DFT coefficients are independent zero-mean Gaussian random variables with variance σ_v^2, the background noise spectral magnitude v_n will be Rayleigh distributed with parameter σ_v. The Rayleigh distribution has been shown to accurately model the spectral amplitudes of a variety of real noise sources [36].

6.1.2 Speech Dynamic Model

Rather than assuming independence of successive frames, we instead model the evolution of the active speech spectral amplitude series as a first-order TVAR model:

$$x_n = a_n x_{n-1} + d_n \tag{6.3}$$

where a_n is the (possibly time-varying) AR coefficient and d_n is the process excitation..

This model gives the prior of x_n expressed in terms of $p_D(d)$, the density of d_n:

$$p(x_n|x_{n-1}) = p_D(x_n - a_n x_{n-1}) \tag{6.4}$$

The TVAR model is simple and can effectively account for the high correlation between successive frames. In this work we consider fixed and time-varying AR(1), as

well as random walk models for the spectral amplitude series. The random walk can be seen as a special case of a fixed AR(1) model where the AR parameter is unity, in this case the excitation, d_n, reduces to the difference between successive spectral amplitudes. The impact of the model choice will be explored in 6.3.

Any distribution can be used for the excitation sequence without substantially changing the algorithm, however in this work a Laplace distribution is assumed. The extreme value behavior of the speech signal present in both the time and frequency domains is well-modeled by the Laplace distribution, and as a result it has been used to model the samples of the speech waveform [91], the real and imaginary parts of the speech DFT coefficients [88] [36], and the complex DFT prediction error [90]. Its use here to model the spectral amplitude excitation is also supported by the following empirical measurements of the normalized differential error. Using $N = 512$ sample Hamming windowed frames with 50% overlap, the difference between successive spectral amplitudes was computed for 2 minutes of continuous speech from four different speakers, and was normalized by the smoothed spectral amplitude to account for signal non-stationarity and differences in signal power between frequency bins:

$$\delta_n \approx \frac{x_n - x_{n-1}}{\overline{|x_n|}} \tag{6.5}$$

Fig. 6.2 plots the empirical histogram along with the maximum likelihood fit Gaussian and Laplace PDFs of normalized spectral amplitude difference errors; it is clear that the sharp modal peak and heavy tails of the Laplace density provide a better fit to the data than the Gaussian distribution.

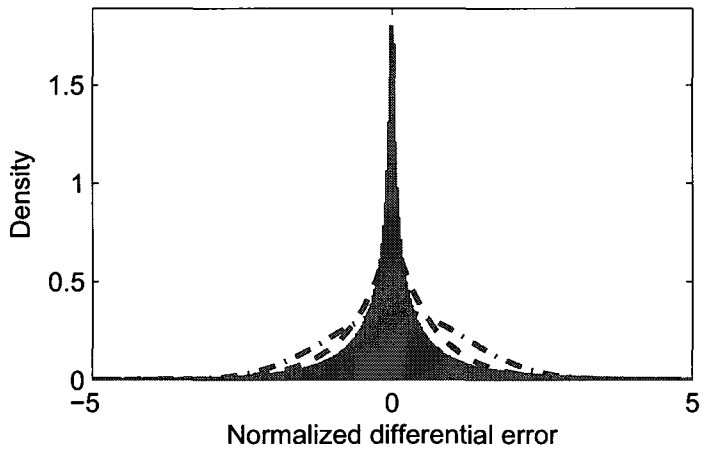


Figure 6.2: Normalized spectral amplitude differential series empirical histogram and maximum likelihood fit Gaussian (dash-dot) and Laplace (dash) densities.

Importance Sampling

Combining (6.3) and (6.2) provides a state-space description of the evolution of the spectral amplitudes, permitting sequential estimation of the posterior density $p(x_n|z_{1:n})$ within an SIR particle filter framework.

Using $\mathcal{L}(x;\mu,b)$ to represent a Laplace density with mean μ and shape parameter b evaluated at x, the conditional density of the clean speech spectral amplitudes under the TVAR model is given by:

$$p_D(x_n - a_n x_{n-1}) = \mathcal{L}(x_n; a_n x_{n-1}, b_{d_n}). \tag{6.6}$$

Candidate spectral amplitudes $x_n^{(i)}$ can be produced by generating samples $d_n^{(i)}$ from the process noise distribution and propagating them along with the particle from the previous sampling interval $x_{n-1}^{(i)}$ through the transition equation (6.3). However, if this approach is used, the un-restricted domain of d_n could result in candidate spectral amplitudes that are negative or greater than the noisy signal spectral amplitude z_n,

leading to an invalid enhanced signal spectrum or noise amplification. In spectral subtraction algorithms and in the decision-directed SNR estimator the estimated spectral magnitude or power is lower-bounded or half-wave rectified to ensure it is non-negative. Particle filters offer several integrated approaches to deal with such constraints.

The simplest approach is to assign a zero importance weight to the invalid draws so that they will have no bearing on the computation of the final estimated signal amplitude. As discussed in [92], this approach can lead to many zero-valued particles, which reduces sample diversity and wastes computational effort. In [92] a rejection-sampling like fast acceptance test is derived to ensure that only valid particles are generated. Alternatively, an auxiliary particle filter [59] could be used to propagate particles with a high predictive likelihood, but invalid draws could still result. Since the range of valid spectral amplitudes is known at the time of sampling, we propose to adapt the density used to draw the process noise samples to take the constraints into account. By limiting the support of the distribution used for sampling $d_n^{(i)}$, the generated particles can be confined to the valid range: $0 \leq x_n^{(i)} \leq z_n$. The density of this constrained excitation signal can be written as:

$$p_D(d_n^{(i)}) \propto \begin{cases} \mathcal{L}(d_n^{(i)}; 0, b_{d_n}), & -a_n x_{n-1}^{(i)} \leq d_n^{(i)} \leq z_n - a_n x_{n-1}^{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{6.7}$$

Laplace distributed excitation samples $d_n^{(i)}$ can be generated by transforming a uniform random variable $u^{(i)}$ using the inverse of the cumulative density function

(CDF) $F_D(d)$. For a zero-mean Laplace distribution:

$$F_D(d) = 0.5[1 + \text{sgn}(d)(1 - e^{-|d|/b})] \tag{6.8}$$

$$F_D^{-1}(d) = -b\ \text{sgn}(d - 0.5)\ \log(1 - 2|d - 0.5|). \tag{6.9}$$

Constraining the excitation samples to the range $(-a_n x_{n-1}, z_n - a_n x_{n-1})$ corresponds to limiting the uniform random draw to the range $(F_D(-a_n x_{n-1}), F_D(z_n - a_n x_{n-1}))$, resulting in the following transformation to generate the samples:

$$u^{(i)} \sim \mathcal{U}(F_D(-a_n x_{n-1}^{(i)}), F_D(z_n - a_n x_{n-1}^{(i)})) \tag{6.10}$$

$$\begin{aligned} d_n^{(i)} &= F_D^{-1}(u^{(i)}) \\ &= -b\,\text{sgn}(u^{(i)} - 0.5)\log(1 - 2|u^{(i)} - 0.5|) \end{aligned} \tag{6.11}$$

Importance Weighting

Employing the SIR particle filter framework the importance density is the prior, and the importance weight reduces to the likelihood:

$$W_n^{(i)} = p(z_n|x_n^{(i)}) \tag{6.12}$$

For Rayleigh distributed noise using the linear measurement model of (6.2) the likelihood is given by:

$$p(z_n|x_n) = p_V(z_n - x_n) \tag{6.13}$$

$$= \begin{cases} \frac{(z_n - x_n)}{\sigma_{v_n}^2} e^{-(z_n - x_n)^2/2\sigma_{v_n}^2} & x_n \leq z_n \\ 0 & x_n > z_n \end{cases} \tag{6.14}$$

Using the particle filter framework gives the proposed algorithm distribution flexibility. If the background noise statistics are known to be non-Rayleigh, different noise distributions can be accommodated by replacing (6.14) with a different probability density function.

Estimation of Signal and Noise Parameters

Particle generation requires the TVAR parameter a_n and the excitation Laplace parameter b_{d_n}, while particle weighting requires the noise Rayleigh parameter σ_{v_n}. Since the algorithm operates in the spectral amplitude domain, the parameters of the noise distribution used in the weight evaluation step can be estimated during speech pauses, or continuously on-line using existing techniques such as the minimum statistics approach in [93].

If the TVAR coefficients are known, the excitation variance can be estimated online using the past enhanced signal. However, the directly-measured prediction error contains noise:

$$e_n = z_n - a_n \hat{x}_{n-1} \tag{6.15}$$

$$\mathbb{E}\{e_n^2\} = \mathbb{E}\{(z_n - a_n x_{n-1})^2\} \tag{6.16}$$

$$= \mathbb{E}\{(x_n - a_n x_{n-1})^2\} + \mathbb{E}\{v_n^2\}. \tag{6.17}$$

For Rayleigh distributed noise with parameter σ_{v_n}:

$$\mathbb{E}\{v_n^2\} = 2\sigma_{v_n}^2 \tag{6.18}$$

the excitation power estimate can therefore be compensated as:

$$\mathbb{E}\{d_n^2\} \approx \mathbb{E}\{e_n^2\} - 2\sigma_{v_n}^2 \tag{6.19}$$

Using recursive smoothing to estimate the differential error variance, $\overline{e^2}$, gives the following steps to update the excitation power:

$$e_n = z_n - a_n \hat{x}_{n-1} \tag{6.20}$$

$$\overline{e^2} = \beta \overline{e^2} + (1-\beta)|e_n|^2 \tag{6.21}$$

$$\sigma_{d_n}^2 = \overline{e^2} - 2\sigma_{v_n}^2 \tag{6.22}$$

$$b_{d_n} = \sqrt{\frac{\sigma_{d_n}^2}{2}}. \tag{6.23}$$

This parameter estimation approach provides a larger variance, and therefore a broader proposal density, when the differential error is high – corresponding to periods where the spectral magnitude is changing rapidly. In contrast, when the error is low, such as during voiced phonemes or speech pauses, a more narrow proposal is used for particle selection, improving sampling efficiency and reducing musical noise.

Several approaches have been proposed for the estimation of the TVAR coefficients a_n. In [89] and [90] the parameters of the DFT coefficient AR process are estimated directly from the past enhanced signal. Among the particle filter algorithms, in [63] fixed AR models are considered for the noise process and in [64] fixed and dynamic models are investigated. In the fixed models, the AR coefficients are estimated offline, while in the dynamic model they are computed online using the particle estimates. In [65] and [60] the TVAR parameters and the log excitation power are assumed to follow random walks and are grouped along with the clean speech vector into an augmented state vector for particle filter estimation.

Using the approach of [65] and [60] to estimate the parameters within the particle filter, we can expand the prior of the augmented state vector $[a_n\ , x_n]$ as:

$$p(x_n, a_n|x_{n-1}, a_{n-1}) = p(x_n|x_{n-1}, a_{n-1}, a_n)p(a_n|x_{n-1}, a_{n-1}) \tag{6.24}$$

Making the following two assumptions: first, the clean speech spectral amplitudes, x_n, are conditionally independent of the past TVAR coefficients, a_{n-1}, given the current coefficients, a_n; second, the current TVAR coefficients, a_n, are conditionally independent of the past clean speech amplitude, x_{n-1}, given the past TVAR coefficients, a_{n-1}, this simplifies to:

$$p(x_n, a_n|x_{n-1}, a_{n-1}) = p(x_n|x_{n-1}, a_n)p(a_n|a_{n-1}) \tag{6.25}$$

Using the SIR framework we can generate particles to jointly estimate the clean spectral amplitudes and the TVAR process coefficients by first simulating the TVAR coefficient transition $a_n^{(i)} \sim p(a_n|a_{n-1}^{(i)})$, and using that TVAR vector to propagate the spectral amplitudes $x_n^{(i)} \sim p(x_n|x_{n-1}^{(i)}, a_n^{(i)})$. In [65] a constrained Gaussian random walk is assumed for the TVAR coefficients:

$$p(a_n|a_{n-1}) = \begin{cases} \mathcal{N}(a_n; a_{n-1}, \sigma_{a_n})^2 & \text{If stable} \\ 0 & \text{otherwise} \end{cases} \tag{6.26}$$

where σ_{a_n} is the variance of the random walk, a hyper-parameter which is chosen in advance, and $\mathcal{N}(x; \mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2 evaluated at x.

This type of estimation within the particle filter framework offers flexibility and

circumvents the difficulty of measuring the parameters from the noise-corrupted signal, however it also requires a meta-model for the parameter evolution which increases complexity and adds extra Monte-Carlo variation. Furthermore, the use of augmented state vectors does not come without cost, as increasing the state dimension requires more particles to cover the state-space [84]. In addition, since the prediction error in (6.20) depends on the TVAR parameters, an excitation parameter must be computed for each particle. In Section 6.3 we compare the performance of particle filter on-line estimation with fixed off-line coefficient estimation.

Combining the importance sampling and particle weighting steps with the parameter estimation steps, an SIR particle filtering algorithm for estimating the spectral amplitude x_n is presented in Algorithm 3. If a static AR process is assumed the TVAR coefficient update step is omitted.

Algorithm 3 STSA-PF

for $n \geq 1$ **do**
    for $i = 1, \ldots, N$ **do**
        (Optionally) draw new TVAR coefficients according to (6.26).
        Update excitation parameter using (6.20)–(6.22).
        Draw constrained process noise samples using (6.11).
        Propagate particles using process noise samples in (6.3).
    end for
    Evaluate the importance weights, up to a constant using (6.12).
    Normalize the entire set of importance weights.
    Compute Bayesian posterior estimate.
    Resample.
end for

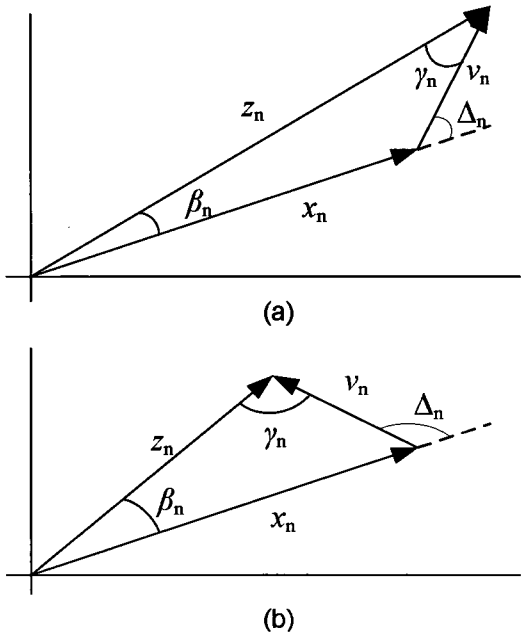
6.2 Algorithm Variations

6.2.1 Phase Estimation

The assumption of (6.2), that the clean signal and noise DFT coefficients are in phase, is restrictive and not realistic in practice. Furthermore, while the phase is not important to the listening quality of enhanced speech, the clean spectral magnitude cannot be recovered exactly without phase information, even if the noise magnitude is known exactly [27]. In this section we develop a variant of the STSA-PF algorithm that incorporates phase information to improve the clean signal magnitude estimate.

Particle filter estimation of the phase requires a state-space model of the phase process. Dynamic modeling is not expected to be beneficial as phase is not predictable between successive frames [90], instead we consider static phase models. It is common to assume that the clean speech DFT phase is uniformly distributed on $[-\pi, \pi]$; for example, in [36] this is shown to provide a good fit to experimental data. A simple approach to phase estimation would be to directly generate clean signal phase samples from this density, however the uniform distribution is an uninformative prior and does not make use of any of the phase information contained in the measurement. For example, at very high SNRs the clean and noisy DFT vectors will be almost co-linear, and the noisy measurement signal phase will approach that of the clean signal; likewise, at very low SNRs the measurement signal phase will be mostly determined by the noise signal [27]. Rather than directly generating candidate phases, we propose to incorporate phase information by modeling the phase difference between the clean signal and noise DFT coefficients; this difference is not SNR dependent, as the signal and noise phases are assumed independent.

Figure 6.3: Geometric representation clean speech, noise and measurement DFT vectors in the complex plane (after [30]), constructive noise (a), destructive noise (b).



Importance Sampling

Let $\Delta_n = \angle X_n - \angle V_n$ be the difference between the clean speech and noise DFT phases. The value of Δ_n determines how the clean speech and noise DFT vectors add in the complex plane which affects the magnitude of their vector sum. If Δ_n lies in the left half plane (ie. $\cos(\Delta_n) < 0$) then the noise will add destructively to the signal vector, giving a measurement signal with a smaller magnitude than the clean signal. If $\cos(\Delta_n) \geq 0$, then the noise will add constructively. Fig. 6.3 presents a geometric representation of the clean speech, noise and noisy speech DFT coefficients in the complex plane for the constructive and destructive noise cases. The magnitudes of the speech noise and measurement DFT vectors are denoted x_n, v_n and z_n; while the relative angles between the speech and noise, speech and measurement and noise and measurement vectors are denoted Δ_n, β_n and γ_n respectively.

Assuming $\angle X_n, \angle V_n$ to be independent and identically distributed $\mathcal{U}[-\pi, \pi]$ random variables, then Δ_n will have a symmetric triangle distribution on $[-2\pi, 2\pi]$, which taken modulo-2π (as the phase is 2π shift invariant) also becomes a $\mathcal{U}[-\pi, \pi]$ random variable. Consecutive phase differences are uncorrelated, and the phase difference can also be assumed to be independent of the clean signal spectral amplitude, therefore we can sample the phase component of the particles as:

$$\Delta_n^{(i)} \sim p(\Delta_n) = \mathcal{U}[-\pi, \pi] \tag{6.27}$$

Having generated the phase difference according to (6.27), it can be used along with the measurement z_n to determine the boundaries for the clean speech candidate spectral amplitude. In the additive noise case, the allowable range of spectral amplitudes remains $[0, z_n]$, as in the simplified linear model of (6.2), however, for the destructive noise case the measurement z_n is the *lower* bound on the range. The upper bound can be found by applying the law of sines to the destructive noise triangle in Fig. 6.3(b), noting that $\sin(\pi - \Delta_n) = \sin(\Delta_n)$. When Δ_n lies in the upper left quadrant, $\pi/2 \leq \Delta_n \leq \pi$, we have:

$$\frac{\sin(\pi - \Delta_n)}{z_n} = \frac{\sin(\gamma_n)}{x_n} \tag{6.28}$$

$$\Rightarrow x_n = \frac{\sin(\gamma_n)}{\sin(\Delta_n)} z_n \tag{6.29}$$

$$\leq \frac{z_n}{\sin(\Delta_n)}. \tag{6.30}$$

A similar analysis can be carried out for the lower left quadrant, in general the range of allowable spectral amplitudes for the destructive noise case is $[z_n, z_n/\sin(|\Delta_n|)]$.

Generating particles by independently drawing the phase difference component $\Delta_n^{(i)}$, and using the appropriate boundaries to generate the state component $x_n^{(i)}$ as

described in section 6.1, gives the joint proposal density:

$$q(x_n, \Delta_n | x_{n-1}^{(i)}, \Delta_{n-1}^{(i)}, z_n) \propto p(\Delta_n) p(x_n | x_{n-1}^{(i)}, \Delta_n) \tag{6.31}$$

Importance Weighting

To derive the weight update equation, we expand the prior term using the following assumptions: first that successive phase differences are independent of one another, which is supported by the results in [90] showing little correlation between successive speech phases; and second, that the phase is independent of the spectral amplitudes, which is supported by the observation from [36] that the empirical speech DFT coefficient distribution is rotationally invariant. Applying these assumptions:

$$p(x_n, \Delta_n | x_{n-1}, \Delta_{n-1}) = p(\Delta_n | x_{n-1}, \Delta_{n-1}, x_n) p(x_n | x_{n-1}, \Delta_{n-1}, \Delta_n) \tag{6.32}$$

$$= p(\Delta_n) p(x_n | x_{n-1}, \Delta_n). \tag{6.33}$$

Using the proposal of (6.31), the prior and the proposal cancel and the weight computation reduces to the SIR form:

$$W_n^{(i)} = p(z_n | x_n^{(i)}, \Delta_n^{(i)}). \tag{6.34}$$

To evaluate this likelihood we require an expression for the measurement z_n in terms of the state x_n and the phase difference Δ_n. Applying the law of cosines to the triangles of Fig. 6.3 we get the non-linear measurement equation:

$$z_n = \sqrt{v_n^2 + x_n^2 + 2 v_n x_n \cos(\Delta_n)} \tag{6.35}$$

where we have used the fact that $\cos(\pi - \Delta_n) = -\cos(\Delta_n)$. Solving for v_n to evaluate the likelihood term gives the likelihood computation in terms of the noise density $p_V(v)$:

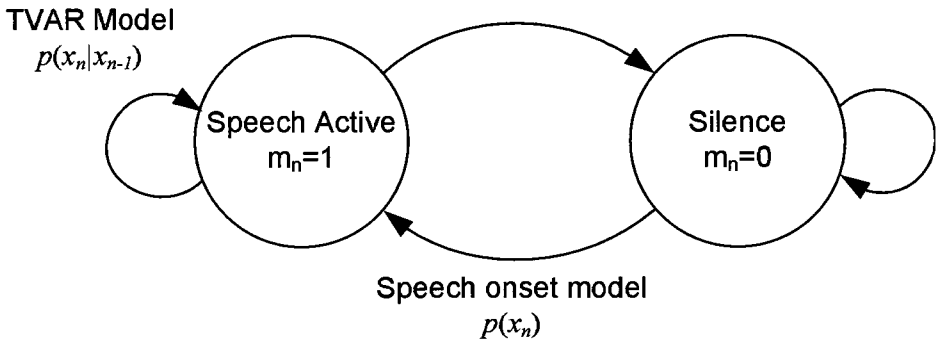
$$p(z_n|x_n^{(i)}, \Delta_n^{(i)}) = p_V\left(-x_n^{(i)}\cos(\Delta_n^{(i)}) + \sqrt{(x_n^{(i)})^2[\cos^2(\Delta_n^{(i)}) - 1] + z_n^2}\right). \quad (6.36)$$

The enhanced amplitude can be combined with the noisy measurement phase to synthesize the output. Alternatively, the particle estimates of x_n and Δ_n can also used to determine β_n, the phase difference between the clean speech and measurement signal DFT vectors, which can then be used to compensate the phase as well as the spectral amplitude in producing the enhanced output. This approach was not found to improve performance compared to using the uncompensated noisy phase.

6.2.2 Interacting Multiple Model

The basic STSA-PF algorithm of section 6.1 assumes speech is always present, however statistical estimation algorithms report increased noise attenuation when speech presence uncertainty is taken into account [12]. Here the dynamics of spectral amplitudes with speech present and absent periods are modeled by a hierarchical state space model, where the system dynamics vary depending on the active regime or mode of operation. We consider speech active and silence states, and define different models to describe the evolution of spectral amplitudes within and between the states. By accounting for the state transitions we seek to achieve less distortion at perceptually important speech onsets and more noise reduction at speech/silence boundaries. Multiple dynamic models of speech evolution are also considered in the cepstral feature enhancement algorithm presented in [94]. A switching linear dynamic model is applied, wherein the model parameters are dependent on a hidden variable

Figure 6.4: State transitions for the active speech/silence model particle filter. No dynamic models are used to model the transitions terminating in the silence state.



representing the index of the parameters in a fixed set. In contrast to the approach we outline here, Gaussian distributions are assumed for the cepstral features, permitting a generalized pseudo-Bayesian estimation approach, and the sets of model parameters are determined offline using training data and do not represent defined speech events.

We define two modes of operation for the spectral amplitude series, $m_n = 1$ corresponding to speech active and $m_n = 0$ to silence. Fig. 6.4 summarizes the dynamic models describing the possible state transitions. Since the spectral amplitudes are identically zero during silence, no dynamic model is needed to describe transitions to and within the silence state; we only require models of the transitions within the speech active state, and from silence to speech active. For transitions within the speech active state we use the TVAR model given by (6.3). Transitions from silence to active speech represent speech onsets, where the spectral amplitudes undergo a sudden transition from zero to the active speech amplitude. At these onsets successive spectral amplitudes are independent and the prior reduces to:

$$p(x_n|x_{n-1}, m_n = 1, m_{n-1} = 0) = p(x_n). \tag{6.37}$$

For $p(x_n)$ we use the exponential distribution as it is peaked towards zero, providing

for the generation of low-amplitude speech onsets, and is linear in the argument of the exponent. In [36] it is shown that densities with thicker tails resulting from the linear argument provide a better fit to empirical spectral amplitude data than distributions with a quadratic argument, such as the Rayleigh distribution that would result from Gaussian distributed DFT coefficient components. The choice also provides for easy sampling, as particles distributed according an exponential distribution with parameter λ_n constrained to the range $[0, z_n]$ can be generated by inverse transform sampling as:

$$x_n^{(i)} = -\lambda_n \log u^{(i)} \qquad u^{(i)} \sim \mathcal{U}[e^{-z_n/\lambda_n}, 1] \tag{6.38}$$

$$= -\lambda_n \log[u^{(i)}(1 - e^{-z_n/\lambda_n}) + e^{-z_n/\lambda_n}] \qquad u^{(i)} \sim \mathcal{U}(0, 1] \tag{6.39}$$

Since consecutive frames are independent at speech onsets, the prediction error approaches the clean signal spectral amplitude. Therefore we approximate the speech onset model exponential parameter, λ_n, as the inverse of the absolute value of the instantaneous prediction error from (6.20):

$$\lambda_n = \frac{1}{|e_n|}. \tag{6.40}$$

Model Mixing

Having defined the dynamic models to describe the mode transitions, we require an algorithm for mode mixing and overall state estimation. The method we use is a modification of the approach from [95], where an extension to the SIR particle filter is proposed for hierarchical state-space estimation. Since keeping track of a history of branched state transitions leads to an exponential memory and computational burden, an interacting multiple model (IMM) approach is used whereby all transitions

into a given state are merged into a single density. To incorporate the multiple model into the SIR particle filter framework, each particle consists of an augmented state vector $[x_n^{(i)}, m_n^{(i)}]$, where $m_n^{(i)}$ is the regime in effect during the sampling interval and determines the dynamic model used to propagate the state estimate. Particles are generated by first simulating the regime transition probability $m_n^{(i)} \sim p(m_n|x_{n-1}^{(i)}, m_{n-1}^{(i)})$, then updating the particle state using the corresponding dynamic model:

$$x_n^{(i)} \sim p(x_n|x_{n-1}^{(i)}, m_{n-1}^{(i)}, m_n^{(i)}) \tag{6.41}$$

The particles are weighted using the likelihood, and resampling of the augmented state vectors is performed using the likelihood weights to produce a multi-modal posterior density. Since each particle contains an estimate of the regime in effect during the sampling interval, the regime probabilities are represented by the proportion of resampled particles from the corresponding mode.

The modes we have defined prevent us from using this approach directly. In the speech absent mode, the spectral amplitudes are identically zero, therefore the mode-conditioned density does not use a particle representation, preventing us from representing the mode probabilities with resampled particle proportions. Instead we use the alternative scheme presented in [96], which allows a fixed number of particles to be allocated to each dynamic model. A particle filter is run for each mode of operation, assuming it to be in effect during the sampling interval; this produces a set of mode-conditioned densities, and the overall posterior is a mixture of these densities weighted by the mode probability. The mode probabilities are propagated using the mode transition probabilities and updated using the mode-conditioned likelihood weights.

Algorithm Description

Mode branching is performed by propagating the posterior mode probability from the previous frame with the *a-priori* mode transition probability:

$$p(m_n, m_{n-1}|z_{1:n-1}) \propto p(m_n|m_{n-1})p(m_{n-1}|z_{1:n-1}) \tag{6.42}$$

We use a Markov switching matrix for the *a-priori* mode transition probabilities:

$$p(m_n|m_{n-1}) = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \tag{6.43}$$

where p_{jk}, $j, k \in \{0, 1\}$, are active-speech-to-silence or silence-to-speech onset transition probabilities. In this work they are assumed to be equal and fixed, however they could also be estimated online.

The proportion of particles allocated to the speech-to-speech and silence-to-speech dynamic models is dictated by the probability of each type of mode transition, as computed using (6.42). Specifically, the number of particles allocated to the speech-to-speech model is given by:

$$N_1 = N_p \cdot \frac{p(m_n = 1, m_{n-1} = 1|z_{1:n-1})}{p(m_n = 1, m_{n-1} = 0|z_{1:n-1}) + p(m_n = 1, m_{n-1} = 1|z_{1:n-1})} \tag{6.44}$$

where N_p is the total number of particles.

As in [96], a state estimate is computed for each mode assuming that mode to be active during the measurement interval. The active-speech mode-conditioned state

estimate, $\hat{x}_n = \mathbb{E}\{p(x_n|m_n = 1, z_{1:n})\}$, is given from the particle filter:

$$\hat{x}_n = \sum_{i=1}^{N_p} W_n^{(i)} x_n^{(i)} \tag{6.45}$$

while the silence mode-conditioned state estimate is identically zero.

The state estimate is a mixture of the active-speech and mode-conditioned estimates, with the mixing weights determined by the mode probabilities. Since the silence mode-conditioned estimate is zero, this is equivalent to the active speech state estimate provided by the particle filter, scaled by the posterior mode probability of active speech:

$$p(x_n|z_{1:n}) = p(x_n|m_n = 1, z_{1:n})p(m_n = 1|z_{1:n}) + p(x_n|m_n = 0, z_{1:n})p(m_n = 0|z_{1:n}) \tag{6.46}$$

$$= p(x_n|m_n = 1, z_{1:n})p(m_n = 1|z_{1:n}). \tag{6.47}$$

The mode probabilities are computed using a predict-update approach. The *a-priori* mode probability $p(m_n|z_{1:n-1})$, is first predicted using the mode transition probabilities in (6.42):

$$p(m_n = j|z_{1:n-1}) = \sum_{l=0,1} p(m_n = j|m_{n-1} = l)p(m_{n-1} = l|z_{1:n-1}) \quad j \in \{0, 1\} \tag{6.48}$$

then updated using the mode-conditioned likelihoods.

$$\tilde{p}(m_n = j|z_{1:n}) = p(z_n|m_n = j)p(m_n = j|z_{1:n-1}) \qquad j \in \{0, 1\} \tag{6.49}$$

and normalized so that the probabilities sum to unity:

$$p(m_n = j|z_{1:n}) = \frac{\tilde{p}(m_n = j|z_{1:n})}{\tilde{p}(m_n = 1|z_{1:n}) + \tilde{p}(m_n = 0|z_{1:n})} \quad j \in \{0, 1\}. \quad (6.50)$$

Where the required mode-conditioned estimate likelihoods are computed using the likelihood model (6.14):

$$p(z_n|m_n = 0) = p_V(z_n) \quad (6.51)$$

$$p(z_n|m_n = 1) = \frac{1}{N_p} \sum_{i=1}^{N_p} p_V(z_n - \hat{x}_n^{(i)}). \quad (6.52)$$

The complete interacting multiple model short-time spectral amplitude (IMM-STSA) speech enhancement algorithm is summarized in Algorithm 4.

6.3 Algorithm Evaluation

The following algorithm variants are considered in the performance comparison:

- Basic spectral amplitude particle filter presented in Algorithm 3 with a random walk model of spectral amplitude evolution. (denoted STSA-PF)
- Basic algorithm with a fixed AR(1) spectral amplitude model. The AR parameter for each frequency bin was estimated independently, using linear prediction of the spectral amplitudes of two minutes of continuous speech different from the test data. (denoted STSA-fAR)
- Basic algorithm with a dynamic TVAR(1) spectral amplitude model using a Gaussian random walk on the TVAR parameter, as in [65], with a random walk variance of 1×10^{-4}. (denoted STSA-dAR)

Algorithm 4 IMM STSA-PF

for $n \geq 1$ **do**
  Propagate mode transition probabilities using (6.42).
  Distribute particles between active speech dynamic models using (6.44).
  Model 1:
  for $i = 1, \ldots, N_1$ **do**
    Update excitation parameter using (6.20)–(6.22).
    Draw constrained process noise samples using (6.11).
    Propagate particles and noise samples using (6.3).
  end for
  Model 2:
  for $i = N_1 + 1, \ldots, N_p$ **do**
    Draw samples from constrained prior using (6.39).
  end for
  Evaluate the importance weights, up to a constant using (6.12).
  Normalize the entire set of importance weights.
  Compute active speech Bayesian posterior estimate.
  Resample.
  Use mode-conditioned likelihoods to compute posterior mode probabilities using (6.49).
  Normalize posterior model probabilities using (6.50).
  Compute enhanced speech output using (6.47).
end for

- IMM variant presented in Algorithm 4 with a random walk for active speech spectral amplitude evolution and Laplace prior model for speech onsets. For Markov mode transitions, the probability of remaining in the current state was $p_{jj} = 0.9$. This value was determined through informal experimentation on a set of sentences separate from the test data, although the objective scores are not sensitive to parameter variations of about 10%. (denoted STSA-IMM)

- Phase estimation variant described in section 6.2.1. (denoted STSA-Phase)

The algorithms were assigned a total of $N_p = 100$ particles and the smoothing parameter for the excitation estimation was set to $\beta = 0.9$, for $N = 512$ sample frames with 50% overlap at 16 kHz sampling rate, this corresponds to a time-constant of approximately 150 ms. These parameters were found, through informal experimentation, to offer good performance: providing enough particles and sufficient smoothing to capture the variations of the speech amplitude process without leading to particle degeneracy or the removal of low amplitude transient speech components.

The DCT-RBPF algorithm from Chapter 5, using 100 particles and a TVAR(1) model for each DCT coefficient, is also included for comparison.

For the clean speech inputs we used the NOIZEUS speech corpus, which contains 30 phonetically balanced IEEE sentences produced by three male and three female speakers [97]. The recordings were downsampled to 16 kHz from the original 25 kHz sampling rate. To ensure consistency between algorithms operating in different domains, noise estimation in all cases was performed by recursive averaging during speech silence periods. Three noise conditions are considered, ranging from stationary to highly non-stationary: synthetic white Gaussian noise (WGN); street noise from passing traffic recorded at an intersection with the listener facing away from the street; and multi-talker babble recorded during a meal at a university cafeteria.

Table 6.1: Objective evaluation scores for WGN.

	WPESQ Score			CSII Score		
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
Noisy	1.03	1.11	1.6	0.14	0.58	0.92
DCT-RBPF	1.26	1.90	2.58	0.23	0.72	0.93
STSA-PF	1.14	1.64	2.49	0.27	0.77	0.95
STSA-fAR	1.18	1.72	2.56	0.27	0.77	0.95
STSA-dAR	1.17	1.69	2.54	0.27	0.77	0.95
STSA-IMM	1.15	1.68	2.57	0.28	0.78	0.95
STSA-Phase	1.23	1.85	2.64	0.29	0.77	0.95

6.3.1 Simulation Results

The results of the simulations at overall SNRs of 0, 10 and 20 dB are presented in tables 6.1 – 6.3. Due to the stochastic nature of the particle filter algorithms, 20 Monte Carlo trials were run for each test condition. The trials verified the stability of the proposed algorithms: while the enhanced signal waveforms exhibit minor variations from trial to trial, these variations do not have a significant impact on the objective scores; the best case, worst case and average case scores for each test condition do not differ from one another within the precision reported. Furthermore, none of the algorithms diverged under any of the test conditions.

The STSA-PF variants compare favorably with the DCT-RBPF. At low SNRs, the Kalman filtering in the DCT-RBPF provides it with an advantage, however in the non-Gaussian noises and at high SNRs the STSA-PF is generally superior.

Using an AR model for the spectral amplitude series rather than the random walk offers a moderate improvement in objective scores, however the dynamic AR model is not found to improve on the fixed AR model. This contrasts with the results in [64]

Table 6.2: Objective evaluation scores for street noise.

	WPESQ Score			CSII Score		
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
Noisy	1.07	1.41	2.15	0.21	0.76	0.97
DCT-RBPF	1.24	1.73	2.45	0.40	0.84	0.96
STSA-PF	1.27	1.81	3.05	0.45	0.89	0.98
STSA-fAR	1.28	1.81	3.06	0.46	0.89	0.98
STSA-dAR	1.27	1.81	3.06	0.46	0.89	0.98
STSA-IMM	1.28	1.84	3.11	0.47	0.89	0.98
STSA-Phase	1.30	1.82	2.97	0.48	0.88	0.97

Table 6.3: Objective evaluation scores for cafeteria babble.

	WPESQ Score			CSII Score		
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
Noisy	1.05	1.24	1.99	0.08	0.45	0.92
DCT-RBPF	1.11	1.59	2.37	0.11	0.61	0.92
STSA-PF	1.09	1.51	2.50	0.10	0.60	0.94
STSA-fAR	1.09	1.53	2.51	0.10	0.60	0.94
STSA-dAR	1.09	1.52	2.51	0.10	0.60	0.94
STSA-IMM	1.09	1.53	2.53	0.10	0.61	0.94
STSA-Phase	1.09	1.53	2.47	0.10	0.60	0.93

that show the dynamic model offering an improvement, but could be due to the fact that the AR matrix employed in this work is diagonal, while the system in [64] uses a full AR matrix to capture across-frequency correlation. The variation with the greatest improvement over the basic algorithm is the STSA-Phase. Incorporating the phase information yields improved scores for all cases except the high SNR fluctuating noise situations. This exception may be due to high SNR time/frequency regions that occur when the fluctuating noise level is low. In these intervals the in-phase assumption of the basic algorithm is sufficiently accurate and is not improved by the phase estimation.

The STSA-IMM also offers a moderate but consistent improvement over the basic algorithm. The improvement is likely due to the slightly higher noise attenuation that the multiple model approach provides during speech to non-speech transitions. The behavior of the multiple models can be seen in Fig. 6.5, which compares the frequency averaged active speech mode probabilities for two different Markov switching probabilities, $p_{ii} = 0.5$ and 0.9. The spectral amplitudes of clean speech averaged across all frequency bins are also included for reference. When the switching probabilities are $p_{ii} = 0.5$ the mode probabilities are equal to the mode likelihoods, while setting $p_{ii} = 0.9$ smooths the fluctuations. Note that since the parameters of the active speech model are estimated online, they decay toward zero during silence periods, resulting in both speech and non-speech modes having the same conditional likelihood. As such the multiple model framework does not provide continuous noise suppression during speech pauses, rather it provides additional attenuation at speech to non-speech transitions and allocates more particles to the speech onset model during transitions and onsets.

The objective measures are not able to evaluate the naturalness of the enhanced speech, however informal listening tests confirm that the speech enhanced by the

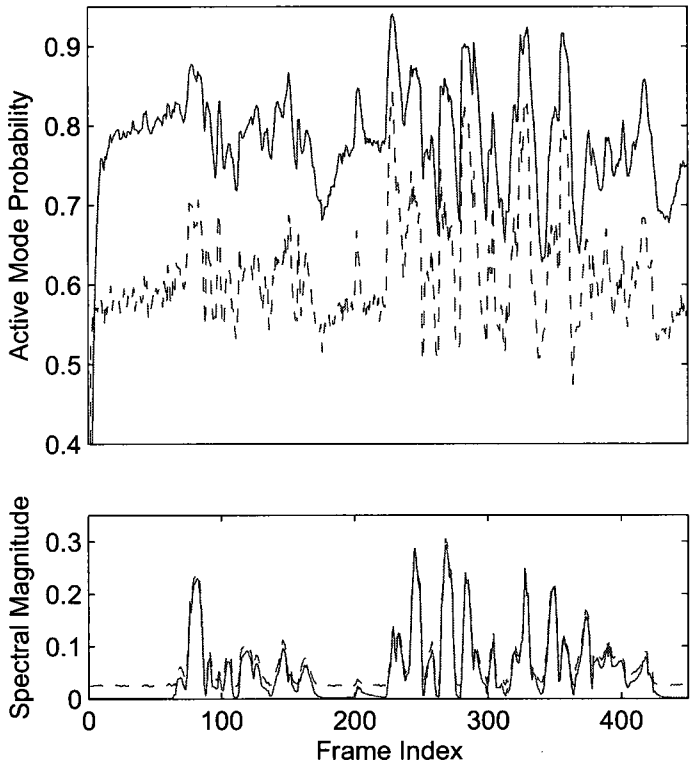


Figure 6.5: Active mode probability for $p_{ii} = 0.9$ (solid) and $p_{ii} = 0.5$ (dashed), clean speech spectral amplitudes averaged across all frequency bins (dotted).

STSA-PF algorithms is free from glitches. As reported in [66] the RBPF enhanced speech exhibits residual noise that preserves the character of the original noise, with an energy envelope that is modulated by the signal energy. In the DCT-RBPF as well as the STSA-PF case, this modulation occurs within each band, therefore the character of the noise is modified; however, both the DCT-RBPF and the STSA-PF approaches exhibit very low level musical noise artifacts. The ability of the STSA-PF algorithms to provide high levels of noise reduction without imposing speech distortion or introducing significant musical noise is attributed to the dynamic model imposed at the sampling step. With both the random walk and AR models, the probability density for the spectral amplitude at time n is centered at the estimated value for time $n-1$. As a result, new candidate spectral amplitudes will tend to fall in a near neighborhood of the previous frame value, encouraging smoothness in the spectral amplitude evolution and reducing spurious amplitude fluctuations that lead to musical noise. To see the influence of the dynamic model on musical noise, Fig. 6.6 shows the spectrograms of a segment of clean and 15 dB WGN corrupted noisy speech, and Fig. 6.7 shows the results of enhancement using: the MMSE-STSA algorithm [12] (decision-directed smoothing factor $\alpha = 0.9$); the STSA-PF algorithm with an AR(1) parameter of zero, ie. applying no dynamic model, assuming successive frames are independent; and the STSA-PF with a random walk on the spectral amplitudes. When the STSA-PF is used without a dynamic model the excitation smoothing provides some musical noise protection, however minor artifacts can be observed, appearing as isolated peaks in the high frequency region. When the random walk model is applied the additional smoothing of the spectral amplitude tracks provided by the dynamic model results in minimal artifacts, even comparing favorably to the MMSE-STSA which is known for its good musical noise performance [8].

Figure 6.6: Spectrograms of clean speech signal (top) and 15 dB WGN corrupted noisy signal.

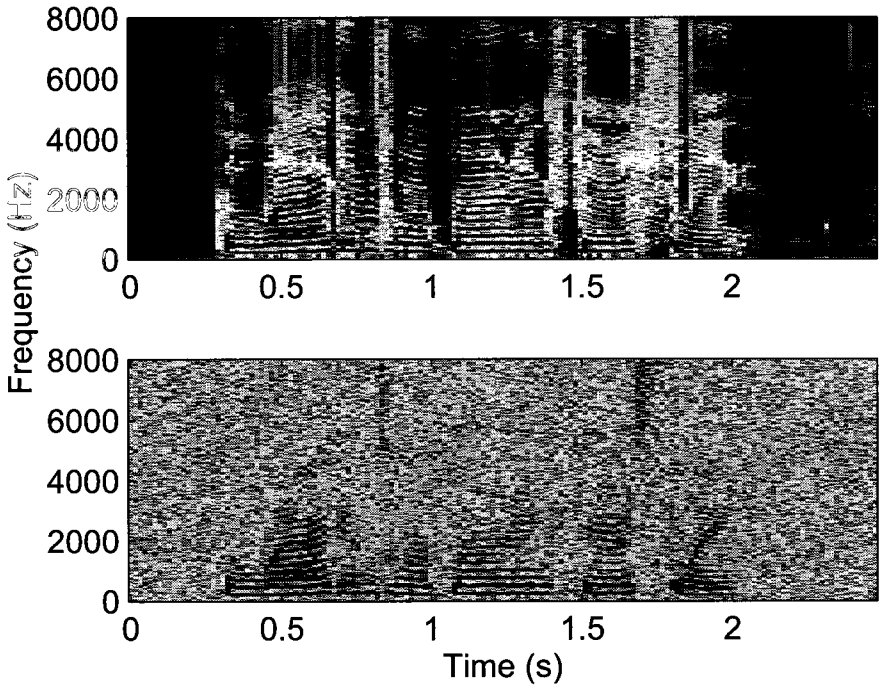
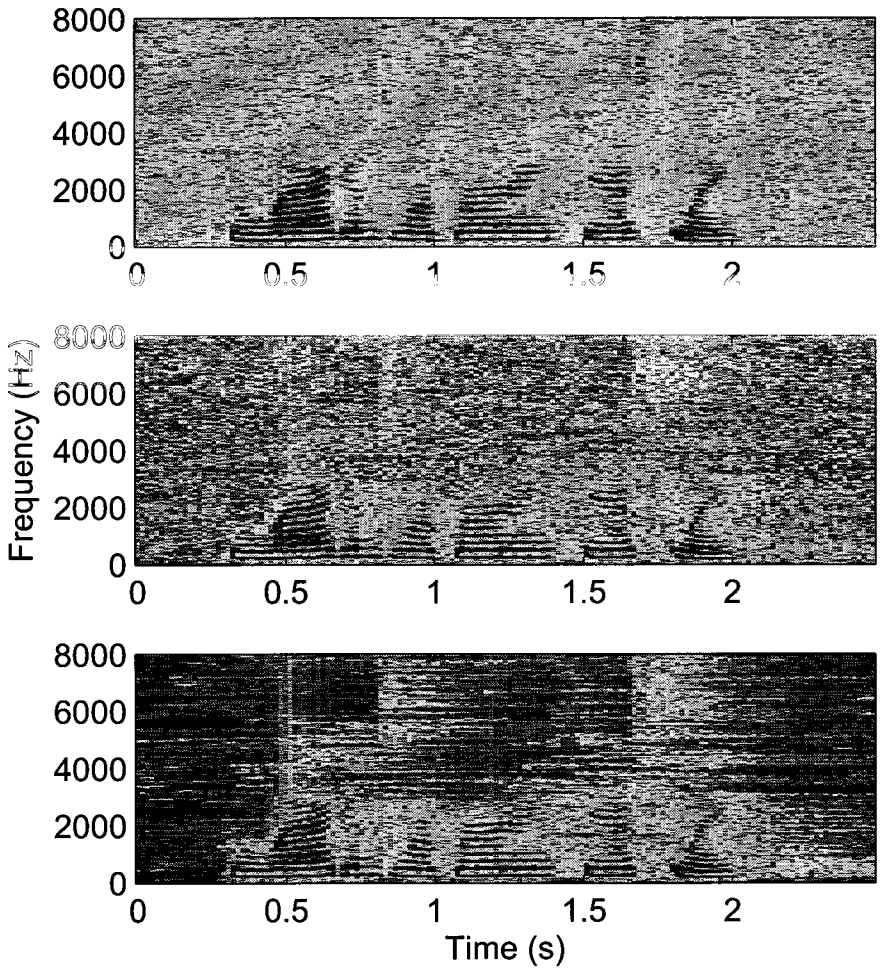


Figure 6.7: Spectrograms of signals from Fig. 6.6 enhanced by: MMSE-STSA (top), STSA-PF with no dynamic model (middle) and STSA-PF with random walk (bottom).



6.3.2 Complexity Comparison

Providing exact computational complexity figures for particle filter algorithms is complicated by implementation-specific factors such as the random number generation algorithm, the choice of sampling technique (inversion versus rejection sampling) and the evaluation of exotic or complex functions. Consequently, rather than exploring the complexity of each algorithm variant in detail, in the following we discuss the general steps that need to be performed and the differences in how each variant computes those steps, then we examine the impact of these differences through measured execution times of the simulated algorithm variants.

Importance Sampling

To perform importance sampling the STSA-PF first requires updating the excitation parameter to define the unconstrained sampling density for all particles. Generation of each particle then requires computing the bounds of constrained Laplace draw using (6.10), generating a uniform random number, and transforming it using (6.11). When the fixed AR(1) model is used in place of the random walk, computing the boundaries for the constrained draw in (6.10) and propagating the particles through (6.3) each require an additional multiplication per particle. When the dynamic AR(1) model is used, a new set of AR parameters must be generated according to (6.26), and the Laplace excitation parameter must be computed for each particle.

The importance sampling for the STSA-IMM involves updating a single set of mode probabilities for all particles, using those probabilities to distribute the particles, then sampling from the active speech and speech onset models. Sampling from the active speech model is the same as the STSA-PF. Sampling from the speech onset model is less computationally intensive as the boundaries of the constrained prior are

the same for all particles, so particle generation reduces to transforming a uniform draw according to (6.39).

The STSA-Phase variant samples the spectral amplitudes in the same manner as the STSA-PF or STSA-IMM, and also requires generating an additional uniform random variable on $[-\pi, \pi]$ for the phase difference estimate.

Importance Weighting and Resampling

The particle importance weighting for the STSA-PF and the STSA-IMM requires computing the likelihood according to (6.14), regardless of whether the random walk or the AR model are used. In contrast the STSA-Phase uses the more complex likelihood expression given by (6.36). Resampling using the set of particle weights is the same for all variants, and the complexity of common algorithms, such as systematic resampling, is linear in the number of particles [57].

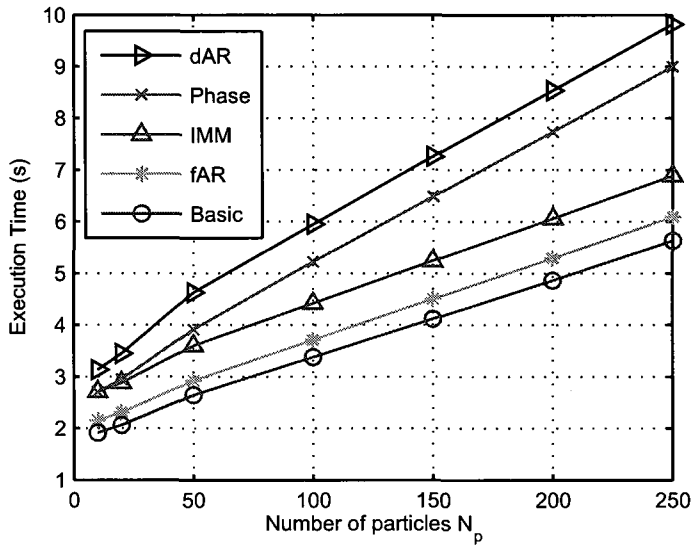
Enhanced Output

The enhanced output for all variants requires computing the average of the resampled candidate spectral amplitudes and combining it with the noisy phase. The STSA-IMM additionally requires updating the posterior mode probabilities and using the active speech probability to scale the enhanced output.

Execution Time

Fig. 6.8 compares the average time over 20 trials for the the algorithm variants to enhance 2 seconds of speech sampled at 16 kHz. The simulations were performed using MATLAB 2007a implementations of the algorithms running on a Linux PC with an Intel Core 2 Duo 2.2 GHz CPU and 4 GB RAM. From the figure it is clear that the performance improvements of the STSA-IMM and STSA-Phase variants come

Figure 6.8: Execution time comparison for particle filter variants processing 2 seconds of speech sampled at 16 kHz.



at a computational cost. However, while most of the additional computations for the STSA-IMM are fixed overhead related to maintaining the mode probabilities, the more complex sampling and weight evaluation steps of the STSA-Phase require significantly more computations per particle. As the number of particles increases the gap between the STSA-IMM and the basic STSA-PF remains consistent, while the gap between the STSA-Phase and the STSA-PF grows, with the STSA-Phase requiring approximately 50% more execution time. Fig. 6.8 also demonstrates that the basic algorithm using a random walk and the fixed-AR version both run at or near real-time for 20 and fewer particles; this number would be expected to increase with an optimized compiled implementation. Comparing Fig. 6.8 to Fig. 5.5, note that the Kalman filter iterations of the RBPF algorithms result in significantly increased running time compared to the STSA-PF.

6.4 Summary

This chapter presented several spectral amplitude speech enhancement estimators that use the particle filter framework to incorporate dynamic models that would make closed-form solutions intractable. This allows us to account for the strong inter-frame correlation and non-Gaussian statistics observed in speech spectral amplitudes. The basic algorithm models the spectral amplitudes as an AR process with Laplace distributed excitation; two variants of the standard algorithm are also presented. The first uses an interacting multiple model approach to account for transitions between active speech and silence intervals, the second accounts for phase differences between the clean speech and noise DFT coefficients. Experiments using wideband speech and real recorded noise demonstrate that modeling the spectral amplitude evolution allows the proposed algorithm to effectively manage the trade-off between speech distortion and sound quality as measured using objective metrics. The enhanced speech is also natural sounding with very low levels of the musical noise artifacts common in spectral amplitude enhancement. Compared to the basic algorithm, the multiple model variant is found to offer an improvement in noise reduction during transitions, while phase estimation is found to be most beneficial in stationary noise and at low SNRs. The proposed algorithm offers comparable performance at a lower complexity than the DCT-RBPF presented in Chapter 5.

Chapter 7

Binaural Coherence-Assisted Speech Enhancement

The speech enhancement approaches outlined in 5 and 6 are monaural algorithms suitable for single channel systems and bilateral hearing aids. Future hearing aids will contain wireless links enabling binaural processing [50]. Binaural hearing enables human listeners to exploit spatial information to assist in hearing in noisy situations. When multiple competing sound sources are present, spatial cues allow normal-hearing listeners to separate the sources and concentrate on the desired signal. In addition to the time, frequency and speech production model information exploited by monaural algorithms, binaural systems have access to the spatial dimension. By mimicking the processing performed by the human auditory system, binaural hearing aids have the potential to offer increased noise reduction and lower speech distortion. Working as a co-operative system binaural hearing aids can achieve this noise reduction while preserving spatial cues, allowing the human auditory system to carry out its own processing. This chapter presents a binaural noise reduction system for use in a diffuse noise environment with a near-frontal target speaker. Knowledge of the

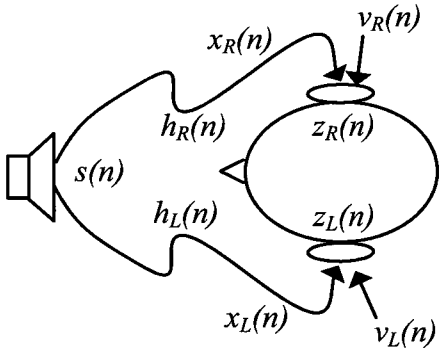


Figure 7.1: Binaural hearing aid speech and noise model.

binaural noise field coherence to is used to derive estimators to extract the clean signal and noise parameters required by the spectral amplitude particle filter of Chapter 6; producing a binaural model-based spectral amplitude speech enhancement system.

7.1 System Setup and Signal Model

Fig. 7.1 presents the binaural system setup. A source signal $s(n)$ is filtered by left and right transfer functions $h_L(n)$ and $h_R(n)$ to give the clean left and right signals $x_L(n)$ and $x_R(n)$. These signals are mixed with independent additive noise yielding the left and right measurement signals $z_L(n)$ and $z_R(n)$. The objective of the binaural speech enhancement task is to recover the clean left and right signals $x_L(n)$ and $x_R(n)$, preserving their spatial information. The system assumes a user with binaural hearing aids with one channel per ear; which could represent two single microphone hearing aids or the outputs of a beamforming front-end. The noise is assumed to be diffuse, having equal power in both channels and a random phase. Many typical noises encountered in speech enhancement can be considered diffuse, such as multi-talker babble and office noise.

7.2 Interaural Coherence

Interaural correlation provides spatial information that can be exploited by both human listeners and speech enhancement algorithms to increase intelligibility in noisy environments. The cross-PSD of two random signals x and y at time n is defined as:

$$\Phi_{xy}(n) = \mathbb{E}\{X(n)Y^*(n)\} \tag{7.1}$$

where the expectation is taken over all realizations of the signals. Assuming a linear transfer function between the signal source and microphone, and assuming independence of the source and the additive noise signals, the auto- and cross- PSDs of the left and right channel measurement signals are:

$$\Phi_{z_L z_L} = \Phi_{x_L x_L} + \Phi_{v_L v_L} \tag{7.2}$$

$$\Phi_{z_R z_R} = \Phi_{x_R x_R} + \Phi_{v_R v_R} \tag{7.3}$$

$$\begin{aligned}\Phi_{z_L z_R} &= \Phi_{ss}|H_L||H_R|e^{j(\angle H_L - \angle H_R)} + \Phi_{v_L v_R} \\ &= \sqrt{\Phi_{x_L x_L}\Phi_{x_R x_R}}e^{j\theta_{x_L x_R}} + \Phi_{v_L v_R}\end{aligned} \tag{7.4}$$

where $\theta_{x_L x_R}$ is the phase difference between the left and right channels of the target speaker.

In practice the signals are assumed ergodic, and the spectra are estimated using first-order recursive smoothing of the periodogram spectral estimates:

$$\Phi_{z_i z_j}(n) \approx \alpha\Phi_{z_i z_j}(n-1) + (1-\alpha)Z_i(n)Z_j^*(n) \tag{7.5}$$

where $i, j \in \{L, R\}$. The choice of the smoothing parameter α, $0 \leq \alpha \leq 1$ presents a trade-off: larger values of α provide a lower variance spectral estimate at the expense

of capturing the time-varying characteristics of the signal.

A measure of the correlation between the left and right channels is given by the complex coherence $\rho_{x_L x_R}$, and the magnitude-squared coherence (MSC), $C_{x_L x_R}$:

$$\rho_{z_L z_R} = \frac{\Phi_{z_L z_R}}{\sqrt{\Phi_{z_L z_L} \Phi_{z_R z_R}}} \tag{7.6}$$

$$C_{z_L z_R} = |\rho_{z_L z_R}|^2 \tag{7.7}$$

In [98] experiments are carried out to determine the interaural coherence of a diffuse noise field. While the coherence of a diffuse noise field in the free-field can be solved theoretically, the diffraction of the head complicates the interaural case. A model assuming a spherical head was found to be insufficient to capture the effects of the irregular head shape, instead curve fitting was performed on the measurement data. The resulting model for diffuse noise interaural coherence with inter-microphone spacing d and speed of sound ν is:

$$\rho_{v_L v_R} = \frac{\operatorname{sinc}\left(\gamma \frac{2\pi f d}{\nu}\right)}{\sqrt{1 + (\beta \frac{2\pi f d}{\nu})^4}} \tag{7.8}$$

where values of $\gamma = 2.2$ and $\beta = 0.5$ were found to provide a good fit for a 15 cm head diameter. The function is real, and its first zero occurs around 500 Hz; at frequencies above this value there is no appreciable interaural coherence, and below the function rises continuously towards unity. Head shadowing reduces the low-frequency region of high coherence; in the free-field case, where $\gamma = 1$ and $\beta = 0$, the first zero occurs around 1.1 kHz. Fig. 7.2 compares the measured MSC of cafeteria babble to the MSC predicted by (7.8), demonstrating that the model is a very good fit to the dummy head measurement data. The additional peaks in the cafeteria babble coherence are attributed to directional noise sources such as clattering plates and nearby talkers.

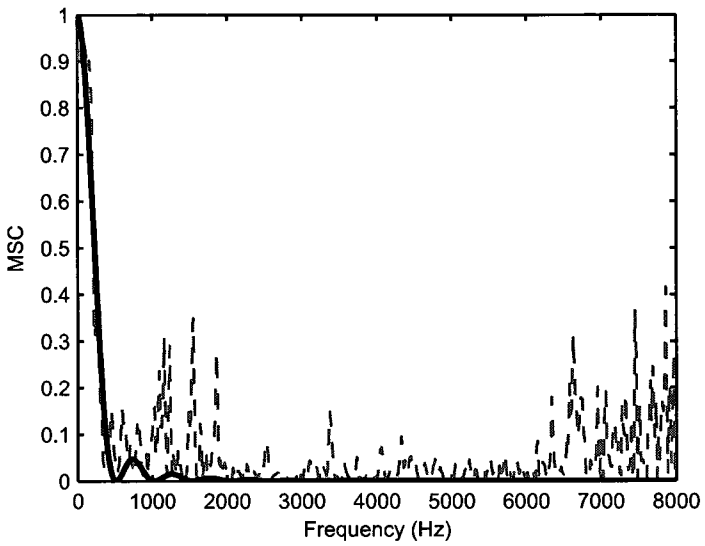


Figure 7.2: Measured MSC of cafeteria babble (dash) compared to theoretical MSC (thick solid).

Unlike diffuse noise, directional targets such as speech have a constant phase delay between the channels, so the MSC will approach unity across all frequencies, even if the target is non-frontal. Fig. 7.3 plots the interaural MSC for 32 seconds of male speech recorded using a dummy head in a cafeteria at a distance of 150 cm and azimuths of 0°, 30° and 90°. Note that reverberation and ambient noise result in MSC values below unity for all azimuths, and for non-frontal targets head shadowing further decreases the MSC in high frequency regions. However, even in the 90° azimuth case the speech MSC is significantly higher than the diffuse noise MSC in Fig. 7.2.

7.2.1 Coherence-Based Speech Enhancement

Knowledge of the noise field coherence has previously been exploited for speech enhancement in multi-microphone environments, however, other than the noise estimation algorithm in [55], the target application is general microphone array processing,

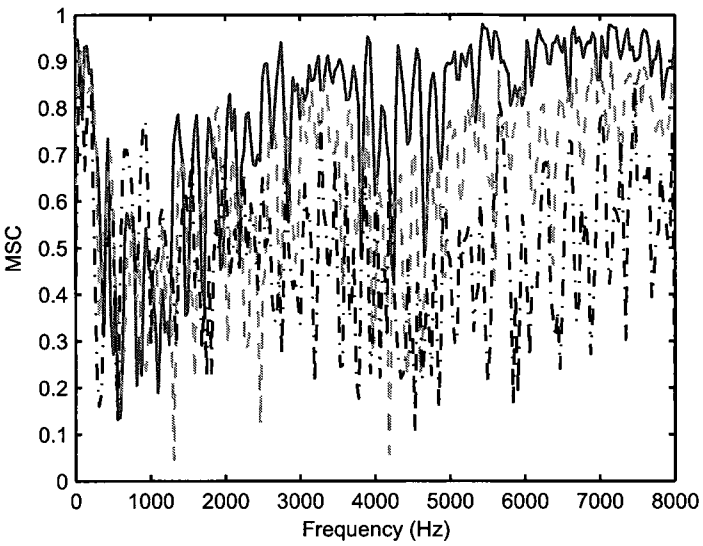


Figure 7.3: Measured MSC of male speech in a cafeteria at 0° azimuth (solid), 30° azimuth (dash), and 90° azimuth (dash-dot).

not a binaural systems. In [54] a diffuse noise field microphone array Wiener post-filter is derived. It is noted that when the noise coherence is known (7.6) can be combined with (7.2)–(7.4) to yield a set of four equations with four unknowns: the left and right speech and auto-PSDs, the speech cross-PSD and the noise auto-PSD. Since the signals in a microphone array are coherent and time-aligned, the cross-correlation of the desired signal between any two channels i and j is real, and equal to the clean signal power spectra:

$$\Phi_{x_i x_i} = \Phi_{x_j x_j} = \Phi_{x_i x_j} \tag{7.9}$$

this assumption reduces the number of variables, resulting in an over-determined system, arithmetic averaging of the solutions provides an estimate of the resulting

Wiener filter from each microphone pair (i, j):

$$\Phi_{xx} = \frac{\Re\{\Phi_{z_i z_j}\} - 0.5\rho_{v_i v_j}(\Phi_{z_i z_i} + \Phi_{z_j z_j})}{1 - \rho_{v_i v_j}} \tag{7.10}$$

$$H = \frac{\Phi_{xx}}{0.5(\Phi_{z_i z_i} + \Phi_{z_j z_j})} \tag{7.11}$$

This post-filter is shown to offer a significant improvement in noise reduction performance compared to a comparable post-filter assuming incoherent noise.

7.3 Estimator Derivation

In this section we use the framework of [54] to derive estimators for the Wiener gain functions that enable separation of the speech and noise spectra for use binaural hearing aid applications. Viewing the PSDs as vectors in the complex plane, geometric arguments and spatial knowledge can exploited to estimate the Wiener filter gain without external noise estimation or the use of a VAD. The target speaker is assumed to approximately frontally located, and the noise is assumed diffuse. These assumptions are satisfied in practice in situations as diverse as multi-talker cocktail party environments and watching television in the presence of a home ventilation system.

7.3.1 Binaural Wiener Gain

Frequency domain Wiener filtering [31] is a standard low-complexity method of speech enhancement. Assuming stationary and jointly Gaussian real and imaginary components of the speech and noise transform coefficients, the Wiener filter provides the

minimum mean-square error (MMSE) estimates of X and V:

$$H_W = \frac{\Phi_{xx}}{\Phi_{xx} + \Phi_{vv}} \tag{7.12}$$

$$\hat{X} = H_W Z \tag{7.13}$$

$$\hat{V} = (1 - H_W) Z \tag{7.14}$$

While theoretically optimal, computation of the Wiener gain in (7.12) requires the clean signal spectrum Φ_{xx} which is not available in practice. Existing approaches, including the binaural Wiener filter proposed in [48], estimate the clean signal spectrum by compensating the measurement signal spectrum with an estimate of the noise. This approach breaks down in fluctuating noise and can lead to musical noise and speech distortion even when the noise is stationary. If the noise-field coherence is known (7.6) and (7.2)–(7.4), hereafter referred to as the measurement equations, can be manipulated to solve for the clean signal PSDs without direct compensation.

In the ideal frontal target case $H_L = H_R$, and the speech auto- and cross-correlations are all real and in phase, ie. $\Phi_{x_L x_L} = \Phi_{x_R x_R} = \Phi_{x_L x_R} = \Phi_{xx}$. For isotropic noise the power spectrum in both channels is also the same, $\Phi_{v_L v_L} = \Phi_{v_R v_R} = \Phi_{vv}$. This reduces the number of variables in the measurement equations, resulting in an over-determined system, and the estimator in (7.11). However, in a binaural hearing aid application, factors such as talker and listener head movement and differences between the left and right ear shapes mean that the left and right signals will never be exactly time-aligned, even for a frontal target, so the clean speech cross-correlation will have an imaginary component. As a result the real part of the measurement cross-correlation in (7.10), and consequently the gain computed according to (7.11), can become negative, producing an invalid enhanced signal spectrum. Half or full-wave

rectification can be used to ensure a valid spectrum, but this discards any contribution of the imaginary component in the measurement cross-correlation. In diffuse and incoherent noise fields, the noise cross-correlation is real, so this imaginary component in the measurement noise cross-correlation is due to the phase-shift between the clean signal HRTFs, and discarding it can lead to under-estimation of the target speech PSD and distortion of the desired signal.

To account for minor time mis-alignment between the left and right clean signals, we make the following approximation for frontal and near-frontal targets in a diffuse noise environment:

$$|\Phi_{z_L z_R}| \approx |\Phi_{x_L x_R}| + |\Phi_{v_L v_R}| \tag{7.15}$$

This approximation is supported as follows. The squared magnitude of the measurement cross-correlation is:

$$|\Phi_{z_L z_R}|^2 = |\Phi_{x_L x_R}|^2 + |\Phi_{v_L v_R}|^2 + 2\Re\{\Phi_{x_L x_R}\Phi_{v_L v_R}\} \tag{7.16}$$

Under the diffuse noise assumption the noise cross-correlation is real and (7.16) becomes:

$$\begin{aligned} |\Phi_{z_L z_R}|^2 &= |\Phi_{x_L x_R}|^2 + |\Phi_{v_L v_R}|^2 + \\ &\qquad 2|\Phi_{x_L x_R}|\cos(\theta_{x_L x_R})|\Phi_{v_L v_R}| \qquad (7.17) \\ &= (|\Phi_{x_L x_R}| + |\Phi_{v_L v_R}|)^2 + \\ &\qquad 2|\Phi_{x_L x_R}||\Phi_{v_L v_R}|(\cos(\theta_{x_L x_R}) - 1) \qquad (7.18) \end{aligned}$$

The first term on the right hand side of (7.18) corresponds to the proposed approximation in (7.15), while the second term is the approximation error.

To estimate the approximation error we assume the left and right HRTFs have a

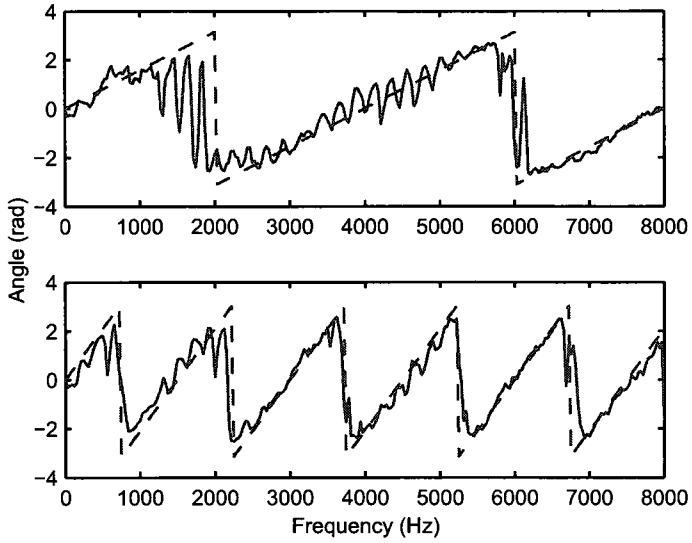


Figure 7.4: Cross-correlation phase measured (solid) and predicted by (7.19) (dashed) for 30° (top) and 90° (bottom) azimuth source locations.

constant time-difference of δ, giving the cross-correlation a linear phase:

$$\theta_{x_L x_R} = 2\pi f \delta \tag{7.19}$$

While this is a simplified approximation which ignores frequency dependent head shadowing and pinna effects, its use in this case is empirically supported by Fig. 7.4, which plots the cross-correlation phase predicted by (7.19) against the value obtained from recorded speech data. The measurement data is from 30 seconds of speech spoken by two male and two female speakers recorded 1 m from a dummy head at 30° and 90° azimuth; while the measured ITDs for the two sets of signals, $\delta_{30°} = 250$ μs and $\delta_{90°} = 667$ μs, were used in (7.19) to plot the model data. The results demonstrate that the constant time delay model is able to capture the average variation of the cross-correlation phase with frequency.

If the target is near-frontal the time delay between the channel signals is small

and the inter-channel phase, $\theta_{x_L x_R} = 2\pi f \delta$, grows slowly with frequency. In this case at low frequencies, where the coherence of diffuse noise is high, we can apply the small angle approximation $\cos(2\pi f d) \approx 1$ and the approximation error term in (7.18) approaches zero. As frequency increases the impact of the phase becomes greater; however at the same time the cross-correlation of diffuse noise, as described by (7.8), is continually decreasing. As a result, at high frequencies $\Phi_{z_L z_R} \approx \Phi_{x_L x_R}$ and the approximation in (7.15) holds as well.

Figs. 7.5 – 7.7 plot the root mean squared error of the approximation in (7.15) as a function of frequency for 30 seconds of speech recorded at azimuths of 0°, 30° and 90°. The average cross-correlation of the measurement signals is also shown, along with the approximation error resulting from assumption in [54], ie. that the speech is perfectly in phase so the cross-correlation is entirely real. The error in this case is the imaginary part of the measurement cross-correlation:

$$|\Phi_{z_L z_R} - \Re\{\Phi_{x_L x_R} + \Phi_{v_L v_R}\}| = |\Im\{\Phi_{z_L z_R}\}| \tag{7.20}$$

Using the model of (7.19) with diffuse noise, the imaginary part is given by:

$$\Im\{\Phi_{z_L z_R}\} = |\Phi_{x_L x_R}| \sin(2\pi f \delta) \tag{7.21}$$

This gives an approximation error that is modulated, with a period determined by the delay between the left and right signals. This modulation can be seen in Figs. 7.6 and 7.7 where the peaks and troughs correspond to the angle variations in 7.4. When $\sin(2\pi f \delta) = 1$ the cross-correlation is completely imaginary and the error of the approximation approaches the measured signal PSD. In contrast the proposed approximation in (7.15) is less frequency dependent, and is much more robust to phase differences between the measurements.

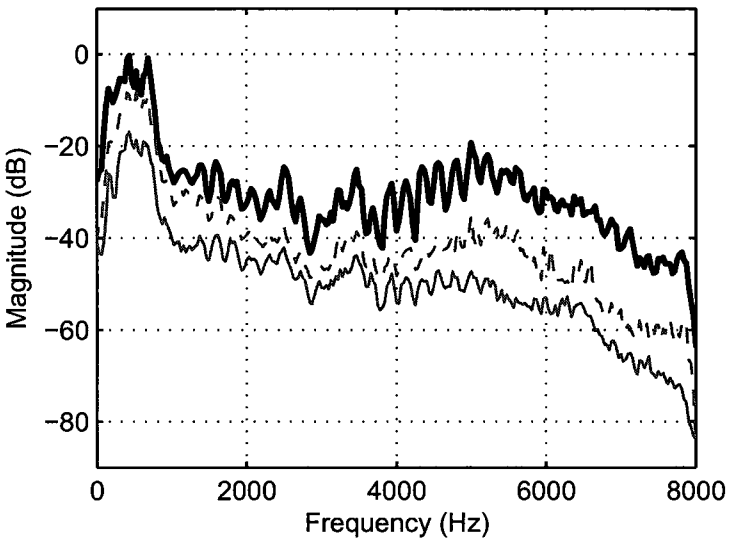


Figure 7.5: Measurement PSD (thick solid), approximation error for proposed model (thin solid) and approximation error for real cross-correlation assumption (dashed); 0° target.

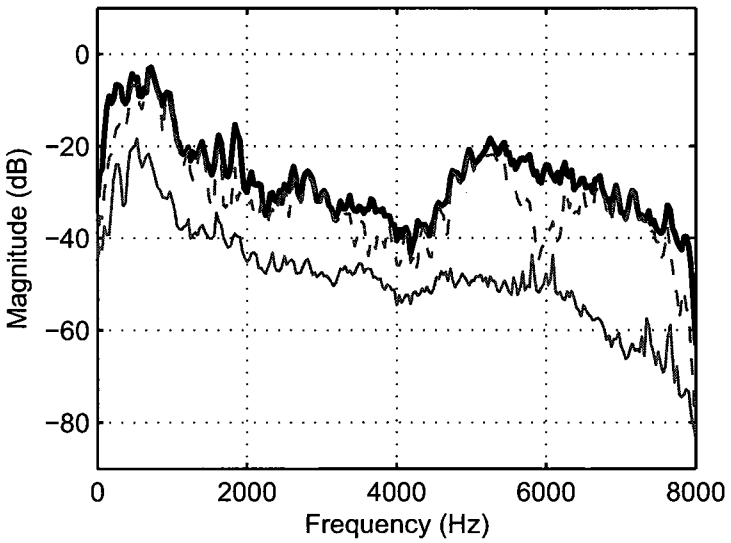


Figure 7.6: Measurement PSD (thick solid), approximation error for proposed model (thin solid) and approximation error for real cross-correlation assumption (dashed); 30° target.

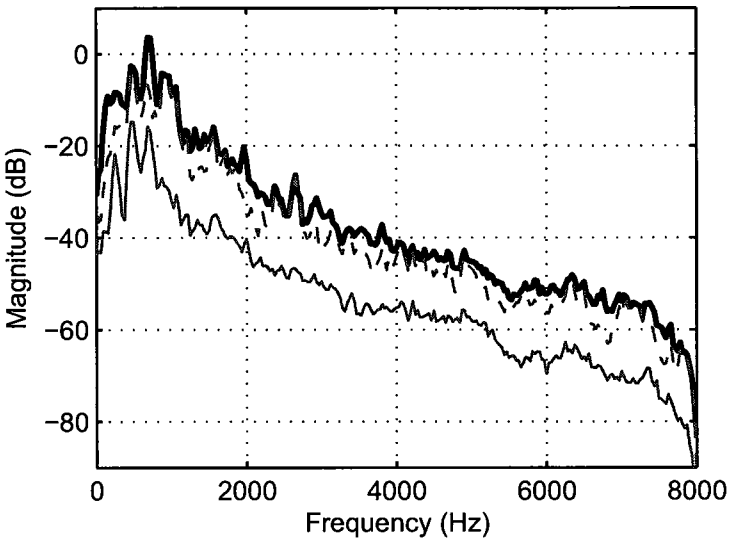


Figure 7.7: Measurement PSD (thick solid), approximation error for proposed model (thin solid) and approximation error for real cross-correlation assumption (dashed); 90° target.

Employing the approximation in (7.15) gives the measurement equations:

$$\Phi_{z_L z_L} = \Phi_{x_L x_L} + \Phi_{vv} \tag{7.22}$$

$$\Phi_{z_R z_R} = \Phi_{x_R x_R} + \Phi_{vv} \tag{7.23}$$

$$|\Phi_{z_L z_R}| = \sqrt{|\Phi_{x_L x_L}||\Phi_{x_R x_R}|} + \rho_{v_L v_R}\Phi_{vv} \tag{7.24}$$

Solving results in the following estimates for the left and right clean signal PSDs:

$$\Phi_{x_L x_L} = \frac{|\Phi_{z_L z_R}| - \rho_{v_L v_R}\Phi_{z_L z_L}}{1 - \rho_{v_L v_R}} \tag{7.25}$$

$$\Phi_{x_R x_R} = \frac{|\Phi_{z_L z_R}| - \rho_{v_L v_R}\Phi_{z_R z_R}}{1 - \rho_{v_L v_R}} \tag{7.26}$$

While the PSD estimates in (7.25) and (7.26) could be used directly to compute the left and right Wiener filter gains, the gains can be intelligently combined to

achieve a more robust solution. The approach used to combine the channel gains will affect the output signal. For example, applying the minimum of the two gains to both channels will result in maximal noise attenuation but may increase speech distortion. We propose to use the geometric average to combine the two signals. While their phases differ, the left and right near-frontal target transfer function magnitudes - and consequently the left and right clean signal PSDs - are approximately equal and averaging can achieve a lower variance gain estimate. In [54] the arithmetic average is used, however the geometric mean is equivalent to arithmetic averaging of the log-spectra, which is desirable since perceived loudness increases logarithmically. The geometric mean also provides more attenuation when one of the gains is small, thereby suppressing inter-speech residual noise and musical tones.

Geometric averaging of the left and right Wiener filters corresponding to the clean signal estimates in (7.25) and (7.26) gives the proposed estimator for near-frontal targets in isotropic noise:

$$H_{w,n} = \frac{|\Phi_{z_L z_R}| - \rho_{v_L v_R}\sqrt{\Phi_{z_L z_L}\Phi_{z_R z_R}}}{(1-\rho_{v_L v_R})\sqrt{\Phi_{z_L z_L}\Phi_{z_R z_R}}}. \tag{7.27}$$

7.4 Computational Complexity

The ability to work without an external VAD makes the proposed estimator very computationally lightweight. For each frequency bin in each frame the following computations are required:

- Auto PSD: 2 complex magnitude, 2 real multiply and 1 real add each for the left and right PSD estimates in (7.2) and (7.3) using (7.5).
- Cross PSD: 1 complex-complex multiply, 2 real-complex multiply, and 1 complex add to update (7.4) using (7.5).

- Gain: 1 absolute value, 2 real multiply, 1 square root, 2 subtract and 1 divide to compute (7.27).

7.5 Robustness Evaluation

A series of trials was carried out to compare the robustness of the proposed near-frontal target estimator and the frontal-target Wiener filter estimator from [54]. Recordings with targets placed at 0°, 30° and 90° azimuth were used to simulate talker and listener head movement. A smoothing parameter of $\alpha = 0.9$ was used to compute all of the required auto and cross-PSDs; for a sampling rate of 16 kHz, a frame length of 512 samples and an overlap of 50%, this corresponds to a time constant of approximately 150 ms. The estimators were tested using approximately 13 seconds of speech containing utterances spoken by two female and two male speakers from the TIMIT database. The diffuse noise source was multi-talker babble. Since multi-talker babble is composed of speech, its variable power envelope and spectral shape are the same as speech, making its removal a challenging task for noise reduction systems.

Tables 7.1 – 7.3 compare the objective scores for the different estimators. Since the measures are both monaural, the left and right channel scores, averaged over all files, are both shown. While both estimators perform well for the 0° case, when the target azimuth is increased to 30° and 90°, the frontal estimator treats the target speech as noise, severely distorting the signal and degrading the objective measure scores. In contrast, the near-frontal estimator degrades more gracefully in the non-frontal target cases. Even at 90° azimuth, the proposed estimator provides good attenuation of diffuse noise with low speech distortion.

Table 7.1: Robustness evaluation objective measure scores for cafeteria noise, 0° target (left/right).

	0 dB	10 dB	20 dB
WPESQ Score			
Noisy	1.08/1.08	1.59/1.56	3.04/2.97
Frontal	1.18/1.19	2.03/2.00	3.31/3.20
Near-frontal	1.18/1.18	1.94/1.88	3.27/3.18
CSII Score			
Noisy	0.16/0.15	0.67/0.65	0.96/0.96
Frontal	0.29/0.27	0.76/0.74	0.94/0.94
Near-frontal	0.23/0.21	0.75/0.73	0.94/0.94

Table 7.2: Robustness evaluation objective measure scores for cafeteria noise, 30° target (left/right).

	0 dB	10 dB	20 dB
WPESQ Score			
Noisy	1.10/1.06	1.73/1.45	3.21/2.76
Frontal	1.08/1.05	1.44/1.28	2.31/1.77
Near-frontal	1.21/1.13	2.04/1.74	3.37/3.01
CSII Score			
Noisy	0.23/0.09	0.76/0.52	0.97/0.93
Frontal	0.19/0.09	0.55/0.42	0.75/0.68
Near-frontal	0.29/0.13	0.78/0.63	0.94/0.90

Table 7.3: Robustness evaluation objective measure scores for cafeteria noise, 90° target (left/right).

	0 dB	10 dB	20 dB
WPESQ Score			
Noisy	1.10/1.05	1.71/1.40	3.17/2.59
Frontal	1.11/1.06	1.67/1.31	2.59/2.31
Near-frontal	1.19/1.11	1.99/1.61	3.09/2.65
CSII Score			
Noisy	0.24/0.09	0.77/0.50	0.97/0.93
Frontal	0.21/0.09	0.61/0.47	0.82/0.74
Near-frontal	0.27/0.11	0.78/0.59	0.94/0.88

7.6 Binaural Spectral Amplitude Particle Filtering

As the simulation results in section 7.5 demonstrate, the estimator in (7.27) could be used directly as a Wiener filter gain to provide noise attenuation. However, as the frame-by-frame nature of this estimation approach does not account for the time-correlation of speech spectral amplitudes, it is vulnerable to musical noise artifacts when the noise source is fluctuating. As an alternative, the instantaneous but high-variance estimates of the signal and noise spectra provided by the coherence-based Wiener filter can be used to estimate the parameters of the model-based spectral amplitude particle filter in Chapter 6. This produces a binaural spectral amplitude particle filter that is capable of providing high fluctuating noise attenuation and natural sounding processed speech. The results from section 6.3 show small performance gains when the more complex models are employed, so this section considers the basic spectral amplitude particle filter with a random walk on spectral amplitudes, and a

linear measurement model.

7.6.1 Parameter Estimation

Noise PSD

Traditional on-line noise estimation approaches like the minimum statistics algorithm [93] rely on long-term statistical information to separate speech and noise signals, resulting in an estimation lag when the noise power changes and degrading system performance in rapidly fluctuating noise. The gain in (7.27) uses spatial coherence rather than statistical information, so it can respond instantaneously to changes in the noise power. The gain $H_{w,n}$ in (7.27) is computed using smoothed PSD estimates; however it is used in (7.14) to provide an estimate of the noise present in the *current* frame. While spectral subtraction algorithms that treat the noise estimate as a direct measurement would require this noise estimate to be smoothed in order to reduce speech distortion and musical noise artifacts, the STSA-PF uses the estimate to parameterize the noise amplitude distribution; since it takes the variability of noise spectral amplitudes into account, the instantaneous noise estimate can be used without smoothing. The coherence-based diffuse noise estimator presented in [55] operates in a similar frame-by-frame manner, and is shown to provide effective tracking of highly non-stationary noises such as babble speech. The per-frame noise estimate given by the Wiener filter is taken as the mode of the noise distribution σ_{v_n} in each channel:

$$\sigma_{v_{i,n}} = |(1 - H_{w,n})Z_{i,n}| \qquad i \in \{L, R\} \tag{7.28}$$

Excitation parameters

The estimate of the spectral amplitude process prediction error, (6.15), in the single-channel STSA-PF uses the noisy amplitude measurement z_n, resulting in a biased estimate that must be compensated in (6.19) by subtracting an estimate of the noise power. In a binaural system the instantaneous estimates of $x_{L,n}$ and $x_{R,n}$ provided by (7.27) can be used in (6.20) in place of the noisy measurement to estimate the spectral amplitude process excitation:

$$e_{i,n} = H_{w,n} z_{i,n} - \hat{x}_{i,n-1} \tag{7.29}$$

$$\overline{e_i^2} = \beta \overline{e_i^2} + (1-\beta)|e_{i,n}|^2 \tag{7.30}$$

$$b_{d_{i,n}} = \sqrt{\frac{\overline{e_i^2}}{2}}. \tag{7.31}$$

where $i \in \{L, R\}$ is the left or right channel. Unlike the single channel estimate of (6.22), this excitation parameter estimate does not require the smoothed prediction error to be compensated with a spectral-subtraction approach.

7.6.2 Interaural Information Fusion

If the left and right channel enhancement were perfect, achieving complete noise removal without any speech distortion, using different levels of attenuation in each ear would not cause any undesired effects. However fluctuations of the random signals mean that ideal enhancement cannot be achieved, consequently there is always some residual noise and speech distortion. Informal listening reveals that when the attenuation levels are set independently, the level of this residual noise is not consistent between the channels, and fluctuations in the relative noise levels can result in a distracting effect whereby the noise source seems to move about the listener. Similarly,

fluctuations in the level of desired speech cancellation can distort ILD spatial cues that assist in localization. Furthermore, the release from masking provided by the ITD in noisy situations is mostly maintained in hearing impaired listeners [9], so an enhancement system seeking to improve or maintain intelligibility should avoid any degradation of ITD cues that would be caused by adjusting the inter-channel phase. In order to prevent the moving noise and distortion of spatial cues, binaural hearing aids should therefore ensure a consistent real-valued gain setting between the left and right channels. The binaural systems in [55] and [99] also use a single real gain to prevent cue distortion. To perform this type of combination, the left and right particle filter clean signal spectral estimates are first converted to spectral gains:

$$G_{i,n} = \frac{\hat{x}_{i,n}}{z_{i,n}} \qquad i \in \{L, R\} \tag{7.32}$$

the gains are then geometrically averaged:

$$G_n = \sqrt{G_{L,n} \cdot G_{R,n}}. \tag{7.33}$$

This gain is applied separately to each channel, producing a binaural output while preserving spatial cues. Fig. 7.8 presents a block diagram of a binaural noise reduction system incorporating a binaural estimator with a single real gain.

Other levels of binuaral information sharing were considered in addition to the gain averaging employed. Since the noise power is assumed the same in both channels, averaging the left and right noise estimates could provide a lower variance estimate; in practice this averaging degraded the performance of the instantaneous noise estimates provided by (7.28). Similarly, averaging the left and right channel excitation parameters from (7.31) was found to degrade performance for targets that were not perfectly frontal.

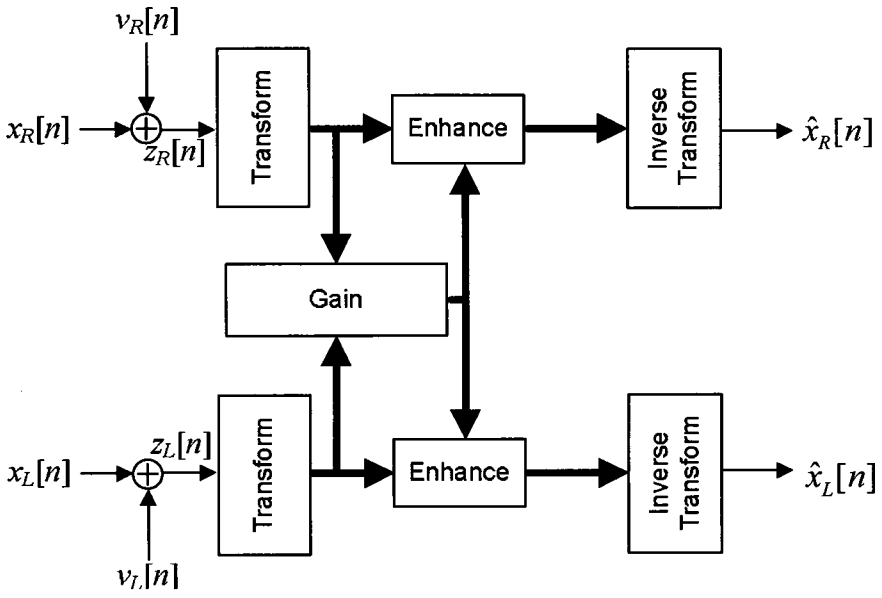


Figure 7.8: Binaural coherence-assisted speech enhancement.

7.7 Evaluation

The proposed coherence-assisted binaural particle filter was compared to: the basic STSA-PF from Chapter 6 operating bilaterally with ideal VAD-based noise estimation; and a standard Wiener filter system using minimum statistics noise estimation [93] and decision-directed SNR estimation with geometric averaging of the left and right channel gains. The results from section 7.5, directly applying the near-frontal target Wiener gain in (7.27), are also included for reference. The test files and parameters used are the same as those from the robustness evaluation, a smoothing parameter of $\alpha = 0.9$ was used to compute all of the required auto and cross-PSDs, and in the decision-directed SNR estimator.

Tables 7.4 – 7.6 compare the left and right channel objective scores for the different algorithms. Both binaural algorithms significantly outperform the standard Wiener filter in both noise reduction and speech intelligibility performance. The intelligibility improvement is most pronounced in low SNR conditions, where the benefit of spatial

information is greatest. Compared to the coherence-based Wiener filter, the binaural PF is able to achieve higher WPESQ scores, indicating higher levels of noise reduction, without degrading the CSII performance. The scores for the bilateral PF are higher than the binaural configuration in almost all cases. This is attributed to two factors: first, the ideal VAD based noise estimation used by the bilateral configuration estimates the coherent as well as non-coherent and diffuse noise, enabling it to target noises sources not considered by the diffuse noise binaural configuration; second, the requirement of the binaural system - that the left and right channels apply the same real gain - reduces the achievable level of noise reduction. However, informal listening reveals that the binaural PF has the most natural sounding output, with minimal fluctuations in the residual noise. In contrast the bilateral configuration exhibits more fluctuations both in the absolute noise level in each channel, and in the relative level of residual noise between the left and right channels.

7.8 Summary

This chapter presented a particle filter spectral amplitude speech enhancement system for use in future binaural hearing aids. The system assumes the target is approximately frontal and that the noise is diffuse, assumptions that are met in common hearing aid listening situations. In contrast to beamforming methods, the proposed algorithm does create a spatial gain pattern to attenuate noise from a specified direction. Instead, using an expression for the coherence of diffuse noise field that takes head-shadowing into account, left and right channel Wiener filter gains were derived to estimate the clean signal and noise spectra. These spectra are used to provide more reliable estimates of the noise and process excitation parameters required for the STSA-PF algorithm presented in Chapter 6. Since the spatial configuration of

	0 dB	10 dB	20 dB
WPESQ Score			
Noisy	1.08/1.08	1.59/1.56	3.04/2.97
Standard Wiener	1.09/1.08	1.67/1.63	3.15/3.08
Coherence Wiener	1.18/1.18	1.94/1.88	3.27/3.18
Bilateral PF	1.20/1.20	2.22/2.08	3.54/3.46
Binaural PF	1.22/1.22	2.03/1.96	3.36/3.25
CSII Score			
Noisy	0.16/0.15	0.67/0.65	0.96/0.96
Standard Wiener	0.18/0.16	0.71/0.67	0.96/0.96
Coherence Wiener	0.23/0.21	0.75/0.73	0.94/0.94
Bilateral PF	0.25/0.19	0.78/0.74	0.97/0.97
Binaural PF	0.23/0.21	0.75/0.73	0.94/0.93

Table 7.4: Objective scores for two channel speech enhancement, 0° azimuth target.

	0 dB	10 dB	20 dB
WPESQ Score			
Noisy	1.10/1.06	1.73/1.45	3.21/2.76
Standard Wiener	1.11/1.06	1.81/1.51	3.25/2.85
Coherence Wiener	1.21/1.13	2.04/1.74	3.37/3.01
Bilateral PF	1.32/1.14	2.41/1.91	3.61/3.31
Binaural PF	1.25/1.17	2.17/1.87	3.41/3.13
CSII Score			
Noisy	0.23/0.09	0.76/0.52	0.97/0.93
Standard Wiener	0.25/0.10	0.78/0.55	0.97/0.93
Coherence Wiener	0.29/0.13	0.78/0.63	0.94/0.90
Bilateral PF	0.36/0.12	0.84/0.65	0.97/0.93
Binaural PF	0.28/0.13	0.75/0.63	0.92/0.88

Table 7.5: Objective scores for two channel speech enhancement, 30° azimuth target.

	0 dB	10 dB	20 dB
WPESQ Score			
Noisy	1.10/1.05	1.71/1.40	3.17/2.59
Standard Wiener	1.12/1.06	1.81/1.46	3.29/2.74
Coherence Wiener	1.19/1.11	1.99/1.61	3.09/2.65
Bilateral PF	1.34/1.11	2.44/1.79	3.68/3.21
Binaural PF	1.26/1.17	2.17/1.79	3.44/3.07
CSII Score			
Noisy	0.24/0.09	0.77/0.50	0.97/0.93
Standard Wiener	0.25/0.10	0.79/0.54	0.97/0.93
Coherence Wiener	0.27/0.11	0.78/0.59	0.94/0.88
Bilateral PF	0.36/0.10	0.85/0.63	0.98/0.93
Binaural PF	0.27/0.12	0.76/0.61	0.92/0.87

Table 7.6: Objective scores for two channel speech enhancement, 90° azimuth target.

desired and interfering signal sources is exploited by the human auditory system to improve hearing in noise, the proposed system converts the left and right channel particle filter signal estimates into a single real-valued gain that is applied to the channels individually, producing a binaural output that preserves spatial cues. In experiments with hearing aid recordings of reverberant speech and multi-talker babble the proposed algorithm was shown to provide higher noise reduction and less speech distortion than a traditional Wiener filter with minimum statistics noise estimation and decision-directed SNR estimation. While the STSA-PF operating in a bilateral configuration achieved higher levels of noise reduction, the output of the binaural configuration was more natural, with minimal fluctuations in the residual noise level. While the algorithm is presented for a binaural configuration, it could be applied to any multi-microphone system where the inter-microphone noise coherence is known.

Chapter 8

Summary and Conclusions

The objective of this thesis was to apply transform-domain model-based algorithms to improve the performance of hearing aid speech enhancement algorithms processing wideband speech in non-stationary environments. The research was motivated by the observation that the potential benefit of hearing devices is highest in noisy situations, however low perceived benefit levels and unsatisfactory sound quality ratings indicate that this benefit is not realized in practice. Feedback and noise control were identified as two bottlenecks to hearing aid user satisfaction: the performance of existing algorithms is insufficient, especially in fluctuating noise, multiple talker environments, and changing acoustic environments. Transform-domain model-based approaches were proposed to account for these issues. Operating in the transform domain allowed for different processing strategies to be employed in different frequency ranges, taking into account the diverse frequency-dependent characteristics of wideband speech and noise. Incorporating models of speech, noise and the acoustic environment allowed the addition of time-domain constraints and a-priori knowledge to improve performance and ensure natural sounding outputs and reduced speech distortion. Statistical dynamic models accounted for speech and noise non-stationarity

to achieve higher levels of noise removal without imposing additional distortion or unnatural noise artifacts. All of the proposed algorithms were tested using real speech and noise signals recorded using commercially available hearing aid devices.

8.1 Summary of Contributions

The contributions of this thesis, and how they relate to the thesis objectives stated in Chapter 1 are summarized as follows:

1. *Subband acoustic feedback compensation:* A subband transform was added to acoustic feedback compensation systems using adaptive modeling of the feedback paths. Traditionally subband adaptive filtering is used to reduce the computational burden of large adaptive filters in applications such as acoustic echo cancellation. The short impulse responses of acoustic feedback paths mean that computational savings are not a motivating factor, so subband structures have not been widely investigated. However, experiments revealed that subband structures can improve the performance and robustness of continuous and non-continuous adaptation feedback compensation systems. The subband configuration was shown to exploit the time and frequency varying properties of changing wideband acoustic feedback paths to offer faster convergence and better tracking of abrupt and gradual feedback path changes. The ability to set the adaptation step-size individually in each band allows the designer more control over the trade-off between the large step-size required for fast convergence and tracking, and the small step size required to prevent system divergence, offering more uniform convergence with less cancellation of wideband speech inputs. This work was outlined in Chapter 4 and has been published in part in [17] and [18].

2. *DCT RBPF speech enhancement:* An AR model-based Rao-Blackwellized particle filter algorithm was proposed for the enhancement of speech DCT coefficients. The use of a dynamic model-based statistically motivated algorithm accounts for fluctuating speech and noise statistics, while the subband decomposition enables the algorithm to better enhance wideband speech. It was shown that the subband structure produces lower residual noise and less speech distortion than the equivalent fullband, both in practical applications and in the ideal case when the AR coefficients are measured directly from the clean speech signal. Smooth curve fitting between spectral peaks causes fullband AR models to over-estimate the signal power in spectral troughs, leading to high levels of intra-speech residual noise. The per-band fitting of the subband structure enables much better fitting of the deep troughs, high peaks, large spectral tilt and noise-like high frequency content of the wideband speech spectrum. In addition to the improved noise reduction performance, the DCT-RBPF has a significantly lower computational burden due to: frame-based processing at a decimated rate and smaller AR models in each band leading to efficient Kalman filter realizations and fewer particles to cover the state-space. The complexity can be further reduced by pairing the DCT-RBPF with a standard low-complexity algorithm: the harmonic content is confined to the lower bands, therefore most of performance gain can be achieved by applying the RBPF in these regions and a simple DCT domain Wiener filter algorithm in the perceptually less relevant upper bands. This work was described in Chapter 5 and has been published in part in [19] and has been submitted for publication [20].

3. *Spectral amplitude particle filter speech enhancement:* A particle filter algorithm

that incorporates a dynamic model into a traditional spectral amplitude estimation framework was proposed for spectral amplitude speech enhancement. Speech spectral amplitudes are modeled as either AR or random walk processes, with the process parameters estimated online. Since incorporating the dynamic model leads to estimators that cannot be solved in closed form, the particle filter framework is used instead. Modeling the evolution of the spectral amplitudes leads to low levels of musical noise without resorting to heuristic smoothing methods. Not using the content independent smoothing leads to better preservation of low-energy high-frequency transient speech components, and less smoothing of speech onsets. The particle filter framework also allows the use of additional information and more complex and accurate models: a multiple model approach to dealing with speech presence uncertainty was found to improve noise reduction at speech/non-speech transitions; while incorporating phase estimation was found to offer an improvement in low SNR environments. The proposed algorithm is flexible and the speech and noise distributions can be changed without modifying the rest of the algorithm. This work was described in Chapter 6 and has been published in [21].

4. *Binaural speech enhancement:* A binaural extension of the spectral amplitude speech enhancement algorithm was proposed, making it suitable for future binaural hearing aids. A model of the noise-field interaural coherence was used to derive Wiener filter expressions capable of separating the speech and noise spectral components in each pair of binaural noisy speech frames. The Wiener filter estimated speech component can be used directly to produce enhanced speech and the output offered good noise reduction with mild musical noise. However, a better result was achieved by using the instantaneous Wiener filter

spectral estimates in the noise and speech process parameter estimation steps of the STSA-PF. The resulting binaural particle filter system does not require external noise estimation or voice activity detection and preserves spatial cues. It was shown to offer good noise reduction and speech preservation with minimal musical noise, offering a more natural sounding output than bilateral Wiener and particle filter approaches. This work was outlined in Chapter 7 and has been published in [22].

8.2 Suggestions for Future Research

There are several areas where the algorithms investigated in this work could be extended. In particular, as the variations of the spectral amplitude particle filter demonstrate, the flexibility of the particle filter framework makes it convenient to incorporate additional information or models. Possible extensions to the work in this thesis are outlined as follows:

1. *Integrated speech and noise process tracking:* The spectral amplitude particle filter algorithm presented in this thesis uses a dynamic model of the speech process, but relies on external estimates of the noise parameters. The algorithms in [61] and [62] achieved improvements in speech recognition performance by using particle filter methods to estimate the noise corrupting the speech features. Combining these approaches into a single particle filter algorithm that jointly estimates the speech and noise processes may improve performance in highly non-stationary environments.

2. *Expanded speech modeling:* The speech enhancement algorithms presented in Chapters 5, 6 and 7 assume independence of the transform coefficients in each

frequency bin. This is a common assumption which is generally made for model simplicity and computational convenience; however it does not hold for the short speech frames used in practice. Expanding the algorithms to model the inter-frequency transform coefficient relationships may reduce musical noise and speech distortion.

3. *Multiple active speech models:* The speech active state of the interacting multiple model STSA-PF algorithm outlined in Chapter 6 could be divided into to several sub-states representing different phoneme classes such as voiced or unvoiced. Additionally, different dynamic models could be applied to different frequency regions. While the AR and random walk models are well suited to the harmonic content present at lower frequencies, performance in the high-frequency regions may be improved with models that account for signal content that is more transient with less time-correlation.

4. *Binaural source tracking:* The binuaral algorithm presented in Chapter 7 assumes the source to be approximately frontally located, in practice the source may be non-frontal or even moving. The proposed near-frontal target algorithm could be used along with non-frontal estimators, such as the one presented in [55], in a system that combines source tracking with speaker location dependent estimation.

List of References

[1] R. Plomp, "Auditory handicap of hearing impairment and the limited benefit of hearing aids," *J. Acoust. Soc. Am.*, vol. 63, no. 2, pp. 533 – 549, 1978.

[2] B. M. Clopton and F. A. Spelman, "Auditory system," in *The Biomedical Engineering Handbook: Second Edition.* (J. D. Bronzino, ed.), CRC Press LLC, 2000.

[3] S. Kochkin, "MarkeTrak VII: Hearing Loss Population Tops 31 Million," *HEARING REVIEW*, vol. 12, no. 7, pp. 16 – 29, 2005.

[4] S. Kochkin, "Customer satisfaction with hearing instruments in the digital age," *Hearing J.*, vol. 58, no. 9, pp. 30 – 37, 2005.

[5] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acous. Soc. Am.*, vol. 19, no. 1, pp. 90 – 119, 1947.

[6] H. Puder and B. Beimel, "Controlling the adaptation of feedback cancellation filters –problem analysis and solution approaches," in *Proc. EUSIPCO'04*, (Vienna, Austria), Sept. 2004.

[7] S. Haykin, *Adaptive Filter Theory.* Prentice Hall, 4 ed., 2002.

[8] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 345 – 349, Apr. 1994.

[9] A. Bronkhorst, "The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker contitions," *Acustica*, vol. 86, pp. 117 – 128, 2000.

[10] J. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, pp. 1725 – 1736, 1990.

[11] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 27, pp. 113 – 120, Apr. 1979.

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

[13] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Proc. IEEE ICASSP*, (Dallas, TX, USA), pp. 177 – 180, 1987.

[14] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, pp. 1732 – 1742, Aug. 1991.

[15] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2915–29, 10/15 2005.

[16] T. Van den Bogaert, T. Klasen, M. Moonen, L. Van Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acous. Soc. Am.*, vol. 119, pp. 515 – 526, Jan. 2006.

[17] B. Laska, R. Goubran, and M. Bolić, "Improved proportionate subband NLMS for acoustic echo cancellation in changing environments," *IEEE Sig. Process. Letters*, vol. 15, pp. 337 – 340, 2008.

[18] B. Laska, M. Bolić, and R. Goubran, "Subband adaptive filtering for acoustic feedback compensation in hearing aids," *Canadian Acoustics*, vol. 37, pp. 134 – 135, Sept. 2009.

[19] B. Laska, R. Goubran, and M. Bolić, "Subband autoregressive modelling for speech enhancement," *Canadian Acoustics*, vol. 37, pp. 62 – 63, Sept. 2009. (CAA 2009 student presentation award winner).

[20] B. Laska, M. Bolić, and R. A. Goubran, "Discrete cosine transform particle filter speech enhancement." Manuscript accepted for publication in *Elsevier Speech Communication*, April 2010.

[21] B. Laska, M. Bolić, and R. A. Goubran, "Particle filter enhancement of speech spectral amplitudes," *To Appear in: IEEE Trans. Audio Speech Lang. Process.*, vol. 18, Aug. 2010.

[22] B. Laska, M. Bolić, and R. A. Goubran, "Coherence-assisted wiener filter binaural speech enhancement." To be published in Proc. IEEE *I²MTC.*, (Austin, TX, USA), May 2010.

[23] V. Hamacher, E. Fischer, U. Kornagel, and H. Puder, "Applications of adaptive signal processing methods in high-end hearing aids," in *Topics in Acoustic Echo And Noise Control: Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise, And Speech Processing* (E. Hänsler and G. Schmidt, eds.), Springer, 2006.

[24] J. Benesty and S. Gay, "An improved PNLMS algorithm," in *Proc. IEEE ICASSP*, vol. 2, (Orlando, FL, USA), pp. 1881 – 1884, May 2002.

[25] J. Kates, "Feedback cancellation in hearing aids: results from a computer simulation," *IEEE Trans. Sig. Process.*, vol. 39, pp. 553 – 562, Mar. 1991.

[26] M. G. Siqueira and A. Alwan, "Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 443–453, 2000.

[27] P. C. Loizou, *Speech Enhancement Theory and Practice.* CRC Press, 2007.

[28] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, vol. 4, pp. 208 – 211, Apr 1979.

[29] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 30, pp. 679 – 681, Aug. 1982.

[30] P. Vary, "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387 – 400, Jul. 1985.

[31] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 28, pp. 137–145, Apr 1980.

[32] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Commun.*, vol. 24, pp. 249 – 257, Jun. 1998.

[33] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 33, no. 2, pp. 443 – 445, 1985.

[34] I. Cohen, "Relaxed statistical model for speech enhancement and a priori snr estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 870 – 881, Sept. 2005.

[35] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 845 – 856, Sept. 2005.

[36] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *Eurasip J. Applied Signal Process.*, vol. 2005, pp. 1110 – 1126, May 2005.

[37] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook* (R. Dorf, ed.), vol. 3, Boca Raton: CRC Press, 2006.

[38] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Commun.*, vol. 10, no. 1, pp. 45 – 67, 1991.

[39] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 251 – 266, Jul. 1995.

[40] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process*, vol. 11, pp. 334 – 341, Jul. 2003.

[41] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process*, vol. 6, no. 4, pp. 373 – 385, 1998.

[42] D. Labarre, E. Grivel, M. Najim, and N. Christov, "Dual h∞ algorithms for signal processing – application to speech enhancement," *IEEE Trans. Sig. Process.*, vol. 55, pp. 5195 – 5208, Nov. 2007.

[43] W.-R. Wu and P.-C. Chen, "Subband kalman filtering for speech enhancement," *IEEE Trans. Circuits Systems II*, vol. 45, pp. 1072 – 1083, Aug. 1998.

[44] H. Puder, "Noise reduction with Kalman-filters for hands-free car phones based on parametric spectral speech and noise estimates," in *Topics in Acoustic Echo And Noise Control: Selected Methods for the Cancellation of Acoustical Echoes,*

the Reduction of Background Noise, And Speech Processing (E. Hänsler and G. Schmidt, eds.), Springer, 2006.

[45] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, (Rosslyn, VA, USA), pp. 209 – 13, 2002.

[46] T. Lotter, C. Benien, and P. Vary, "Multichannel speech enhancement using bayesian spectral amplitude estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, (Hong Kong, China), pp. 880 – 883, 2003.

[47] A. Kaps, "Acoustic noise reduction using a multiple-input single-output kalman filter," in *Proc. IWAENC*, (Eindhoven, The Netherlands), pp. 133 – 136, Sept. 2005.

[48] T. Klasen, M. Moonen, T. Van den Bogaert, and J. Wouters, "Preservation of interaural time delay for binaural hearing aids through multi-channel wiener filtering based noise reduction," in *Proc. IEEE ICASSP*, vol. Vol. 3, (Philadelphia, PA, USA), pp. 29–32, Mar 2005.

[49] T. den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel wiener filter," in *Proc. IEEE ICASSP*, vol. 4, (Honolulu, HI, USA), pp. 565 – 568, April 2007.

[50] V. Hamacher, "Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids," in *Proc. IEEE ICASSP*, vol. 4, (Orlando, FL, USA), pp. 4008 – 4011, May 2002.

[51] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912 – 915, 1977.

[52] R. Le Bouquin and G. Faucon, "Using the coherence function for noise reduction," *IEE Proc. Commun. Speech Vision*, vol. 139, pp. 276 – 280, Jun. 1992.

[53] R. Le Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 484 – 487, Sep. 1997.

[54] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 709 – 716, Nov. 2003.

[55] A. H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 521 – 533, May 2009.

[56] T. Wittkop, *Two-channel noise reduction algorithms motivated by models of binaural interaction.* PhD thesis, Carl von Ossietzky University Oldenburg, Mar. 2001.

[57] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, pp. 174 – 88, Feb. 2002.

[58] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/nongaussian bayesian state estimation," in *IEE Proc. -F (Radar and Sig. Process.)*, vol. 140, pp. 107 – 113, Apr. 1993.

[59] M. K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *Journal of the American Statistical Association*, vol. 94, pp. 590–599, Jun. 1999.

[60] W. Fong, S. Godsill, A. Doucet, and M. West, "Monte carlo smoothing with application to audio signal enhancement," *IEEE Trans. Sig. Process.*, vol. 50, pp. 438 – 449, Feb. 2002.

[61] K. Yao and S. Nakamura, "Sequential noise compensation by sequential monte carlo method," in *Adv. Neural Inform. Process. Syst.* (T. G. Dietterich, S. Becker, and Z. Ghahramani, eds.), vol. 14, (Cambridge, MA), pp. 1205 – 1212, MIT Press, 2002.

[62] R. Singh and B. Raj, "Tracking noise via dynamical systems with a continuum of states," in *Proc. IEEE ICASSP*, vol. 1, pp. 396 – 399, Apr. 2003.

[63] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," in *Proc. IEEE ICASSP*, vol. 1, pp. 965 – 968, May 2004.

[64] M. Wolfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 312 – 323, Feb. 2009.

[65] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill, "Particle methods for bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 173 – 185, Mar. 2002.

[66] F. Mustière, M. Bouchard, and M. Bolić, "Quality assessment of speech enhanced using particle filters," in *Proc. IEEE ICASSP*, (Honolulu, HI, USA), pp. 1197 – 1200, Apr. 2007.

[67] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus." Linguistic Data Consortium, Philadelphia `http://www.ldc.upenn.edu/Catalog/LDC93S1.html`, 1993.

[68] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, pp. 2224 – 2237, Apr. 2005.

[69] ANSI, "American national standard: Methods for the calculation of the speech intelligibility index," ANSI S3.5-1997, American National Standards Institute, New York, 1997.

[70] K. Rhebergen and N. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, pp. 2181 – 2192, Apr. 2005.

[71] ITU-T, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," Recommendation P.862.2, International Telecommunication Union, Nov 2007.

[72] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Language Process.*, vol. 16, pp. 229 – 238, Jan. 2008.

[73] M. M. Sondhi and W. Kellermann, "Adaptive echo cancellation for speech signals," in *Advances in speech signal processing* (S. Furui and M. M. Sondhi, eds.), pp. 327 – 356, New York: Marcel Dekker, 1992.

[74] D. Morgan, "Slow asymptotic convergence of LMS acoustic echo cancelers," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 126 – 36, Mar. 1995.

[75] P. Eneroth and T. Gänsler, "Analysis of subband impulse responses in subband echo cancelers," tech. rep., Department of Applied Electronics, Signal Processing Group Lund University, 1999.

[76] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. An application of very-high-order adaptive filters," *IEEE Sig. Process. Mag.*, vol. 16, pp. 42 – 69, Jul 1999.

[77] M. Siqueira, R. Speece, E. Petsalis, A. Alwan, S. Soli, and S. Gao, "Subband adaptive filtering applied to acoustic feedback reduction in hearing aids," in *Proc. Asilomar Conf. Sig. Sys. Comp.*, vol. 1, pp. 788 – 792, Nov 1996.

[78] M. Harteneck, S. Weiss, and R. W. Stewart, "Design of near perfect reconstruction oversampled filter banks for subband adaptive filters," *IEEE Trans. Circ. Sys. II*, vol. 46, no. 8, pp. 1081 – 1085, 1999.

[79] M. Schnell, R. Geiger, M. Schmidt, M. Multrus, M. Mellar, J. Herre, and G. Schuller, "Low delay filterbanks for enhanced low delay audio coding," in *Proc. IEEE WASPAA*, (New Paltz, NY, United states), pp. 235 – 238, 2007.

[80] S. Weiss, L. Lampe, and R. Stewart, "Efficient implementations of complex and real valued filter banks for comparative subband processing with an application to adaptive filtering," in *Proc. Int. Symp. Comm. Sys. DSP*, vol. 1, (Sheffield, UK), pp. 32 – 35, 1998.

[81] O. Hoshuyama, R. A. Goubran, and A. Sugiyama, "A generalized proportionate variable step-size algorithm for fast changing acoustic environments," in *Proc. IEEE ICASSP*, vol. 4, (Montreal, Canada), May 2004.

[82] J. A. Maxwell and P. M. Zurek, "Reducing acoustic feedback in hearing aids," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 304 – 313, 1995.

[83] Y. Deng and V. J. Mathews, "Subband particle filtering for speech enhancement," in *Proc. Eurasip EUSIPCO*, (Florence, Italy), Sep. 2006.

[84] F. Daum and J. Huang, "Curse of dimensionality and particle filters," in *Proc. IEEE Aerospace Conf.*, vol. 4, pp. 1979–1993, Mar. 2003.

[85] S. Rao and W. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *IEEE Trans. Info. Theory*, vol. 42, pp. 1160 – 1178, Jul 1996.

[86] F. Mustière, M. Bolić, and M. Bouchard, "A modified Rao-Blackwellised particle filter," in *Proc. IEEE ICASSP*, vol. 3, (Toulouse, France), pp. 21 – 24, May 2006.

[87] N. Ma, M. Bouchard, and R. Goubran, "Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations," *IEEE Trans. Audio Speech Language Process.*, vol. 14, pp. 19 – 32, Jan. 2006.

[88] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Speech Enhancement* (J. Benesty, S. Makino, and J. Chen, eds.), pp. 43 – 64, Berlin: Springer, 2005.

[89] E. Zavarehei, S. Vaseghi, and Q. Yan, "Inter-frame modeling of dft trajectories of speech and noise for speech enhancement using kalman filters," *Speech Commun.*, vol. 48, pp. 1545 – 1555, Nov. 2006.

[90] T. Esch and P. Vary, "Speech enhancement using a modified kalman filter based on complex linear prediction and supergaussian priors," in *Proc. IEEE ICASSP*, (Las Vegas, USA), pp. 4877 – 4880, Apr. 2008.

[91] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 498 – 505, Sept. 2003.

[92] F. Faubel and M. Wolfel, "Overcoming the vector taylor series approximation in speech feature enhancement - a particle filter approach," in *IEEE ICASSP*, vol. 4, (Honolulu, USA), pp. 557 – 560, Apr. 2007.

[93] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[94] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. IEEE ICASSP*, vol. 1, (Montreal, Canada), pp. 953 – 956, May 2004.

[95] S. McGinnity and G. Irwin, "Multiple model bootstrap filter for maneuvering target tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, pp. 1006 – 12, Jul. 2000.

[96] H. Driessen and Y. Boers, "Efficient particle filter for jump markov nonlinear systems," *IEE Proc. Radar Sonar Nav.*, vol. 152, pp. 323 – 326, Oct. 2005.

[97] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588 – 601, 2007.

[98] I. Lindevald and A. Benade, "Two-ear correlation in the statistical sound fields of rooms," *J. Acoust. Soc. Am.*, vol. 80, pp. 661 – 664, Aug. 1986.

[99] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *Eurasip Journal on Applied Signal Processing*, vol. 2006, pp. 1 – 14, 2006.